



**Community Analysis of Cyber Security
Experts on Online Social Networks**

By

Mohamad Imad Mahaini

October 2023

*A thesis submitted to the University of Kent, School of Computing
in the subject of Computer Science for the degree of
Doctor of Philosophy (PhD)*

Supervised By

Prof. Shujun Li

Dedicated to ..

my blue sky, my beloved mother:

Raghdaa

my firm ground, my good father:

Hussam Eddin

the journey partner, my devoted wife:

Rama

my radiant stars, my children:

Raghad & Hussam

my pillars, my siblings:

Youser, Yumen, Mohamad & Beshar

Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement under Reference Number 675320 (NeCS, [2015](#)).



“I was taught that the way of progress was neither swift nor easy.”

- Marie Curie -

1867-1934

Acknowledgements

I am filled with enormous gratitude as I begin my acknowledgements for my PhD thesis, a journey that has been made possible through the unwavering support and love of countless individuals.

Achieving a PhD has been a long-standing goal I aspired to reach a few years ago. Today, I proudly present this thesis, a culmination of years of relentless effort and perseverance, even amidst uncertain and challenging times like the COVID-19 pandemic. Throughout this journey, I have been fortunate to receive immense encouragement and support from numerous kind-hearted individuals. Their belief in my abilities has been a driving force behind my success, and I am deeply grateful for their unwavering support.

First and foremost, I extend my heartfelt thanks to my beloved family: my mother (Raghdaa Al-Hennawi), my father (Hussam Mahaini), and my siblings, whose endless encouragement, sacrifices, and belief in my abilities have been the foundation of my academic pursuit since the very beginning when I had my BSc, through my MSc and now the PhD. Your unwavering support has been my lighthouse, guiding me through the high waves and strong winds of this life, and I am forever indebted to you both.

I extend my deepest appreciation to my dear wife, Rama, and my children, Raghad and Hussam. Your constant encouragement, understanding, and patience have been my greatest source of strength and motivation throughout this challenging journey. Your presence has been a beacon of hope, reminding me of the greater purpose of my work and giving me every reason to carry on.

I am grateful to Professor Shujun Li, my esteemed research supervisor, for his exceptional guidance, boundless knowledge, and mentorship. His insights, constructive feedback, and dedication to my growth as a researcher have been invaluable in shaping my thesis. I would like to thank also Professor David Chadwick for his supervision and feedback during the first year and a half of my research journey.

I am deeply indebted for the grant from the NeCS project, a network funded by

the European Union Horizon2020 Marie Skłodowska-Curie Actions program. This life-changing opportunity has made a significant difference and contributed massively to my research, training, career, and networking endeavours. I sincerely thank all my research colleagues, especially the NeCS project people, ESRs, and leaders, who have been integral to this journey. The collaborative spirit, diverse perspectives, and enriching discussions within our research community have broadened my horizons and inspired me to pursue excellence. Additionally, I want to thank my colleagues from the Institute of Cyber Security for Society (iCSS) at the School of Computing, University of Kent. Their support and contributions have been invaluable throughout this endeavour.

A heartfelt thanks to my dear friends, Adham and Kenan, for being supportive during the ups and downs of this arduous path. Your unwavering belief in my abilities, encouraging words, and shared moments of joy have made this journey memorable. To all those who have supported me, directly or indirectly, I offer my sincerest thanks. Your encouragement, kind words, and belief in my potential have been instrumental in helping me overcome challenges and reach this significant milestone.

In closing, I am deeply humbled and grateful for the love and support I have received from my family, friends, and academic community. Without you all, this achievement would not have been possible. Your presence in my life has made it richer; I am forever thankful for that.

With heartfelt appreciation and warm regards.

Yours,

Mohamad Imad Mahaini

Declaration

I, Mohamad Imad Mahaini, hereby declare that:

- The work presented in this thesis is entirely original and represents my own efforts.
- Any material or ideas obtained from other sources, including published or unpublished works, have been fully acknowledged with proper citations and references.
- No part of this material has been submitted for a degree at this or any other university, either in whole or in part.
- All sources consulted during the research process have been appropriately cited and listed in the references section.
- Any contributions made by others to the research project or to the preparation of the thesis have been duly acknowledged.
- I understand the consequences of academic dishonesty and take full responsibility for the content and presentation of this thesis.

A rough estimate of the word count is 51,357 words.

Signed: [M.I.M](#)

Date: [30 July 2024](#)

Abstract

The advent of Online Social Networks (OSNs) has started a new era of communication and information dissemination, fundamentally altering the way individuals and organisations interact in the digital age. OSNs such as Twitter, Facebook, LinkedIn and YouTube allow billions of people across the globe to create online communities based on similar interests. With almost everyone and everything being touched by cyber space in recent times, and despite the advancement in technology and cyber security, malicious attacks are still targeting individuals, organisations and systems on a large scale and utilising new channels like social media platforms. Cyber security is a field dedicated to safeguarding computer systems, networks, and data from theft, damage, or unauthorised access. Within the complex interconnected nature of OSN users, a distinct category of accounts has gained particular attention and significance: cyber security accounts. These accounts encompass a wide spectrum of groups, such as activists, hacktivists, cyber criminals and cyber security experts.

Cyber security experts include researchers, practitioners, innovators, vendors, etc. While previous studies have explored various types of cyber security related accounts and their communities on OSNs, the activities of cyber security experts have not been sufficiently investigated. This thesis addresses this research gap by designing, developing and testing tools to support studying cyber security experts on OSNs, with a particular focus on cyber security researchers. The tools developed and tested include 1) a general cyber security taxonomy, 2) multiple Machine Learning (ML) classifiers, and 3) some generalisable methods for collecting and analysing data from OSNs. Therefore, the thesis encompasses three main studies as follows.

First, the thesis introduces a novel human-machine teaming-based process for building taxonomies. Many previous studies relied solely on manual efforts, leading to limitations in covering diverse topics and rapidly evolving concepts. The proposed process was applied to the cyber security domain as an example, which allowed human experts to collaborate with automated Natural Language Processing (NLP) and Information Retrieval (IR) tools to co-develop a general cyber security taxonomy

from relevant textual sources with reasonable human effort.

Second, the thesis presents the design and development of several ML classifiers to detect the needed Twitter accounts. They include a baseline classifier for detecting cyber security related accounts in general, and four sub-classifiers for detecting other related sub-groups of accounts (individuals, hackers, academia and research). The classifiers were trained and tested using a systematic approach involving the cyber security taxonomy built earlier, real-time tweet sampling, and crowdsourcing for dataset labelling. By considering a richer set of features than previous studies, the classifiers achieved promising performance, with the Random Forest model outperforming others, with the F1-score reaching 93% for the baseline classifier and 83-91% for the sub-classifiers. Feature reduction analysis demonstrated that a subset of just six features maintained the same performance levels, providing efficient and effective classifiers for detecting cyber security accounts on Twitter.

Third, in the last part, the cyber security researchers were analysed as an example of a sub-group of cyber security experts. As a case study of a research community, the presence of the UK's Academic Centres of Excellence in Cyber Security Research (ACEs-CSR) on Twitter was analysed. Several machine learning classifiers were utilised to identify cyber security and research related accounts, and a social graph was constructed using the friends and followers of the ACE-CSR accounts. Then, a comprehensive analysis was carried out, including community detection, social structural analysis, influence analysis, topic modelling, and sentiment analysis, revealing interesting insights about the research community around ACEs-CSR, such as their sub-communities, top influencers, the influence distribution, the topics being discussed by cyber security researchers, and last but not least the sentiment towards the ACE-CSR programme and accounts. Twitter was used as an example in this thesis, but all the presented methodologies can be applied to other OSNs.

List of Publications

The following peer-reviewed conference papers were published during the course of this research. Each of these publications was based on the material and results discussed in a corresponding chapter as follows.

- Mohamad Imad Mahaini, Shujun Li, and Rahime Belen-Sağlam (2019). “Building Taxonomies based on Human-Machine Teaming: Cyber Security as an Example”. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. ACM, 30:1–30:9. DOI: [10.1145/3339252.3339282](https://doi.org/10.1145/3339252.3339282).
→ This paper influenced the material presented in Chapter 4 and was utilised later in Chapter 5.
- Mohamad Imad Mahaini and Shujun Li (2021). “Detecting Cyber Security Related Twitter Accounts and Different Sub-Groups: A Multi-Classifer Approach”. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, pp. 599–606. DOI: [10.1145/3487351.3492716](https://doi.org/10.1145/3487351.3492716).
→ This paper influenced the material presented in Chapter 5 and set the basis for the research presented in Chapter 6.
- Mohamad Imad Mahaini and Shujun Li (2023). “Cyber Security Researchers on Online Social Networks: From the Lens of the UK’s ACEs-CSR on Twitter”. In: *Security and Privacy in Social Networks and Big Data: 9th International Symposium, SocialSec 2023, Canterbury, UK, August 14–16, 2023, Proceedings*. Vol. 14097. Lecture Notes in Computer Science. Springer, pp. 129–148. DOI: [10.1007/978-981-99-5177-2_8](https://doi.org/10.1007/978-981-99-5177-2_8).
→ This paper influenced the material presented in Chapter 6.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research Aims	5
1.3	Research Questions	6
1.4	Contributions	8
1.5	Thesis Structure	9
2	Background	11
2.1	Online Social Networks (OSNs)	11
2.1.1	What are networks?	11
2.1.2	Definition	12
2.1.3	Six Degrees Concept	13
2.1.4	Genesis & Growth	15
2.1.5	Online Communities	17
2.1.6	Twitter, a major research platform	18
2.2	Social Network Analysis (SNA)	19
2.2.1	Definition	19
2.2.2	Power and Influence	21
2.2.3	General SNA Metrics	21
2.2.4	Network Centrality	23
2.3	Taxonomies	29
2.3.1	Definition and Use	29
2.3.2	The importance of Classification	29

2.3.3	Structure	30
2.3.4	Terminology	32
2.3.5	Components	33
2.3.6	Development Process	34
2.4	Machine Learning (ML)	35
2.4.1	Definition	35
2.4.2	Learning Types	36
2.4.3	Statistical Inference & Machine Learning	38
2.4.4	Supervised Learning Overview	39
2.4.5	Classification vs Regression	40
2.4.6	Supervised Learning Algorithms	41
2.4.7	Classifiers Performance Evaluation	47
2.5	Natural Language Processing (NLP)	50
2.5.1	n-gram	50
2.5.2	Text Metrics	51
2.5.3	Text Readability Scores	56
2.5.4	Topic Modelling	57
3	Related Work	59
3.1	Cyber Security Taxonomies and Ontologies	59
3.1.1	Automatic Taxonomy and Ontology Building	59
3.1.2	Selected Taxonomies in Cyber Security	60
3.1.3	Selected Ontologies in Cyber Security	63
3.2	Detecting Cyber Security Related Accounts	65
3.3	Studying Cyber Security Communities	67
4	Building General Cyber Security Taxonomy	69
4.1	Introduction	69
4.1.1	Cyber Security Taxonomy	70
4.1.2	The Necessity of a “General” Taxonomy	71

4.2	Methodology	72
4.2.1	Stage A: Data Collection	74
4.2.2	Stage B: Text Analysis	75
4.2.3	Stage C: taxonomy-building	75
4.3	Data Collection	77
4.3.1	Building Cyber Security Dataset	78
4.3.2	Building Non-Cyber Security Corpora	79
4.3.3	Text Extraction	80
4.4	Text Analysis	80
4.4.1	Text Processing	80
4.4.2	N-grams Metrics Calculation	82
4.4.3	Terms Selection	83
4.5	Taxonomy Building	87
4.5.1	Defining Taxonomy Structure	87
4.5.2	Assigning Terms to Taxonomy Structure	91
4.5.3	Taxonomy Refinement	93
4.6	Taxonomy Visualisation	94
4.6.1	Initial Taxonomy Structure Visualisation	94
4.6.2	Final Taxonomy Structure Visualisation	95
4.7	Potential Applications	100
4.8	Taxonomy Validation	101
4.9	Conclusion	105
5	Automatic Detection of Cyber Security Accounts on OSNs	107
5.1	Introduction	108
5.2	Methodology	110
5.3	Data Collection	111
5.3.1	Harvesting Tweets	112
5.3.2	Sampling Cyber Security Related Tweets	113
5.3.3	Selecting Candidate Accounts	114

5.4	Dataset Construction	117
5.4.1	Crowdsourcing Labelling	118
5.4.2	Research Ethics and GDPR	121
5.4.3	Participant Recruitment	122
5.4.4	Labelled Dataset	124
5.5	Feature Extraction	125
5.5.1	Profile Features (P)	125
5.5.2	Behavioural Features (B)	127
5.5.3	Content Statistics Features (C)	127
5.5.4	Linguistic Features (L)	128
5.5.5	Keyword-based Features (K)	128
5.6	Classification Tasks & ML Models	132
5.6.1	Detecting Cyber Security Related Accounts	133
5.6.2	Detecting Cyber Security Individual Accounts	133
5.6.3	Detecting Hacker Related Accounts	133
5.6.4	Detecting Cyber Security Academia Related Accounts	134
5.7	Experimental Results	134
5.7.1	Classifier Results	135
5.7.2	Features Importance	138
5.7.3	Features Reduction	140
5.7.4	Comparison to Past Studies in the Literature	142
5.7.5	Crowdsourcing Effects on Overall Results	144
5.8	Conclusion	145
6	Studying Cyber Security Communities on OSNs	147
6.1	Introduction	148
6.2	Cyber Security Communities	150
6.2.1	Analysing Cyber Security Communities	150
6.2.2	Case Study: ACEs-CSR Network on Twitter	150
6.3	Research Questions	152

6.4	Methodology	152
6.5	Data Collection	154
6.5.1	Seed Accounts	154
6.5.2	Friends and Followers of the Seed Accounts	155
6.5.3	Account Timelines	156
6.6	Machine Learning Classifiers	156
6.6.1	Cyber Security Related and Individual Classifiers	157
6.6.2	Research Related Classifier	162
6.7	Social Network Analysis	168
6.7.1	Social Graph Construction	168
6.7.2	Graph Visualisations	170
6.7.3	Network Statistics	170
6.7.4	Community Detection & Analysis	173
6.7.5	Influence Analysis	185
6.8	Topic Modelling Analysis	195
6.8.1	Documents Preparation	195
6.8.2	Applying LDA Algorithm	196
6.8.3	Extracted Topics	198
6.9	Sentiment Analysis	200
6.9.1	Dataset Creation	200
6.9.2	Sentiment Analysis Algorithms	201
6.9.3	Sentiment Analysis Results	202
6.10	Conclusion	203
7	Conclusion & Future Work	205
7.1	Conclusion	205
7.2	Revisiting Research Questions	207
7.3	Summary of Contributions	208
7.3.1	Data-Driven Human-Machine Teaming Approach for Constructing Taxonomies	208

7.3.2	Built General Cyber Security Taxonomy	208
7.3.3	Improved Methodology for Automatic Detection Of Cyber Security Related Accounts	208
7.3.4	Additional Classifiers to Detect Different Types of Cyber Security Accounts	209
7.3.5	Tested the ML Classifiers in a Real-World Setting	209
7.3.6	ACEs-CSR Research Network Analysis	210
7.3.7	Approach to Study Cyber Security Expert Networks	211
7.4	Limitations & Future Work	211
7.4.1	Extend the Taxonomy-Building Approach	211
7.4.2	Enhance the Automatic Detection of Cyber Security Related Accounts and other Sub-groups	213
7.4.3	Further applications for the developed ML classifiers	214
7.4.4	Open Opportunities for Studying Cyber Security Communities on OSNs	214
7.4.5	Extend the Work to Other Languages	215
Appendix A General Cyber Security Taxonomy Webpage		217
Appendix B Labelling Experiment Questionnaire		219
B.1	Part 1: Basic Demographics	219
B.2	Part 2: Cyber Security Experts on OSNs	221
B.3	Part 3: Labelling Process Feedback	223
Bibliography		225

List of Figures

1.1	Thesis Research Question Relations	7
2.1	Six Degrees of Separation (Ma, 2015)	14
2.2	Six Degrees of Kevin Bacon (Reynolds, 1999)	15
2.3	Network example to calculate centrality scores	27
2.4	Hierarchical Taxonomy, example (Laubheimer, 2022)	31
2.5	Faceted Taxonomy, example (Laubheimer, 2022)	32
2.6	Taxonomy components, example (Unterkalmsteiner and Adbeen, 2023)	33
2.7	Taxonomy Building Traditional Approach (Nai Fovino et al., 2019)	35
2.8	Machine Learning Types (CogNub, 2018)	37
2.9	Statistical inference types and their relations (V. N. Vapnik, 1999)	38
2.10	The supervised machine learning process (Grieve, 2020)	39
2.11	SL, Classification & Regression (Krzyk, 2018)	41
2.12	SL, Logistic Regression classifier (Kanade, 2022)	42
2.13	SL, Support Vector Machine classifier (Saini, 2023)	43
2.14	SL, Decision Tree, an example (Navlani, 2023)	44
2.15	SL, Random Forest (Sharma, 2023)	46
4.1	Human-machine teaming-based methodology for taxonomy-building	74
4.2	A visualisation of the initial cyber security taxonomy structure	90
4.3	Assigning terms to taxonomy classes	92
4.4	Visualisation interface: different visualisation methods and features	95
4.5	Visualisation [Flat Tree], showing the “Vulnerability” branch	96
4.6	Visualisation [Sun Burst], showing the “Security Control” branch	97

4.7	Visualisation [Space Tree], showing the “Hacking” branch	98
4.8	Visualisation [Radial Graph]	99
4.9	Visualisation [Hyper Tree], showing the “Attack” branch	100
5.1	Proposed methodology for detecting cyber security accounts	111
5.2	Accounts Selection Verification - Results	116
5.3	Labelling Interface, Tasks List (left) & Information Sheet (centre) . .	119
5.4	Crowdsourcing Labelling Application, Labelling Sheet	121
5.5	Labelling Experiment - Participant Demographics	123
5.6	Labelling Experiment - Participants Cyber Security Experience . . .	124
5.7	The top 20 features of two classifiers, ranked by χ^2 significance values	139
5.8	Feature reduction analysis for the baseline classifier	141
6.1	The Constructed Graph of ACEs-CSR Network (Random View) . . .	171
6.2	The Constructed Graph of ACEs-CSR Network (Focused View) . . .	172
6.3	The communities within the ACEs-CSR social graph ($\gamma = 1$)	176
6.4	Six different visualisations of the ACEs-CSR network with different clustering parameters (C : the number of communities, M : modularity)	177
6.5	Location distribution per continent for ACE-CSR communities	184
6.6	Top 5% nodes of the ACEs-CSR graph using eigenvector centrality .	190
6.7	The degree centrality scores distribution, ACEs-CSR Network	192
6.8	The in-degree centrality scores distribution, ACEs-CSR Network . . .	192
6.9	The out-degree centrality scores distribution, ACEs-CSR Network . .	193
6.10	The betweenness centrality scores distribution, ACEs-CSR Network .	193
6.11	The eigenvector centrality scores distribution, ACEs-CSR Network . .	194
6.12	The PageRank centrality scores distribution, ACEs-CSR Network . .	194
6.13	Visualisation of the estimated topics by the LDA algorithm	197
6.14	Size comparison of the estimated topics by the LDA algorithm	198
6.15	Sentiment Analysis, Statistics on Tweets Related to ACEs-CSR . . .	202
A.1	Taxonomy Visualisation Webpage	217

List of Tables

2.1	Market appearance timeline of some selected OSNs from 1997 to 2020	16
2.2	Comparison of centrality measures	27
2.3	ML Binary Classification, Confusion Matrix	48
4.1	Data sources statistics in the cyber security corpus	79
4.2	Statistics of all five corpora used	79
4.3	Statistics of extracted n -grams for each corpus	81
4.4	Statistics of the initial filtering step	84
4.5	Examples of n -grams eliminated by the coverage rules	86
5.1	Statistics of Tweets Sampling step	112
5.2	Statistics of the n -grams list review process	114
5.3	List of all extracted features and their main groups	126
5.4	Top unigrams by each keyword metric from the cyber security corpus	131
5.5	Overall experimental results for the four classifiers	136
6.1	The 19 UK universities recognised as ACEs-CSR	151
6.2	Statistics of accounts and levels in the initial ACE-CSR network . . .	156
6.3	Overall experimental results for the four old classifiers	159
6.4	The prediction results for the four old classifiers	160
6.5	Manual validation results for the classifier predictions	161
6.6	The Researcher Keywords in the Twitter Name & Description fields .	164
6.7	Experimental results of the Research classifier	166
6.8	The prediction results using the Research classifier	168
6.9	The different graphs within the ACEs-CSR Twitter network	169

6.10	Network Statistics of the ACE-CSR directed graph	173
6.11	Resolution, Modularity and Communities using Leiden Algorithm . .	179
6.12	Statistics of discovered communities ($\gamma = 1$)	180
6.13	Twitter location field statistics in the studied datasets	182
6.14	ACE-CSR community members distribution per continent	183
6.15	Top 40 nodes ranked by different centrality measures	187
6.16	LDA topics with top 20 keywords, ranked in descending order by size	199
6.17	Examples of tweets wrongly classified by TextBlob sentiment analyser	201
6.18	Sentiment Analysis, Statistics on Tweets Related to ACEs-CSR . . .	203
A.1	Cyber Security Taxonomy Files	218

List of Abbreviations

ACE-CSR	A cademic C entres of E xcellence in C yber S ecurity R esearch
API	A pplication P rogramming I nterface
BC	B etweenness C entrality
CC	C loseness C entrality
CI	C ritical I nfrastructure
CTI	C yber T hreat I ntelligence
CVE	C ommon V ulnerabilities and E xposures
CWE	C ommon W eakness E numeration
DC	D egree C entrality
DF	D ocument F requency
DT	D ecision T ree
EC	E igenvector C entrality
ENISA	E uropean N etwork and I nformation S ecurity A gency
ET	E xtra T rees
F-K	F lesch- K incaid
GDPR	G eneral D ata P rotection R egulation
GUI	G raphical U ser I nterface
IAM	I ntity and A ccess M anagement
IDC	I n D egree C entrality
IDF	I nverse D ocument F requency
IR	I nformation R etrieval
LIWC	L inguistic I nquiry and W ord C ount
LR	L ogistic R egression
ML	M achine L earning
NCSC	N ational C yber S ecurity C entre
NeCS	E uropean N etwork for C yber S ecurity
NIS	N etwork and I nformation S ecurity

NIST	N ational I nstitute of S tandards and T echnology
NLP	N atural L anguage P rocessing
ODC	O ut D egree C entrality
OSN	O nline S ocial N etwork
PC	P ageRank C entrality
PI	P ersonal I nformation
PIS	P articipant I nformation S heet
RBF	R adial B asis F unction
REAG	R esearch E thics A dvisory G roup
RF	R andom F orest
RQ	R esearch Q uestion
SecOps	S ecurity O perations
SL	S upervised L earning
SM	S ocial M edia
SMOG	S imple M easure of G obbledygook
SNA	S ocial N etwork A nalysis
STS	S ocio- T echnical S ystem
SVM	S upport V ector M achine
SVM-L	SVM with L inear kernel
SVM-R	SVM with RBF kernel
TF	T erm F requency
TF-IDF	T erm F requency- I nverse D ocument F requency
UGC	U ser G enerated C ontent
VADER	V alence A ware D ictionary and s Entiment R easoner
XGBoost	e Xtreme G radient B oosting

Chapter 1

Introduction

“A goal without a plan is just a wish.”

- Antoine de Saint-Exupéry

1.1 Motivation

DURING the first decade of this century, online social networks (OSNs) like Facebook and Twitter have witnessed remarkable expansion in user registrations and social engagement (Bilge et al., 2009). These platforms enable individuals to share a wide array of information, encompassing news, photos, videos, emotions, personal details, and even research endeavours (Adewole et al., 2016). These platforms have seamlessly integrated into our daily routines, becoming pervasive in modern society and fundamentally shaping how people connect, share, and consume information. Some industry experts argue that you are not part of cyber space if you are not using Facebook, YouTube or Second Life, as everything nowadays is about social media (Kaplan and Haenlein, 2010).

Numerous studies have highlighted that a significant portion of people’s time is spent on various online platforms like Facebook, MySpace, Twitter, YouTube, and other blogosphere (Hajeer et al., 2013). Thanks to the fast increase in mobile internet access, individuals can now engage with OSNs anytime and anywhere. These

platforms have given rise to an extensive body of User Generated Content (UGC), which has evolved into a significant form of electronic word of mouth in today's digital landscape (Mao, Zhou, and Xiong, 2020) and information sharing across the globe. This evolving landscape highlights how online social media applications have transcended their initial roles and are now instrumental in shaping modern interactions and information dissemination strategies.

These networks not only connect people on a massive scale but also play a pivotal role in the formation of communities centred around specific topics or interests. The prevalence of OSNs has given rise to diverse online communities, enabling individuals to engage with like-minded peers, stay informed, and participate in discussions aligned with their interests and professional pursuits. In essence, social networks are composed of various connections, such as friendships and other acquaintances among individuals. It is a common observation that these networks display a community structure characterised by groups of vertices with dense internal connections and sparser connections between these groups (Girvan and Newman, 2002).

The proliferation of online platforms, facilitating the gathering and dissemination of information, has brought about a concurrent increase in cyber crimes targeting not only individuals but entire communities. Consequently, researchers and practitioners have embarked on a quest to comprehend this virtual landscape and the methods through which socially and digitally connected individuals can be exploited (Carley, 2020). This ongoing exploration has given rise to a novel scientific and engineering field known as "social cyber security", as described by Carley in their study. This term is commonly accepted as the "socio-technical aspects of cyber security" based on the Socio-Technical Systems (STS) theory, which revolves around an approach aimed at optimising the coordination and harmony between the social and technical aspects of a system, all while taking into account the system's external environment (Beekun, 1989). STS theory emphasises the interplay between social and technical elements in any system, including the cyber security domain, since cyber security is not solely a technical concern but also a social one.

Social cyber security differs significantly from traditional cyber security, which primarily concerns the security of machines, computers, and databases against potential compromises. In contrast, social cyber security places its focus on the human element, exploring how individuals can be compromised, influenced, or relegated to the unimportant in the digital realm (Carley, 2020). While cyber security experts are typically well-versed in technology, computer science, and engineering, social cyber security experts require expertise in fields such as social communication, statistics, Social Network Analysis (SNA), community building, and Machine Learning (ML).

OSNs are no longer exclusively utilised by individuals; instead, they have increasingly attracted the attention of various organised collectives, including legitimate enterprises. When it comes to cyber security accounts on OSNs, we can distinguish two main categories: the adversaries or nefarious accounts and the cyber security expert accounts. For the first category, notably, groups such as activists, hacktivists, and cyber criminals have recognised the potential of these platforms as effective communication tools for disseminating their ideologies and messages (Nouh and Nurse, 2015). For the second category, within these networks a myriad of accounts and profiles are dedicated to cyber security, ranging from individual cyber security experts and research organisations to governmental and commercial entities. These accounts often engage in discussions related to different topics in cyber security such as emerging threats, cyber attacks, data breaches, best practices, research findings, privacy laws, and policy matters. Thus, OSNs have become pivotal arenas for discourse, collaboration, dissemination of knowledge and, more importantly, for fighting and mitigating the harm caused by nefarious activities of hackers and different cyber criminal groups (Adewole et al., 2016; Ferrara et al., 2016).

It is crucial to identify the accounts used by cyber criminals or hackers and understand these accounts and their activities. First, they pose significant threats to individual users, who can become victims of fraud, identity theft, or any other form of cyber crime. Second, they can disrupt the normal functioning of OSNs and degrade user trust in these platforms. Lastly, through their influence on public opin-

ion, these accounts can have broader societal impacts, including the manipulation of political processes and the spread of harmful ideologies (Bradshaw and Howard, 2019). Studying the communities formed by these accounts can provide insights into their strategies, behaviours, and targets, which can be used to develop more effective defence mechanisms. This is an active area of research, with various methods being proposed, including network analysis, machine learning, and user behaviour analysis (Jones, Nurse, and Li, 2022; Aslan, Li, et al., 2020; Jones, Nurse, and Li, 2020; Tavabi et al., 2019; Kigerl, 2018).

On the other hand, cyber security experts encompass a diverse group of individuals, including researchers, practitioners, innovators, vendors, and more. While prior research has delved into different categories of cyber security related accounts and their respective communities on OSNs, there remains a gap in our understanding of the activities and interactions of cyber security experts. Studying the activities and communities of experts on social media can provide valuable insights into their communities, activities, communication patterns, discussions, influence, etc.

The motivation behind the research presented in this thesis stems from the need to design a practical approach and develop and test the needed tools and methods that allow studying cyber security expert accounts and communities on OSNs. Thus, the following pieces of research were presented in the course of this thesis.

First, the thesis introduces a novel human-machine teaming-based process for building taxonomies that further contributes to the advancement of knowledge organisation and systematisation in the dynamic and interdisciplinary field of cyber security. Using this methodology, a general cyber security taxonomy was built, which helped capture and organise the knowledge in the cyber security domain.

Second, the thesis presents an enhanced methodology for detecting cyber security related accounts and other sub-groups (individuals, hackers, academia and research) on OSNs. Designing and building reliable ML classifiers was necessary to automatically identify any required group of cyber security related accounts that we intend to study. It is important to note that the classes and concepts of the created

taxonomy were used as additional features to improve the developed classifiers.

Third, using the tools mentioned above and other techniques, we studied the cyber security experts on OSNs, focusing on a specific sub-type of experts, which was the cyber security research accounts. As a case study, we analysed the presence of the UK's Academic Centres of Excellence in Cyber Security Research (ACEs-CSR) (NCSC, 2019) on Twitter. See Section 6.2.2 for more about ACEs-CSR.

Overall, the research conducted throughout this thesis offers a comprehensive analysis of cyber security experts on OSNs, focusing on cyber security researchers with potential implications for Cyber Threat Intelligence (CTI), research collaboration, and cyber security awareness activities.

1.2 Research Aims

Studying cyber security experts' activities, communities and discussions on OSNs is quite important for many reasons, we will mention a few below.

- **Understanding Threat Landscape:** By analysing the activities of cyber security experts, researchers and practitioners can gain insights into emerging threats, attack patterns, and vulnerabilities. This understanding is essential for devising effective defence strategies and enhancing overall cyber security posture.
- **Identifying Trends and Patterns:** Studying the behaviours and interactions of cyber security experts allows for the identification of trends and patterns within the cyber security community. This knowledge can inform decision-making processes, policy development, and resource allocation in combating cyber threats.
- **Enhancing Collaboration:** By finding online communities of cyber security experts, it becomes possible to identify potential collaborators, thought leaders, and influencers within the field. This facilitates knowledge sharing, collaboration on research projects, and the dissemination of best practices and lessons learned.

- **Monitoring and Early Warning:** Monitoring the activities and discussions of cyber security experts on OSNs can serve as an early warning system for emerging threats and attacks. Timely identification of such threats enables proactive measures to be taken to mitigate risks and protect against potential cyber attacks.
- **Informing Policy and Regulation:** Insights gathered from studying cyber security experts' activities and discussions can inform the development of policies, regulations, and guidelines related to cyber security. This includes areas such as data protection, incident response, threat intelligence sharing, and ethical considerations in cyber security research and practice.

The thesis aims to set a practical approach and develop the needed tools and methods that facilitate studying cyber security experts' activities and communities on OSNs, which is vital for staying abreast of evolving cyber threats, fostering collaboration and knowledge sharing, and informing decision-making processes in the field of cyber security. It is worth mentioning that this thesis does not set any hypotheses to prove or disprove about the cyber security experts, their behaviours or their communities. The thesis can be considered an exploratory study to identify cyber security experts on OSNs, analyse their communities, and learn insights about their discussions and behaviours.

1.3 Research Questions

We broke the main research aim into sub-research problems; each was addressed in a dedicated chapter. Thus, we set our high-level research questions as follows.

- **RQ1:** How to build a general cyber security taxonomy using a data-driven approach?
- **RQ2:** How to develop reliable machine learning classifiers to identify cyber security related accounts on OSNs and other related sub-groups?

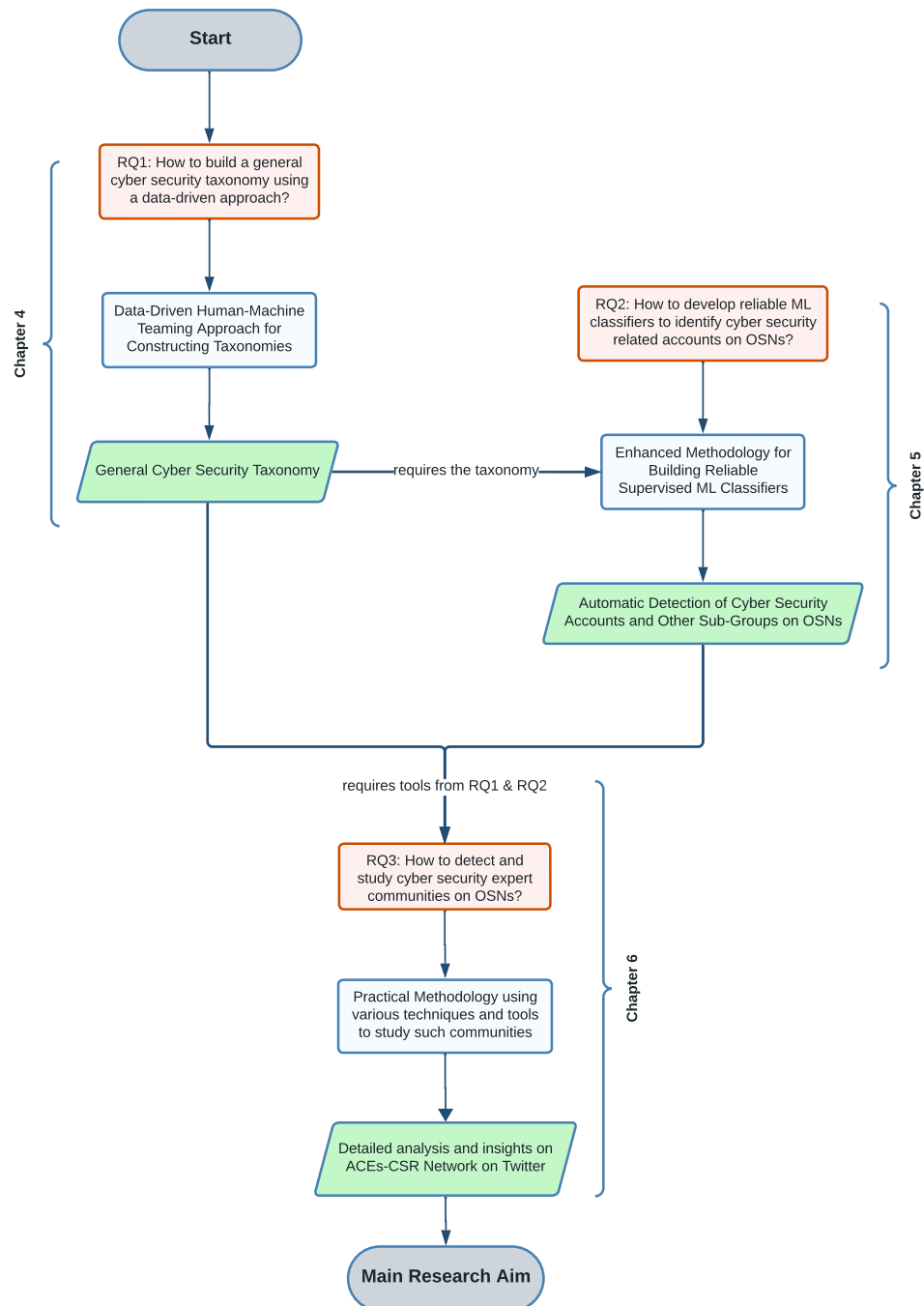


Figure 1.1: Thesis Research Question Relations

- **RQ3**: How to detect and study cyber security expert communities on OSNs?

RQ3 corresponds to the main research aim of this thesis. However, **RQ3** requires some tools and methods to enable the automatic detection of cyber security related

accounts and other sub-groups on OSNs. Thus, we had to finish **RQ1** and **RQ2** before **RQ3**. Also, **RQ2** requires the completion of **RQ1**. See Figure 1.1 for a visual representation of the thesis research questions, their outputs and how they are connected to each other to achieve the research aim. To address **RQ1**, a data-driven human-machine teaming approach for constructing taxonomies is introduced, while for **RQ2**, an enhanced methodology is presented to build reliable supervised machine learning classifiers to detect such accounts and other related sub-groups.

1.4 Contributions

This thesis makes several major contributions in response to the research questions, which are listed below.

- A data-driven human-machine teaming approach for constructing taxonomies was introduced. A general cyber security taxonomy was built using diverse textual sources as an example and direct application of this approach.
- A methodology based on machine learning classifiers to improve the automatic detection of cyber security related accounts was designed and implemented.
 - Three other ML classifiers were created to capture different types of cyber security accounts (individuals, hackers, and academia).
 - The labelled dataset used for training and testing the ML classifiers was built using a crowdsourcing method leveraging the expertise of cyber security experts and knowledgeable participants in the field.
 - The features were analysed based on their importance to the classification process. This resulted in identifying a smaller set of features that achieved the same performance, suggesting lightweight classifier versions can be built.
 - The ML classifiers for detecting cyber security accounts and other sub-groups were tested in a real-world setting, showing that they maintain good performance when they are used on a new dataset.

- An additional ML classifier was developed to detect cyber security research accounts in order to capture the required group of accounts that were considered for the analysis of the ACEs-CSR network on Twitter.
- A practical methodology to study cyber security expert communities was introduced, taking the cyber security research accounts as an example of an expert group. Thus, as a case study, the ACEs-CSR network on Twitter was analysed using a wide range of tools and methods, revealing interesting insights and proving the practicality of conducting such studies on cyber security expert communities.

1.5 Thesis Structure

To address the research questions, the rest of the thesis is organised as follows.

Chapter 2 covers the background, which presents some selected topics covering the following areas: i) Online Social Networks (OSNs), ii) Social Network Analysis (SNA) and network centrality, iii) Taxonomies (classification, structure, components, development process,..etc), iv) Natural Language Processing (NLP), including text metrics and Topic Modelling (TM) and v) Machine Learning, presenting the basics of supervised ML and focusing on the models that were used in this thesis. The background topics cover the required basic knowledge that readers of this thesis should know before going through the presented work. However, experts in the aforementioned topics can skip most of this chapter.

Chapter 3 presents the related work, which includes selected studies about: 1) cyber security taxonomies and ontologies, 2) automatic classification of accounts on OSNs in general and cyber security related classification tasks in particular, and 3) studying cyber security accounts' behaviours and communities on OSNs.

Chapter 4 addresses the first research question **RQ1** by introducing a human-machine teaming-based process for constructing taxonomies. The presented process is data-driven, where human experts can utilise automated NLP and IR tools to develop taxonomies from relevant textual documents. The process can be generalised

to support non-textual documents and to build complex ontologies as well. The cyber security domain serves as an illustrative example, showcasing the effectiveness of the proposed taxonomy-building process in creating a general and data-driven cyber security taxonomy with moderate human involvement and reasonable time.

Chapter 5 answers **RQ2**, presenting our efforts in developing ML classifiers to identify cyber security related accounts on Twitter. We created a baseline classifier to detect cyber security accounts and three sub-classifiers for specific sub-groups of cyber security accounts (individuals, hackers, and academia). To train and test the classifiers, we employed a systematic approach that involved constructing a labelled dataset with multiple tags for each account, using a cyber security taxonomy, real-time tweet sampling, and crowdsourcing. A richer set of features was utilised compared to previous studies. Among the evaluated ML models, the Random Forest model performed the best, achieving 93% for F1-score for the baseline classifier and 88-91% for the three sub-classifiers. Additionally, we investigated feature reduction and found that using only six features maintains comparable performance.

Chapter 6 addresses **RQ3** and bridges the gap in the literature about studying cyber security expert communities on OSNs. The chapter presents an analysis of the UK's ACEs-CSR presence on Twitter, where ML classifiers were employed to identify cyber security and research related accounts. Starting from the 19 ACE-CSR seed accounts, a network graph of 1,817 research-related accounts was constructed. The study used social structural analysis, influence analysis, topic modelling, and sentiment analysis, revealing insights like sub-community detection, key discussion topics, and positive sentiment towards ACE-CSR. The results demonstrated the value of automated analysis of cyber security expert networks on OSNs.

Chapter 7 provides a comprehensive conclusion to this thesis, summarising the key findings and contributions made in the preceding chapters. We discuss the implications of the presented work and how the research questions were addressed. Furthermore, we highlight the limitations and challenges encountered throughout the thesis research and provide insights into several areas for future investigation.

Chapter 2

Background

“Live as if you were to die tomorrow. Learn as if you were to live forever.”

- Mahatma Gandhi

2.1 Online Social Networks (OSNs)

2.1.1 What are networks?

NETWORKS provide a way of thinking about examining social systems by directing our focus towards understanding the connections between the entities within the system (referred to as actors or nodes). These nodes possess unique attributes, which can include categorical traits (e.g. work, gender, or marital status) or continuous characteristics (e.g. age, salary, or height). Additionally, the links or ties between nodes also possess distinct characteristics. For instance, the relationship between Bill (male, 47 years old) and Jane (female, 43 years old) can be characterised by various aspects, such as marital status, living together, business partnership, shared friendships, and numerous other relational features referred to as ties. These relational attributes can take on various forms, including continuous or ordinal values, such as the duration of their acquaintance or the frequency of disagreements (Borgatti, Everett, and Johnson, 2018).

2.1.2 Definition

A social network can be defined as a social structure comprising nodes that represent individuals or organisations. These nodes are interconnected by various attributes, including but not limited to friendships, shared values, common visions, ideas, business affiliations, and shared interests (Bilge et al., 2009). This intricate web of connections forms the foundation of social networks, allowing for the exchange of information, ideas, and interactions among their members. OSN is a case of a general social network which in turn consists of nodes and edges (the connections). The nodes can be Facebook accounts, Twitter users, Email Accounts or mobile numbers, while for the edges, they can be friendships (Facebook), following (Twitter), emails, messages, mobile calls or SMS.

Social Network Site (SNS), as defined by Boyd and Ellison, is a web-based platform that offers individuals the capacity to A) set up a profile (public/semi-public) within the system, B) create a list of fellow users with whom they have a connection, and C) view and navigate through their connections alongside those curated by other users within the same system. Notably, the characteristics and designations of these connections might exhibit variability from one site to another (Boyd and Ellison, 2007). SNS such as Facebook, MySpace, Twitter and Bebo experienced a surge in user numbers in the mid-2000s. These platforms gained substantial media attention due to their massive expansion, large user base (particularly among the younger demographic), and other issues like the posting of inappropriate content by minors. Also, the potential for SNSs to be exploited in identity fraud (Thelwall, 2009) attracted media attention.

As SNSs continue to spread and develop, the task of precisely defining what characterises an SNS has become increasingly difficult. Some of the attributes that once set them apart have lost their prominence, while other features have been imitated by different types of social media. Platforms that centre on media-sharing, gaming, and location-based media, for example, all motivate users to list their contacts and “Friends” rendering this feature an inadequate criterion for distinguishing SNSs

from other genres of online platforms. Furthermore, other characteristics, such as the media streams on Facebook (i.e., “News Feed”), have emerged as more central aspects of the SNS user experience. Adding to this complexity, open Application Programming Interfaces (APIs) and other platform technologies have enabled numerous third-party websites to build upon SNSs or integrate the social connections of popular SNSs into various tools and platforms (Ellison and Boyd, 2013).

2.1.3 Six Degrees Concept

In 1909, Guglielmo Marconi, the Italian inventor of radio communication, envisioned that technological advancements would ultimately enable contact with any human on the planet through approximately 5.83 connections, although he was referring to a network of radio stations for global communication rather than social connections. This early idea hints at the concept of a closely connected world (Bradley, 2008).

In 1929, the concept of “Six Degrees of Separation” (see Figure 2.1) was proposed by the Hungarian writer Frigyes Karinthy in his story “Chains”, where he suggested that any two individuals could be connected through a chain of no more than five intermediaries (Karinthy, 1929). However, the formal proposition and development of the theory were attributed to the American sociologist Stanley Milgram, who conducted a series of experiments known as the “Small World Experiment” in 1967 (Ma, 2015). He aimed to assess Americans’ connectivity and explore the presence of a separation factor. In these experiments, Milgram asked participants to send a package to a selected target person – located in another part of the United States – by passing the message through a chain of acquaintances (Milgram, 1967). In this study, he introduced the now-famous concept of “Six Degrees of Separation”, suggesting that when considering the interpersonal relationships of individuals, social distances between people are remarkably short despite the vastness of the world (Ma, 2015), and it appears that in a network like the World Wide Web, there are several large hubs (individuals or websites) that possess an extensive number of connections, and serve as critical points on which the connectivity of countless websites and individuals

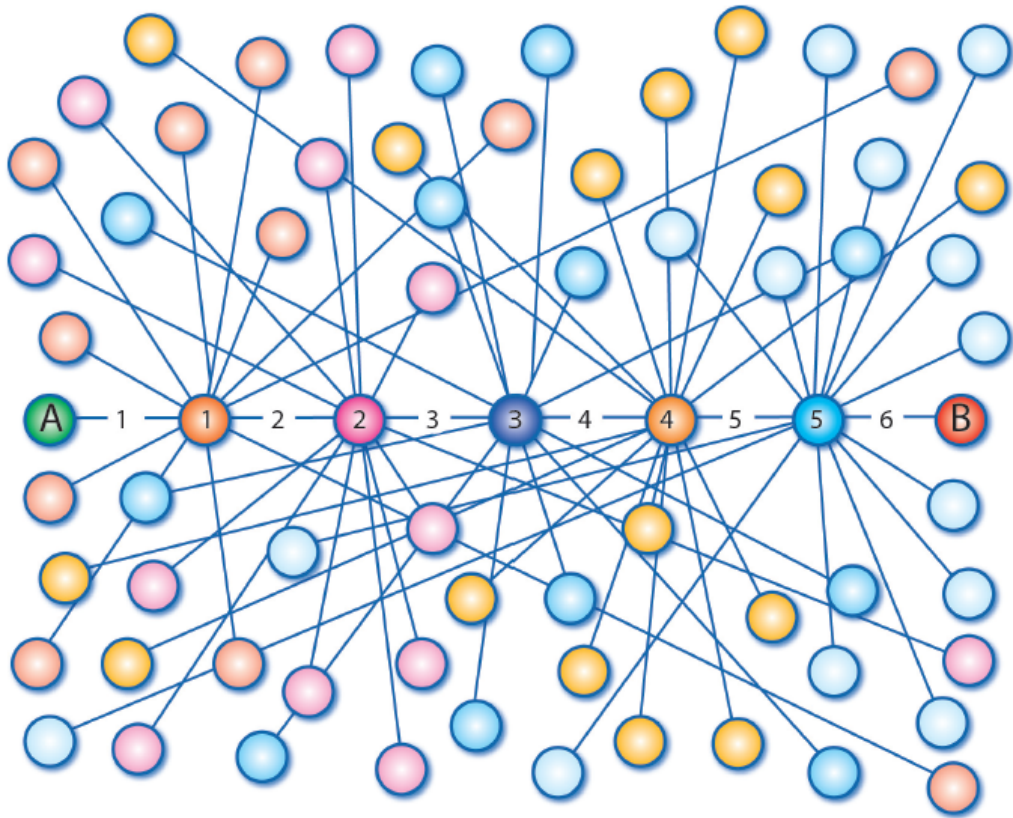


Figure 2.1: Six Degrees of Separation (Ma, 2015)

depends (Bradley, 2008).

In 1994, during an interview with the American actor Kevin Bacon, he said that he “had worked with everybody in Hollywood or someone who has worked with them” (Hedden, 2019). This comment inspired three students at Albright College, Craig Fass, Brian Turtle, and Mike Ginelli, to create a parlour game called “Six Degrees of Kevin Bacon” (Fowler, 2019). This game focused on connecting any actor to Kevin Bacon within six or fewer steps based on their movie roles. Figure 2.2 shows the concept of such a game, and it is available on this website (Reynolds, 1999).

In 2001, three decades after Milgram’s experiment, Duncan Watts and two other researchers re-created a similar experiment by conducting a global online social-search experiment project to prove Six Degrees of Separation and address some limitations found in Milgram’s work. More than 60,000 email users registered online to participate in the experiment. The participants were asked to reach one of the 18

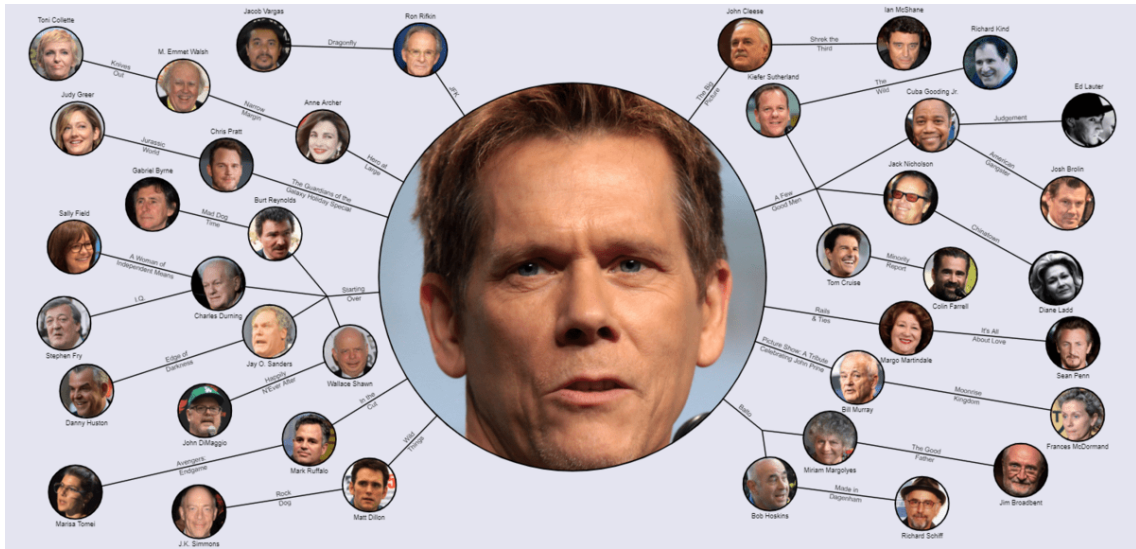


Figure 2.2: Six Degrees of Kevin Bacon (Reynolds, 1999)

target persons located in 13 countries. They were instructed to relay a message to their assigned target by passing it to a social acquaintance they deemed “closer” to the target than themselves. Finally, a total of 384 emails arrived at the destinations through a chain of five to seven people (Dodds, Muhamad, and Duncan J. Watts, 2003). The study revealed that effective social search primarily occurs through intermediate to weak strength connections. Surprisingly, it does not require highly connected individuals (“hubs”) to succeed. Intriguingly, in contrast to unsuccessful social searches, successful ones notably lean towards professional relationships as a key factor in reaching their targets (Dodds, Muhamad, and Duncan J. Watts, 2003).

2.1.4 Genesis & Growth

Andrew Weinreich, often referred to as “the father of social networking” (Warner, 2014; Hines, 2022), was inspired by the vision of Guglielmo Marconi regarding connecting people through a small number of links. In line with this vision, he founded **SixDegrees** in 1997, which was the first OSN (Bedell, 1998). This marked the beginning of a transformative era in digital interaction.

Over the following years, various social networking platforms, including but not

Table 2.1: Market appearance timeline of some selected OSNs from 1997 to 2020

			Year			
AsianAvenue	Bolt		1997	CaringBridge	SixDegree	
Care2	Fotki		1998	Open Diary	Xanga	
AsianAve	Advogato	BlackPlanet	1999	Cyworld	LiveJournal	
DeviantArt	Faceparty	Habbo	2000	LunarStorm	MiGente	Trombi
CozyCot	Cyworld	Jappy	2001	Kwick	Partyflock	Ryze
Fotolog	Friendster	Last.FM	2002	Reunion	Skyblog	StayFriends
Couchsurfing	Hi5	LinkedIn	2003	Multiply	MySpace	Netlog
Nexopia	Tribe.net	WAYN		Xing	Zorpia	
aSmallWorld	Catster	Dodgeball	2004	Dogster	Facebook	Flickr
Hyves	Mixi	Orkut		Piczo	Tagged	
Bebo	Buzznet	Lokalisten	2005	MocoSpace	myYearbook	Ning
Reddit	Renren	Xiaonei		Yahoo!360	YouTube	
CafeMom	Jaiku	MyChurch	2006	Odnoklassniki	Tencent QQ	Tuenti
Twitter	Vkontakte	Wer kennt wen		Windows Live Spaces		
Bahu	Flixster	Fuelmyblog	2007	Platinnetz	Ravelry	Tumblr
Academia.edu		MeinVZ	2008	ResearchGate		
DailyBooth	Foursquare	Sina Weibo	2009	Skoob	WhatsApp	
Audimated	Folkdirect	Friendica	2010	Instagram	Jiebang	Quora
Google+	Pinterest	Snapchat	2011	WeChat	Wellwer	
Cucumbertown	Sgrouples	Spot.IM	2012	Stage 32	Vine	
	Smartican	Spring.me	2013	Telegram		
		Ello	2014	Poolwo		
	Blab	Discord	2015	Periscope		
		TikTok	2017			
		Clubhouse	2020			

limited to Facebook, Twitter, and LinkedIn, have risen in prominence, serving as vital channels for individuals to connect and engage with one another (Heidemann, Klier, and Probst, 2012). In Table 2.1, we listed a timeline of the market appearance of some selected OSNs from the year 1997 to 2020 (Heidemann, Klier, and Probst, 2012; Adewole et al., 2016; Hines, 2022). We can notice that during the years from 2003 to 2006, a lot of OSNs appeared (Boyd and Ellison, 2007), and several ones were quite successful until this day, e.g. LinkedIn (2003), MySpace (2003), Facebook

(2004), Flickr (2004), Reddit (2005), YouTube (2005), and Twitter (2006).

OSNs have evolved into crucial platforms facilitating global communication between people (Adewole et al., 2016). The widespread adoption of social media is evident in the remarkable user growth. As of 2021, nearly half of the world’s population uses social media, reflecting an increase of over a billion users within the preceding five years (Kemp, 2021). A recent report by We Are Social Inc highlights the continued surge in OSN users, with “active users” reaching 4.76 billion in January 2023 (We Are Social Inc., 2023), compared to 4.20 billion in 2021 (Kemp, 2021). This exponential expansion underscores the profound influence of OSNs on contemporary society.

2.1.5 Online Communities

As early as 400 BC, Aristotle described man as a *Zoon Politikon*, a being with an inherent need to seek community and form communities (Martins, 2019). The social networks that Milgram experimented on illuminated that the world is highly clustered, where many of our friends are also friends of each other (Duncan J Watts, 2004). Particularly within the realm of social sciences, the widespread inclination to participate in a community has been a thoroughly examined phenomenon for an extended period (Bagozzi and Dholakia, 2006).

Through the use of platforms such as Twitter, Facebook, LinkedIn, Instagram, and more, billions of individuals are creating online communities and establishing connections with one another (Thakur, Hayajneh, and Tseng, 2019). OSNs facilitate the creation of these online communities comprising individuals who share similar interests, engage in common activities, have comparable backgrounds, or maintain friendships. The majority of OSNs nowadays operate on web platforms, enabling users to create profiles by uploading various content types such as text, images, and videos. Moreover, these networks provide diverse means for users to interact with one another (Schneider et al., 2009). Therefore, there are also content communities whose primary purpose is to facilitate the exchange of media content among users.

These communities cover various types of media, such as text (e.g., BookCrossing), images (e.g., Flickr), videos (e.g., YouTube), and PowerPoint presentations (e.g., SlideShare) (Kaplan and Haenlein, 2010).

The emergence of SNSs represents a transformation in the structure of online communities. While websites centred around communities of shared interests continue to thrive, SNSs are predominantly organised around individuals rather than topics. In the early days of public online communities like Usenet and public discussion forums, the organisation was topic-based, often following hierarchical structures. On the contrary, SNSs adopted a personal (i.e., “ego-centric”) network structure, where the individual serves as the focal point of their own online community (Boyd and Ellison, 2007).

Online communities represent a virtual organisational structure where knowledge collaboration can take place on an extensive scale and across a wide range of subjects, often in novel ways not previously envisioned. One notable aspect is the potential for collaboration among individuals who may be strangers to each other, possess diverse interests, and engage in knowledge sharing without direct communication.

2.1.6 Twitter, a major research platform

Twitter ranks as the third most widely used OSN, following Facebook and Instagram (Antonakaki, Fragopoulou, and Ioannidis, 2021). In contrast to other OSNs, Twitter boasts a straightforward data model and accessible data retrieval API (Application Programming Interface). It has solidified its position as a significant research platform, serving as the focal point of study in over ten thousand research papers within the past decade (Antonakaki, Fragopoulou, and Ioannidis, 2021). These studies encompass but are not limited to, the dynamics of information propagation and its credibility, the exploration of user mobility patterns within the platform, the identification of surges in collective attention towards specific topics or events, and the comprehensive analysis of prevailing trends in public sentiment. Researchers across various disciplines have recognised the immense potential of Twitter data as

a valuable resource for probing these aspects of online communication and societal interaction (González-Bailón et al., 2014). In contrast to user-declared networks like Facebook, Twitter is uniquely focused on the dissemination of information. Twitter users can subscribe to broadcasts from others, making it possible to reconstruct the network of “who listens to whom” by crawling the corresponding “follower graph” (Bakshy et al., 2011).

Twitter integrates features from both SNSs and blogs, incorporating unique characteristics. Similar to SNSs, profiles are interconnected through a defined network structure (Boyd, Golder, and Lotan, 2010). However, these connections are directed rather than undirected; users can choose to link to (i.e., “follow”) others and view their tweets without the obligation for the other user to reciprocate the connection, which is unlike Facebook, where connections are mutual and require confirmation from both parties (Kwak et al., 2010). Nonetheless, the growing interest among researchers in Twitter can be attributed to the platform’s relatively straightforward data accessibility. In contrast to other prominent SNSs such as Facebook, Twitter is inherently public, and its messages can be readily downloaded on a large scale through its API (González-Bailón et al., 2014).

2.2 Social Network Analysis (SNA)

2.2.1 Definition

SNA is an approach for analysing social structures using methods and techniques based on Graph Theory and network measures (Groshek, Mees, and Eschmann, 2020; Tsvetovat and Kouznetsov, 2011). SNA can be characterised as an approach that displays four properties: structural intuition, systematic relational data, graphic imagery, and mathematical and computational models (Freeman, 2004). SNA is not a formal theory within sociology; rather, it is a methodology used to explore and examine “social structures” using concepts and metrics from the graph theory (Wellman and Berkowitz, 1988; Otte and Rousseau, 2002).

SNA focuses on understanding relationships and interactions between individuals or entities in a network, whether it is in a social, organisational, or online context (Wasserman and Faust, 1994). SNA is a greatly adaptable strategy that has come before Twitter and Facebook for at least three decades. In SNA, the emphasis is placed on the connections between nodes (individuals or entities) rather than solely on the attributes of the nodes themselves (Basu, 2014). This approach enables researchers to uncover patterns, information flow, and dynamics within networks. SNA employs various measures, metrics, and visualisation techniques to quantify and visualise relationships, helping researchers gain insights into the structure, functioning, and influence within networks (Wasserman and Faust, 1994).

While SNA itself is not a formal theory, it is a valuable strategy that can be applied across various disciplines, including sociology, anthropology, psychology, business, and even computer science. It provides a lens through which researchers can study the complex interplay of connections and relationships that shape social systems. Researchers, advertisers and political activists see enormous social networks as a representation of interactions that can be used to analyse the propagation of ideas, thoughts, social bond dynamics and viral marketing among individuals (Huberman, Romero, and Wu, 2008).

In a few words, SNA can be considered as a study of relationships between individuals using the graph theory (Tsvetovat and Kouznetsov, 2011). Unlike other analytical methods where the focus is on individual behaviour, SNA looks at the interaction between those individuals, which is an important attribute of SNA. The analysis on the network level -instead of the node level- gives researchers the opportunity to study how networks' structures affect the way individuals, groups, organisations, and systems work and interact inside such networks. It characterises networks in terms of nodes and edges. Depending on the application, the nodes can be individual actors, OSN accounts, etc, while the edges (links) can be relationships or interactions that connect these nodes together.

2.2.2 Power and Influence

Influence – as defined by Oxford dictionary¹ – “*is the capacity to have an effect on the character, development, or behaviour of someone or something, or the effect itself*”. When you influence someone, that means to affect or change how someone (or something) develops, behaves, or thinks according to Cambridge University².

Maxwell says “Leadership is influence, nothing less, nothing more, nothing else.” (Maxwell, 2019). All sociologists would concur that power is a key property of social structures. There is much less agreement about its definition, description and how we can analyse its causes and outcomes (Hanneman and Riddle, 2005).

As for SNA, power and influence can be measured by studying the social structures, which means they depend on the structure’s shape. There are different kinds of social networks, and accordingly, there are variations in the concept and how power is computed (Pineiro, 2011). The different structures we see in social networks rely basically on the underlying connections between nodes. Thus, the measure of power is strongly correlated to the relationships between nodes within the social network. A particular member (node) of the social network (structure) has power exclusively as a result of its connections inside the network (Pineiro, 2011).

2.2.3 General SNA Metrics

2.2.3.1 Graph Density

Graph density refers to the extent to which a graph’s edges (connections) are realised from all possible edges that could exist in a complete graph of the same number of nodes. In simple terms, it measures how interconnected the nodes of a graph are (West, 2001).

Mathematically, the density D of graph G can be calculated using Formula (2.1).

$$D(G) = \frac{2E}{V(V-1)} \quad (2.1)$$

¹<https://en.oxforddictionaries.com/definition/influence>

²<https://dictionary.cambridge.org/dictionary/english/influence>

where E is the number of edges and V is number of nodes. The value of graph density ranges from 0 (corresponds to no edges) to 1 (when all possible edges are present). Obviously, a higher density indicates a more interconnected graph.

2.2.3.2 Graph Diameter

The graph diameter is a measure used in graph theory to quantify the longest distance between any two nodes (vertices) within a graph. In the context of a network, it represents the greatest number of edges that must be traversed to connect any pair of nodes. In other words, it is the longest path one would need to take to move from one node to another while covering the fewest number of edges (Wasserman and Faust, 1994).

For example, in a social network, the graph diameter might represent the maximum number of friendships or connections one would need to traverse to reach any two people in the network. It provides insights into the overall efficiency of communication, information flow, or influence propagation across the network.

2.2.3.3 Average Path Length

The average path length is a key metric in network analysis, particularly within graph theory. It refers to the average number of steps or edges it takes to travel between any two nodes in a network, considering all possible pairs of nodes. In simpler terms, it measures the average distance between nodes in terms of the minimum number of edges that need to be traversed to connect them. To calculate the average path length, we need to find the shortest path between every possible pair of nodes in the network, sum up the lengths of these paths, and then divide by the total number of pairs. This metric is valuable for understanding the overall efficiency of information or influence transfer within a network (Newman, 2018).

A smaller average path length implies that the network is more connected and efficient in terms of information flow. Nodes are generally closer to each other,

making it easier to transmit information or influence. In contrast, a larger average path length indicates that the network might be less efficient in terms of direct connections between nodes. The average path length complements the concept of graph diameter. While graph diameter gives the longest “shortest path” in the network, the average path length provides an overall picture of the average distance between nodes. Both metrics help in assessing the network’s structure, efficiency, and potential for information propagation (Newman, 2018).

2.2.4 Network Centrality

The basic idea behind the centrality is to find the most “important” or the “central” node (individual) in a given network. Centrality is vital since it reflects who has the critical position in the network. Central positions usually correspond with opinion leadership or celebrities, which are related to adoption behaviours.

There are several types of centrality. Each centrality is calculated differently and reflects a certain role of importance. We will discuss the most-used ones in the literature: Degree, Closeness, Betweenness, Eigenvector and PageRank.

2.2.4.1 Degree Centrality

Degree centrality (C_D) is the degree of a node in a graph, which means the count of links that node has (Tsvetovat and Kouznetsov, 2011). In the OSN context and taking Facebook as an example, the C_D reflects the number of friends, but for Twitter, it is a little bit different. The “Friendship” relation on Facebook is a symmetric connection, while on Twitter, we have the “Following” relation, which is a one-way arrow from “Follower” and thus, we need to consider the direction of the connection, i.e., dealing with directed graphs and thus we can distinguish between:

- In-Degree Centrality (C_{ID}): which corresponds to the in-degree a node has in a directed graph. This is the number of people who follow you on Twitter.
- Out-Degree Centrality (C_{OD}): which corresponds to the out-degree a node has

in a directed graph. This is the number of people you follow on Twitter.

If $G = (V, E)$ is a graph, and $A = (a_{ij})$ is its adjacency matrix, and $deg_G(v)$ is the degree of node v in G then the degree centrality is given in Formula (2.2), according to (Newman, 2008).

$$C_D(v) = deg_G(v) = \sum_{i=1}^n (a_{iv}) \quad (2.2)$$

where n is the total number of nodes. The centrality scores can be normalised by dividing them by $n - 1$.

In-degree and out-degree centralities can be calculated using Formula (2.3) and Formula (2.4), (Kiss, Scholz, and Bichler, 2006).

$$C_{ID}(v) = deg_G^-(v) = \sum_{i=1}^n (a_{iv}) \quad (2.3)$$

$$C_{OD}(v) = deg_G^+(v) = \sum_{i=1}^n (a_{vi}) \quad (2.4)$$

2.2.4.2 Closeness Centrality

Closeness centrality C_C highlights the distance of a node to all other ones in a graph by calculating the path from each node to all others (Hanneman and Riddle, 2005). C_C measures the mean geodesic from a node to others in the graph, and it increases when these distances decrease.

If $G = (V, E)$ is a graph, and $d_G(v, t)$ is the length of a geodesic path from v to t , meaning the count of ties along that path. The sum of all paths from v to all other nodes is called “Farness”, and it is not a centrality, but it reflects how far a node is from all others in the network, and thus the C_C can be given as an inverse of “Farness” in Formula (2.5), (Sabidussi, 1966).

$$C_C(v) = \frac{1}{\sum_{t \neq v}^n (d_G(v, t))} \quad (2.5)$$

2.2.4.3 Betweenness Centrality

Introduced in 1977 by (Freeman, 1977), Betweenness centrality C_B is a network centrality measure that quantifies the significance of a node in facilitating connections between other nodes within the network. It calculates the extent to which a node falls on paths between other nodes in the graph. (Freeman, 1978) presented it as a measure for evaluating the control of an individual over the information flow between others in a graph.

If $G = (V, E)$ is a graph, σ_{st} is the count of the shortest paths from s to t , and $\sigma_{st}(v)$ is the count of these paths which contains v , thus the C_B will be given in Formula (2.6), according to (Freeman, 1978).

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2.6)$$

2.2.4.4 Eigenvector Centrality

While C_D simply counts the edges a node has, Eigenvector centrality C_E acknowledges that not all these edges are the same (Newman, 2008). The eigenvector approach is an effort to locate the most central individuals (i.e. those nodes with the smallest farness from other ones) regarding the “global” network’s structure and to focus less on the “local” patterns (Hanneman and Riddle, 2005). A node which has a high score of C_E is connected to many nodes which are themselves well-connected. A variant of C_E called “PageRank” was used by “Google” to rank webpages (Newman, 2008).

If $G = (V, E)$ is a graph, and $A = (a_{ij})$ is its adjacency matrix, then the eigenvector centrality is given in Formula (2.7), according to (Newman, 2008).

$$C_E(i) = x_i = \frac{1}{\lambda} \sum_{j=1}^n a_{ij} x_j \quad (2.7)$$

where λ is a constant. Writing the centralities’ vector as $x = (x_1, x_2, \dots)$, then Equation (2.7) can be rewritten in a matrix form to be Equation (2.8), according

to (Newman, 2008), so X is an eigenvector of A with λ as eigenvalue.

$$\lambda X = A \cdot X \tag{2.8}$$

2.2.4.5 PageRank Centrality

PageRank centrality (PR) is the algorithm that is used by the Google search engine to rank webpages. PR considers the count and quality of incoming links to a page to estimate how important the webpage is. In other words, the rank of a page is given by the rank of those pages that have links to it. Then, their ranks can be given by the ranks of pages with links to them, and so on. Therefore, the PR of a page is always calculated recursively by the ranks of other pages. For that reason, PR is a variant of eigenvector centrality (Brin and Page, 1998).

PR can be applied to any graph as long as it is directed. If $G = (V, E)$ is a graph, then PR centrality for node v is given in Formula (2.9), (Brin and Page, 1998).

$$C_{PR}(i) = x_i = (1 - d) + d \sum_{j \in M(i)} \frac{x_j}{L(j)} \tag{2.9}$$

where d is a damping factor, $M(i)$ is a set which contains all the nodes that have a connection to i , and $L(j)$ is the count of connections that j has.

2.2.4.6 Centrality Measures Comparison

To compare the different centrality measures presented earlier, the network in Figure 2.3 was used to calculate the centrality scores of all nodes using the NodeXL plugin for MS-EXCEL (Smith et al., 2010). The results are shown in Table 2.2.

For **degree** C_D , nodes B , D , F , and G are the strongest because each one of them has the highest score of C_D and that is simply because they have more connections than other nodes, while nodes A and H hold the lowest value because each one of them has only one connection. The noticeable drawback of C_D is that only the direct links are considered, unlike other centralities.

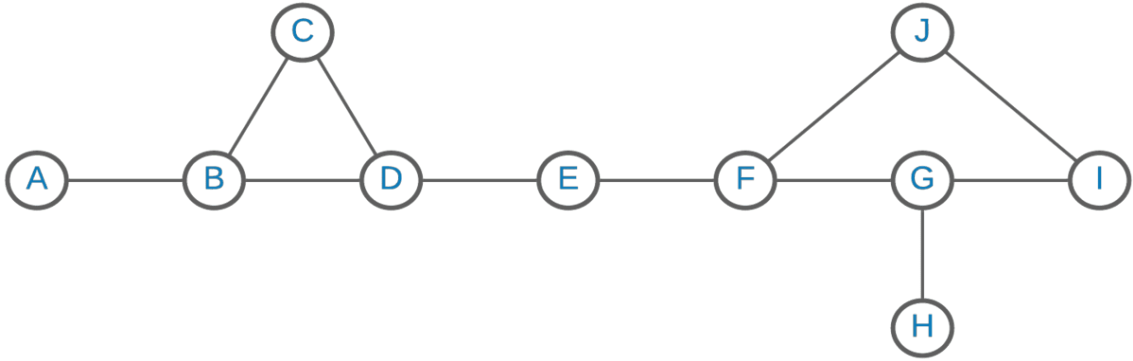


Figure 2.3: Network example to calculate centrality scores

Table 2.2: Comparison of centrality measures

Centrality	A	B	C	D	E	F	G	H	I	J
Degree	1	3	2	3	2	3	3	1	2	2
Rank (D)	9	1	5	1	5	1	1	9	5	5
Closeness	0.029	0.038	0.037	0.048	0.053	0.053	0.043	0.032	0.034	0.04
Rank (C)	10	6	7	3	1	1	4	9	8	5
Betweenness	0	8	0	18	20	21	11	0	1	3
Rank (B)	8	5	8	3	2	1	4	8	7	6
Eigenvector	0.057	0.138	0.121	0.155	0.114	0.121	0.098	0.04	0.074	0.081
Rank (E)	9	2	3	1	5	3	6	10	8	7
PageRank	0.532	1.349	0.898	1.29	0.884	1.301	1.369	0.538	0.926	0.912
Rank (PR)	10	2	7	4	8	3	1	9	5	6

For **closeness** C_C , nodes E and F are the strongest because each one of them has the highest score since they are positioned in the middle, which makes them closer to all other nodes. A has the lowest score because it is far from all other nodes. The C_C relies only on the location of a node inside the network.

For **betweenness** C_B , node F has the highest score because it is positioned in the middle of the network, which makes it a member of all the shortest paths that connect any two nodes. Nodes A , C , and H are the worst ones in terms of C_B because none of them is a member of any shortest paths that connect any other pair

of nodes. C_B depends on how well situated the node is between other ones.

For **eigenvector** C_E , node D is the best for two reasons: 1) it has the highest score of C_D and 2) it is connected to another strong node, B . Node H has the lowest score because it has only one connection with G , which is also a weak node.

2.2.4.7 Which centrality measures to use?

The measurement of a person's centrality, power, prestige, or influence within a network can take various forms, and the choice of the most suitable measure often hinges on the specific context and objectives of the analysis. Different network centrality measures offer unique insights and interpretations, making their selection dependent on the particular application (Bloch, Jackson, and Tebaldi, 2023).

Here are some common use cases for the centrality we presented earlier. Degree centrality is often applied in social networks to find the most popular individuals, while closeness centrality is suitable for analysing information diffusion. Betweenness centrality is useful for studying the flow of information or influence. Eigenvector centrality is useful for identifying nodes with indirect influence, as it extends the degree centrality concept by considering not only a node's immediate connections but also the influence of their connections, and this influence cascades through the entire network. Finally, PageRank centrality can also be used to find influential nodes. It is similar to Eigenvector centrality but in a directed graph and it considers the weights of connections as well (Disney, 2020).

The choice of which centrality measure to use should be aligned with the research question and the specific characteristics of the network subject to analysis. A combination of centrality measures can also provide a more comprehensive understanding of network structure and dynamics.

2.3 Taxonomies

2.3.1 Definition and Use

The term “taxonomy” originates from the fusion of two Greek roots, “taxis”, denoting arrangement or order, and “nomia”, signifying distribution or method. This etymology encapsulates the essence of taxonomy as a systematic approach to organising and categorising information or entities within various disciplines. According to the Cambridge dictionary, a taxonomy is “*a system for naming and organizing things, especially plants and animals, into groups that share similar qualities*”³. The glossary of the National Institute of Standards and Technology (NIST) defined taxonomy as a “*scheme of classification*”⁴.

Taxonomies succinctly encapsulate knowledge about a specific domain, fostering a common understanding among peers. Researchers utilise taxonomies to communicate information about a particular field of knowledge or to aid in automation tasks. Practitioners employ them to facilitate communication beyond organisational boundaries (Unterkalmsteiner and Adbeen, 2023).

2.3.2 The importance of Classification

The inherent tendency to classify, known as the taxonomic impulse, plays a pivotal role in how humans perceive and comprehend the world around them. This innate inclination towards classification extends across diverse realms of human activity, serving as a foundational element in our cognitive processes. Taxonomies, in essence, act as cognitive lenses through which we organise, interpret, and communicate our experiences and observations of the world (Lambe, 2014). They provide structure and coherence to our understanding, allowing us to navigate the complexities of our environment and articulate our insights effectively. Classification “is almost the methodological equivalent of electricity, we use it every day, yet often consider it to

³<https://dictionary.cambridge.org/dictionary/english/taxonomy>

⁴<https://csrc.nist.gov/glossary/term/taxonomy>

be rather mysterious” (Bailey, 1994).

Simply, classification is ordering entities into classes based on their similarity. Knowledge classification has significantly facilitated the advancement of various knowledge domains, primarily through four key mechanisms (Usman et al., 2017). 1) Classification of the elements within a knowledge domain establishes a shared vocabulary (unified terminology), thereby facilitating the exchange and dissemination of knowledge among researchers and practitioners. 2) Classification aids in elucidating the interconnections among the elements within a knowledge domain, which enhances the comprehension of the subject matter. 3) By systematically organising information, classification methodologies can highlight areas within a knowledge domain that are underdeveloped or lacking in research. 4) Classification contributes to informed decision-making processes by providing structured insights into the relationships and attributes of the elements within a knowledge domain. In essence, classification serves as a valuable tool for researchers and practitioners alike, facilitating the generalisation, communication, and practical application of knowledge gathered from a given field.

2.3.3 Structure

Lambe described various representations of taxonomies such as lists, trees, hierarchies, poly-hierarchies, matrices, facets and system maps (Lambe, 2014), while Kwasnik outlined four primary methods for structuring a classification scheme: hierarchy, tree, paradigm, and faceted analysis (Kwasnik, 1999).

1. **Hierarchy:** This approach results in taxonomies with a single overarching class that encompasses all subordinate classes in a hierarchical relationship, characterised by inheritance, i.e., “is-a” relationship (Kwasnik, 1999). See Figure 2.4 for an example of a hierarchical taxonomy for vehicles.
2. **Tree:** While similar to the hierarchical structure, tree-based taxonomies do not have inheritance relationships among classes. Instead, typical relationships in-

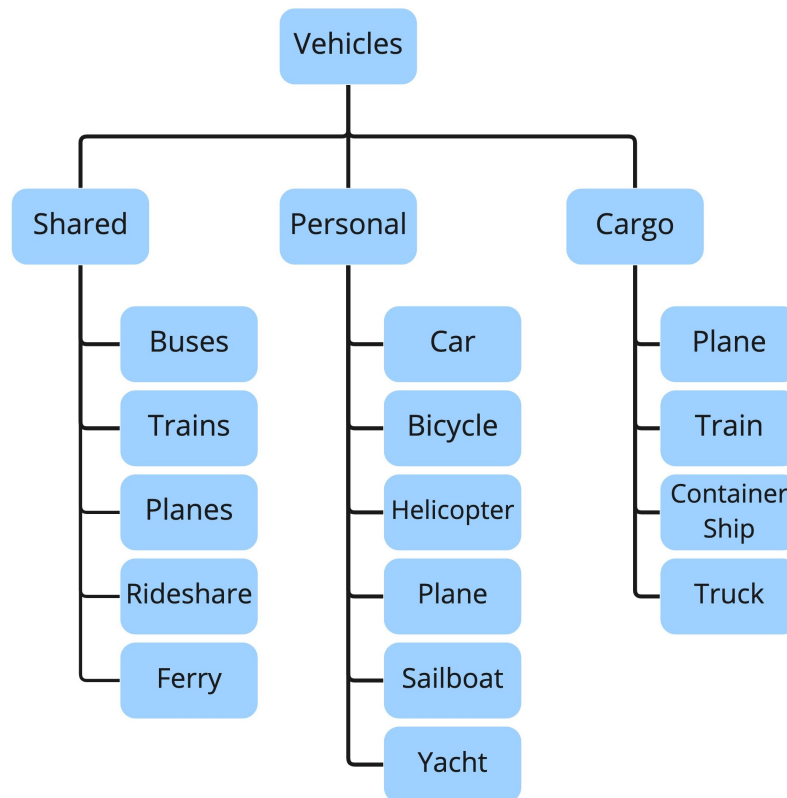


Figure 2.4: Hierarchical Taxonomy, example (Laubheimer, 2022)

clude “part-whole”, “cause-effect” and “process-product” (Kwasnik, 1999).

3. **Paradigm:** This method creates taxonomies with bidirectional hierarchical relationships between classes. Each class is defined by a combination of two attributes at a time (Kwasnik, 1999).
4. **Faceted Analysis** This approach classifies subjects from multiple perspectives or facets (Laubheimer, 2022). Each facet represents an independent viewpoint with its own set of classes, allowing for flexible and adaptive taxonomies that can evolve over time (Prieto-Díaz, 1991; Kwasnik, 1999). See Figure 2.5 for an example of a faceted taxonomy for vehicles.

A hierarchical taxonomy of vehicle types (Figure 2.4) uses a single organising principle, such as categorising vehicles as shared, personal, or cargo. In contrast, a

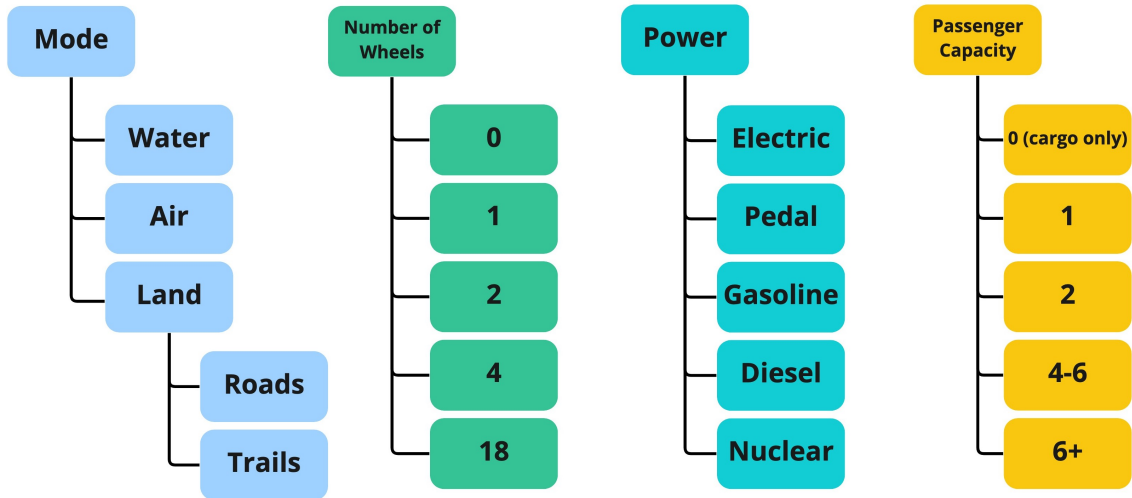


Figure 2.5: Faceted Taxonomy, example (Laubheimer, 2022)

faceted taxonomy (Figure 2.5) employs a distinct small hierarchy for each facet or attribute, enabling detailed combinations of characteristics.

2.3.4 Terminology

While there are various ways to structure a taxonomy, the most common method - and the one we are interested in for our research - is a hierarchy. Figure 2.6 depicts a hierarchical structure using a tree diagram. A tree consists of nodes connected by edges, where the edges represent the parent-child relationships within the hierarchy. We can distinguish three types of nodes in a tree.

1. **Root node:** This is the single root node from which all other nodes descend.
2. **Intermediate nodes:** Positioned between the root node and the leaf nodes, these nodes are referred to as categories in the context of taxonomy constructs.
3. **Leaf nodes:** These are nodes without any children, referred to as characteristics in the context of taxonomy constructs.

Each node, except the root node, has exactly one parent. Additionally, each node has a depth, defined as the number of edges from the node to the tree's root node. Thus, the root node has a depth of 0.

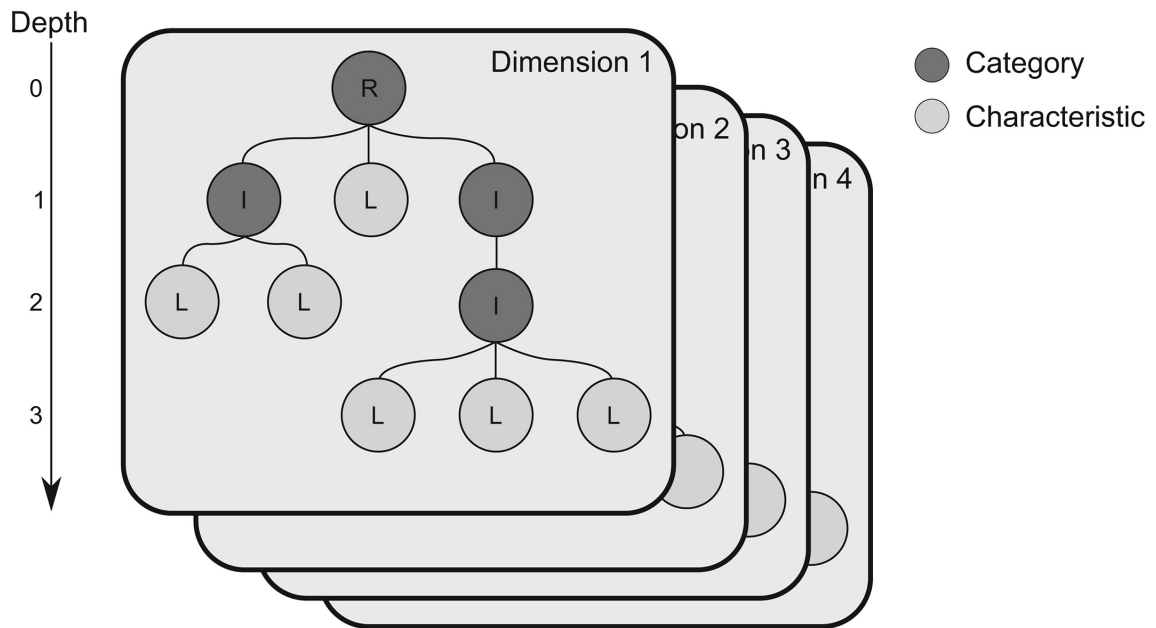


Figure 2.6: Taxonomy components, example (UnterkaImsteiner and Adbeen, 2023)

2.3.5 Components

Taxonomy components are fundamental elements that structure the classification systems used in various fields. These components ensure that taxonomies effectively capture and communicate knowledge about a specific domain. Here are the primary components of taxonomies.

- **Dimensions** refer to the broadest categories or aspects of the domain that the taxonomy aims to organise. They represent different perspectives or angles from which the domain can be understood. For example, in cyber security, dimensions could include threat types, attack vectors, and defence mechanisms.
- **Categories** are more specific groupings within each dimension. They represent distinct classes or clusters of entities that share common characteristics. Categories help break down dimensions into manageable and meaningful parts.
- **Characteristics** are the specific attributes or properties used to define and differentiate the entities within each category. They provide the criteria for inclusion in a particular category and help in distinguishing one category from another.

- **Relationships** describe how different categories or entities are connected or related to one another within the taxonomy. They can include hierarchical relationships (e.g., parent-child), associative relationships (e.g., linked entities), and other forms of connections.
- **Levels of Abstraction.** Taxonomies often include multiple levels of abstraction, ranging from broad generalisations to specific details. These levels help users navigate the taxonomy and find the appropriate level of detail for their needs.
- The **scope and boundaries** define the limits of what the taxonomy covers. They clarify what is included and excluded from the taxonomy, ensuring it remains focused and relevant.

These components collectively form the backbone of a well-structured taxonomy, enabling it to serve as a powerful tool for organising and communicating knowledge across various domains.

2.3.6 Development Process

When developing taxonomies, it is important to recognise that there is not a single, universally correct taxonomy for any domain. Instead, some taxonomies may be more fitting or informative in specific contexts. As explained by (Nai Fovino et al., 2019), the traditional process of creating a taxonomy involves several well-defined steps as follows (See Figure 2.7).

1. **Define Subject Scope:** Identify the purpose and scope of the taxonomy.
2. **Identify Sources:** Select widely recognised sources, including standards, activities from international working groups, and scientific literature.
3. **Collect Terms and Concepts:** Analyse sources to extract relevant concepts, sub-domains, and terminology.
4. **Group Similar Concepts Together:** Cluster related concepts.

5. **Add Term Relationships and Details:** Identify commonalities and simplify the taxonomy structure. Create a glossary using definitions from international standards or scientific references where available.

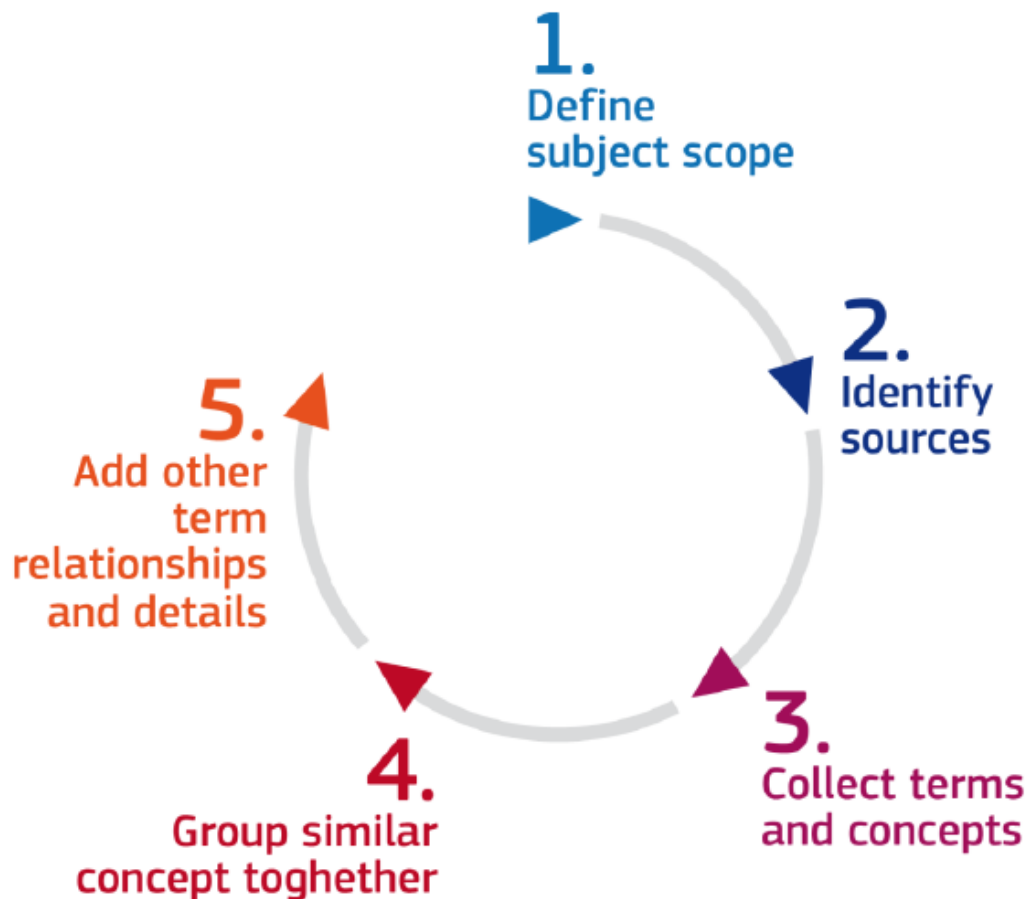


Figure 2.7: Taxonomy Building Traditional Approach (Nai Fovino et al., 2019)

2.4 Machine Learning (ML)

2.4.1 Definition

The task of identifying patterns in data is fundamental and has a rich and successful history. In the 16th century, the comprehensive astronomical observations of Tycho Brahe paved the way for Johannes Kepler to derive the empirical laws governing

planetary motion. This, in turn, served as a critical foundation for the advancement of classical mechanics (Bishop, 2006). The science of “learning” is undeniably pivotal in various fields, including statistics, data mining, and Artificial Intelligence (AI). It represents a multifaceted domain that often intersects with various other disciplines, including engineering and beyond (Trevor, Robert, and Jerome, 2009).

ML is a prominent branch of AI with the primary objective of leveraging data to enhance performance across a range of tasks, including prediction (Mitchell, 1997). ML provides a set of powerful tools to address challenges that traditional statistical methods may struggle with (T. Jiang, Gradus, and Rosellini, 2020). The type of challenges solved by ML surpasses the capabilities of fixed programs created and designed by humans (Mohri, Rostamizadeh, and Talwalkar, 2018). It has become a vital field as it enables computers to learn and adapt without being explicitly programmed for every specific task. This capability has opened up many applications across various domains, from healthcare and finance to NLP and image recognition.

ML is a computational approach that utilises past information, termed “experience” to enhance performance and achieve precise predictions. It involves developing efficient and accurate prediction algorithms (Mohri, Rostamizadeh, and Talwalkar, 2018). Like other computer science domains, evaluating these algorithms includes assessing their time and space complexity. However, in ML, we introduce the concept of sample complexity to determine the sample size necessary for the algorithm to grasp a set of concepts. Given that the algorithm’s success hinges on the utilised data, ML is intrinsically linked with data analysis and statistics. In a broader sense, learning techniques are data-driven methods that integrate fundamental computer science principles with insights from statistics, probability, and optimisation (Mohri, Rostamizadeh, and Talwalkar, 2018).

2.4.2 Learning Types

ML algorithms can be divided into categories: supervised learning, unsupervised learning and reinforcement learning, as depicted in Figure 2.8. The main tasks

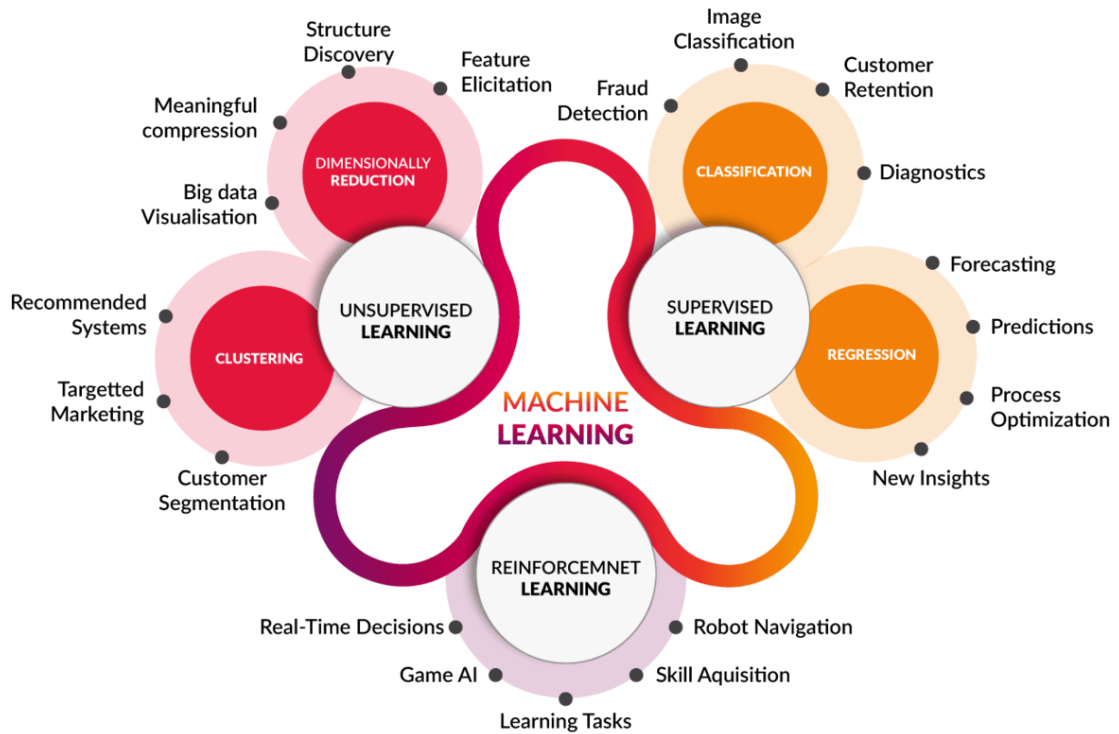


Figure 2.8: Machine Learning Types (CogNub, 2018)

under each category and some common applications were also mentioned.

Supervised Learning (SL) (or supervised machine learning) involves training a model on labelled data to make predictions or decisions based on input variables. In this approach, a dataset is provided to the model, consisting of input features (also known as independent variables) and corresponding output labels (also known as dependent variables). The goal in SL is to train the model to learn the underlying patterns and relationships in the data, allowing it to generalise and make accurate predictions on new unseen data (James et al., 2023). It is often used for tasks like classification and regression.

Unsupervised Learning presents a different scenario compared to SL, where the focus shifts from predicting a response variable to uncovering patterns, structures, or relationships within a dataset where there are no associated response labels, i.e., unlabelled dataset (James et al., 2023). Common applications include clustering and dimensionality reduction.

Reinforcement Learning involves the process of learning how to effectively associate situations with actions to optimise a numerical reward signal. The learner is not provided explicit instructions on which actions to take but must ascertain which actions result in the highest rewards through experimentation. In complex scenarios, actions can influence not only immediate rewards but also subsequent situations and, consequently, all future rewards. These defining aspects, trial-and-error exploration and delayed-reward impact, are fundamental in distinguishing reinforcement learning as a paradigm (Sutton and Barto, 2018). It is often used in areas like robotics and gaming.

The research conducted in this thesis primarily involves SL, as discussed in Chapters 5 and 6. Consequently, the upcoming sections will focus only on the principles, techniques, and models relevant to classification tasks from SL.

2.4.3 Statistical Inference & Machine Learning

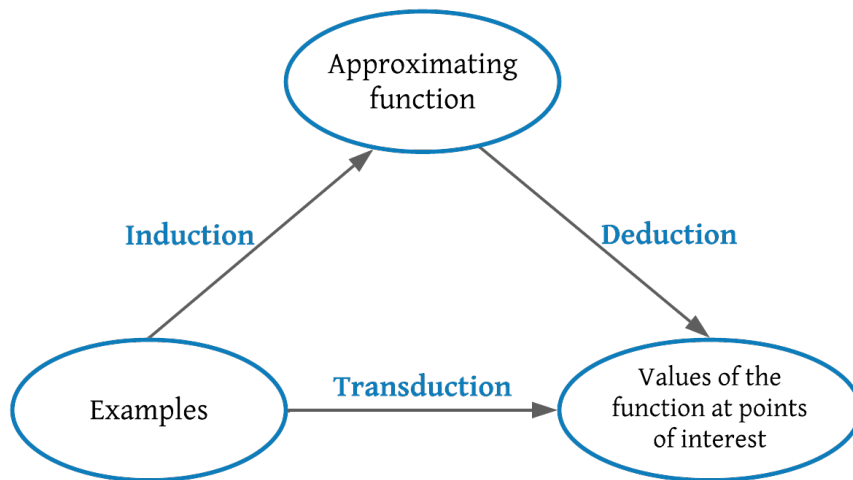


Figure 2.9: Statistical inference types and their relations (V. N. Vapnik, 1999)

In statistics, we can distinguish three types of inference as follows: A) **Induction** inference, which involves deriving a function or a model from given data. B) **Deduction** inference, on the other hand, involves deriving specific values or conclusions from a given function or model. C) **Transduction** inference which is a

more specialised form of inference that involves deriving the values of an unknown function for specific points of interest using the given data. It combines aspects of both induction and deduction (V. N. Vapnik, 1999). These inference types and the relationships between them are depicted in Figure 2.9.

The aforementioned types of statistical inference capture different phases of the ML process: **induction** involves the learning phase when models generalise from data, where **deduction** occurs after learning when models are applied to new data, and **transduction** is related to cases where predictions are made for specific examples (Brownlee, 2019).

2.4.4 Supervised Learning Overview

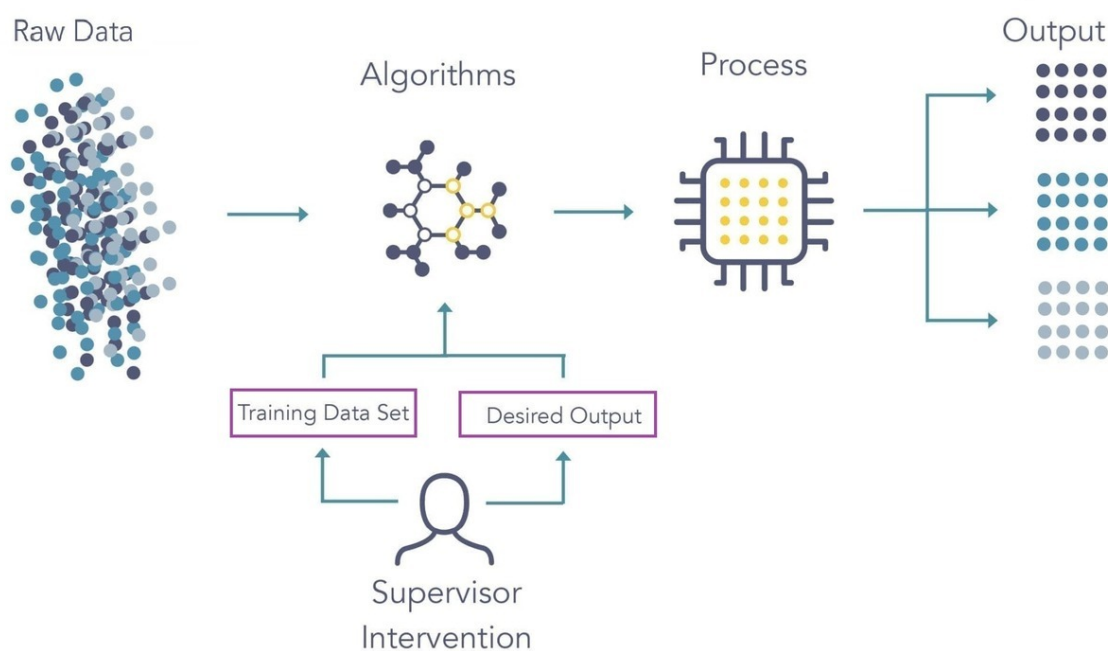


Figure 2.10: The supervised machine learning process (Grieve, 2020)

The classical approach often suggests a two-step process for deriving values for points of interest: first, use *induction* to learn a general function from data, and then use *deduction* to apply the learned function to specific cases, rather than getting the values in one step (V. N. Vapnik, 1999). This approach is common in many ML

workflows, where models are trained on historical data and then used for real-world predictions (Mohri, Rostamizadeh, and Talwalkar, 2018).

The process of SL, as depicted in Figure 2.10, commences with the induction phase by creating a labelled dataset through the annotation of the training dataset, a task typically performed by a human supervisor. These labels correspond to the desired output classes. Subsequently, an ML algorithm is chosen and trained using the labelled training dataset. Throughout this training phase, the model acquires the ability to map input feature values to their corresponding output values or classes. Upon the completion of training and the generation of the model, the deduction phase is initiated. In this phase, the model is applied to predict outcomes using new and previously unseen data sources.

These primary steps constitute the core of the SL process, which is used to build predictive models for various applications, from image recognition to fraud detection and beyond. However, each primary step includes further sub-steps, which will be described in detail in Chapter 5.

2.4.5 Classification vs Regression

In ML, variables can be classified into two main types: quantitative and qualitative (which is also known as categorical). Quantitative variables are those that are expressed in numerical values. They encompass attributes such as age, height, income, the value of a house, or the price of a stock. In contrast, qualitative variables assume values from a set of distinct classes or categories, denoted as K . Examples of qualitative variables include marital status (married or not), the brand of a purchased product (brand A, B, or C), a person's default status on a debt (yes or no), or a medical diagnosis (like Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia). Problems involving quantitative responses are typically referred to as “*regression*” problems, while those dealing with qualitative responses are often termed “*classification*” problems (James et al., 2023).

Figure 2.11 shows two example models for classification and regression problems.

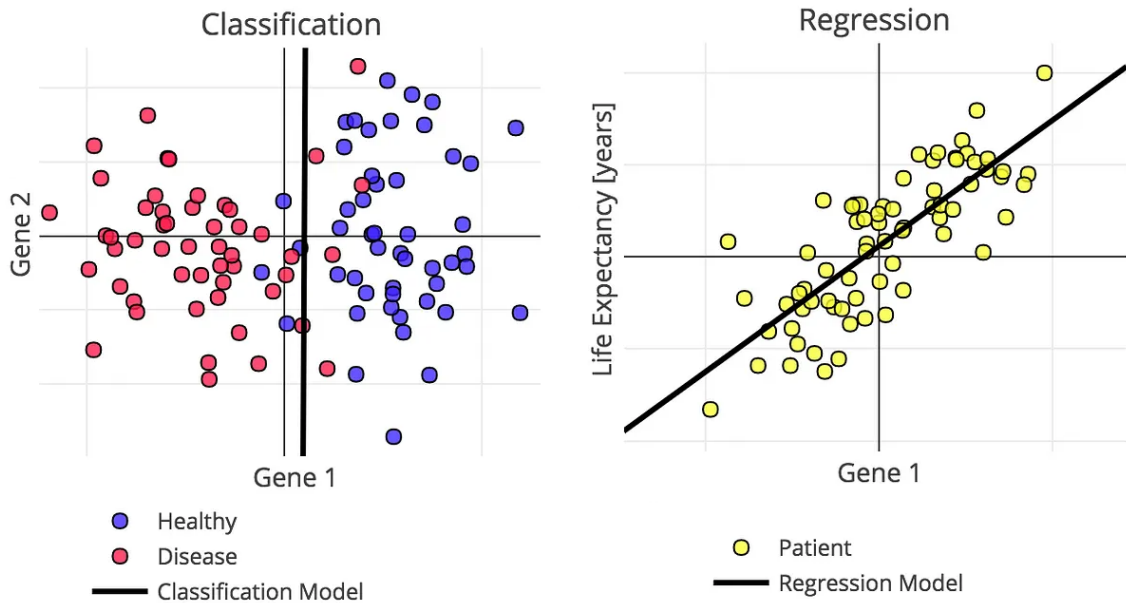


Figure 2.11: SL, Classification & Regression (Krzyk, 2018)

We can notice in the classification case that the output variable has discrete values (Healthy, Disease), whereas in the regression case, the output variable has infinite values (number of years) (Krzyk, 2018).

2.4.6 Supervised Learning Algorithms

SL field encompasses a range of algorithms that are employed for classification tasks. Classification involves assigning input data points to predefined categories or classes. Below are some of the algorithms used for classification tasks presented in this thesis.

2.4.6.1 Logistic Regression (LR)

Despite the naming, Logistic Regression (LR) is primarily a classification model rather than a regression one. It serves as a straightforward and highly efficient solution for binary and linear classification tasks. This model, known for its simplicity and effectiveness, excels when dealing with linearly separable classes. LR is widely adopted in industrial applications, where it has proven to be a valuable algorithm for classification. Furthermore, the LR model (such as Adaline and perceptron) is a

statistical approach primarily designed for binary classification, with the capability to extend its utility to multi-class classification scenarios (Subasi, 2020). Figure 2.12 shows an example of a logistic regression classifier where only one independent variable was considered.

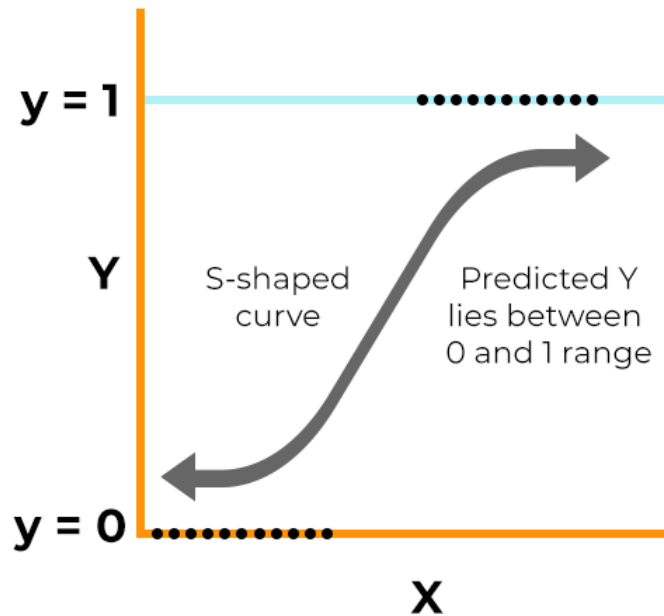


Figure 2.12: SL, Logistic Regression classifier (Kanade, 2022)

The primary idea behind LR is to model the probability that a given input belongs to a particular class. It accomplishes this by using a sigmoid function, also known as an S-shaped curve, to transform predictions into probabilities by converting real-valued inputs into a bounded range spanning from 0 to 1 (Kanade, 2022). When the sigmoid function produces an estimated probability greater than a predefined threshold, the model assigns the instance to a particular class. Conversely, if the estimated probability falls below the threshold, the model predicts that the instance belongs to the other class. During the training process, LR optimises its sigmoid function using techniques such as maximum likelihood estimation. Once trained, the model can be used to predict the probability of a data point belonging to a particular class (Kanade, 2022).

2.4.6.2 Support Vector Machines (SVM)

Support Vector Machine (SVM) stands as an effective ML approach, originating from the foundations of statistical learning theory, as introduced by Vapnik in 1995 (Cortes and V. Vapnik, 1995). The fundamental principle behind SVM is to find an optimal hyperplane that maximises the margin of separation between distinct classes within a dataset. In binary classification, this hyperplane aims to create a clear boundary between two classes by maximising the distance between the nearest data points from each class to the hyperplane (Cortes and V. Vapnik, 1995). These nearest data points, known as support vectors, significantly influence the determination of the hyperplane.

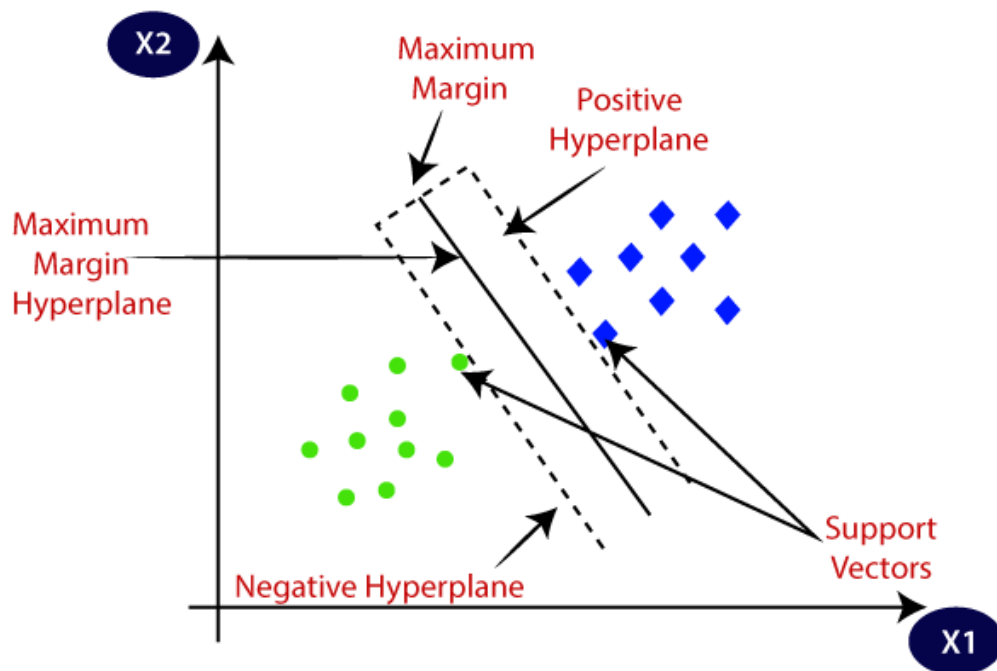


Figure 2.13: SL, Support Vector Machine classifier (Saini, 2023)

Mathematically, SVM seeks to solve an optimisation problem that involves finding the weight vector and bias term for the hyperplane. The objective is to maximise the margin while minimising the classification error. This process can be linear or nonlinear, depending on the choice of the kernel function used in the SVM (Bishop, 2006). Common kernel functions include linear, polynomial, Radial Basis Func-

tion (RBF), and sigmoid. SVM models work exceptionally well in high-dimensional spaces, making them suitable for a wide range of applications, including text classification, image recognition, and bioinformatics. One of the strengths of SVMs is their ability to handle non-linear data by using a mathematical technique called the kernel trick. This allows SVMs to implicitly map data into a higher-dimensional space, where a linear hyperplane can separate non-linearly separable classes.

2.4.6.3 Decision Trees (DTs)

Decision Tree (DT) provides a structured tree-like representation of decision rules. DTs recursively split the data into subsets based on the most informative features, leading to a hierarchical decision-making process. DTs are interpretable, and their ensembles (e.g., Random Forests) often yield robust performance. Figure 2.14 shows an example of a DT model used for the classification of heart attack risk.

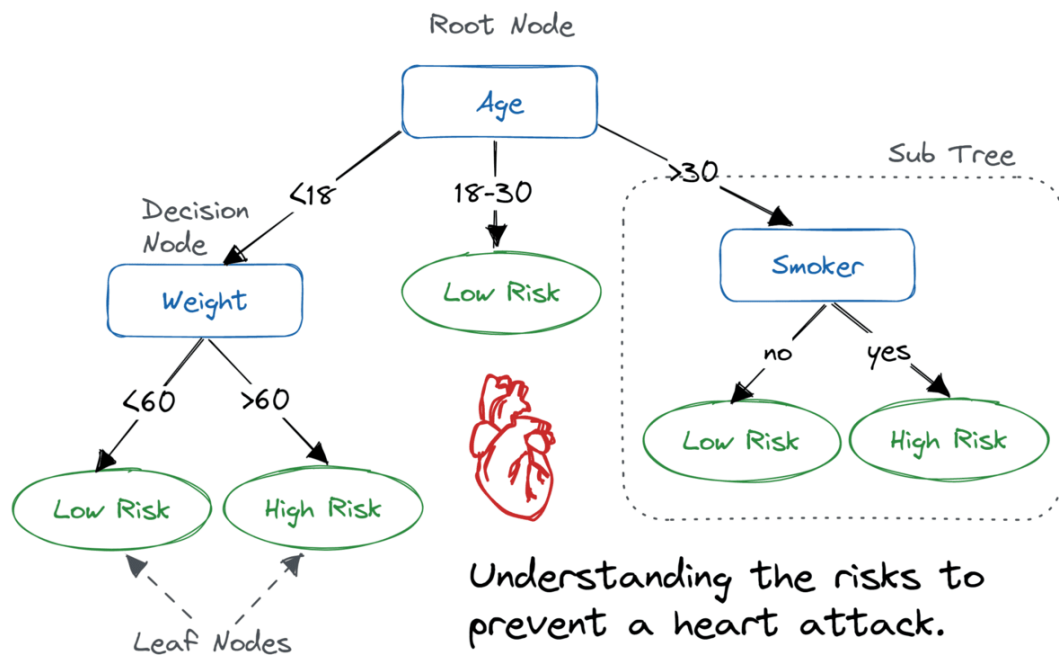


Figure 2.14: SL, Decision Tree, an example (Navlani, 2023)

Using its recursive structure, the decision tree articulates a sequential classification procedure. In this process, a case, characterised by a set of attributes, is

allocated to one of several distinct classes. Each terminal node (i.e., leaf) within the tree corresponds to a specific class, while an internal node signifies a test involving one or more attributes. For each potential result of this test, there is a subsidiary decision tree (Quinlan, 1987). To classify a given case, the process initiates at the tree's root. If the starting point is a leaf, the case is ascribed to the designated class. Conversely, if it is an evaluative test, the case's outcome is determined, and the procedure proceeds with the appropriate subsidiary tree aligned with that outcome.

2.4.6.4 Random Forest (RF)

Random Forest (RF) is an “ensemble learning” method that combines multiple DTs to improve classification accuracy and reduce overfitting. It introduces randomness both in data sampling and feature selection (Breiman, 2001). Since 1998, there has been a notable surge of interest in employing ensemble methods to enhance classifiers learning. These approaches typically involve taking a fundamental “base” learning algorithm and iteratively applying it to re-weighted versions of the initial training dataset, which yields an array of hypotheses that are subsequently amalgamated into a final aggregate classifier through a weighted linear voting mechanism (Grove and Schuurmans, 1998). The enthusiasm around ensemble methods stems from their ability to substantially improve predictive performance by leveraging the collective wisdom of multiple hypotheses generated during this iterative learning process.

As explained by Breiman, in the RF algorithm, a multitude of DTs is generated during the training phase (Breiman, 2001). These trees are constructed using subsets of the training data, chosen randomly with replacement. Additionally, when constructing each tree, a random subset of features is considered for splitting at each node. This feature selection approach introduces diversity among the trees, mitigating the risk of overfitting the model to the training data. During the prediction phase, each tree in the RF provides a class prediction (in classification tasks) or a numerical prediction (in regression tasks). The final prediction is determined through a voting mechanism for classification (where the class with the most votes

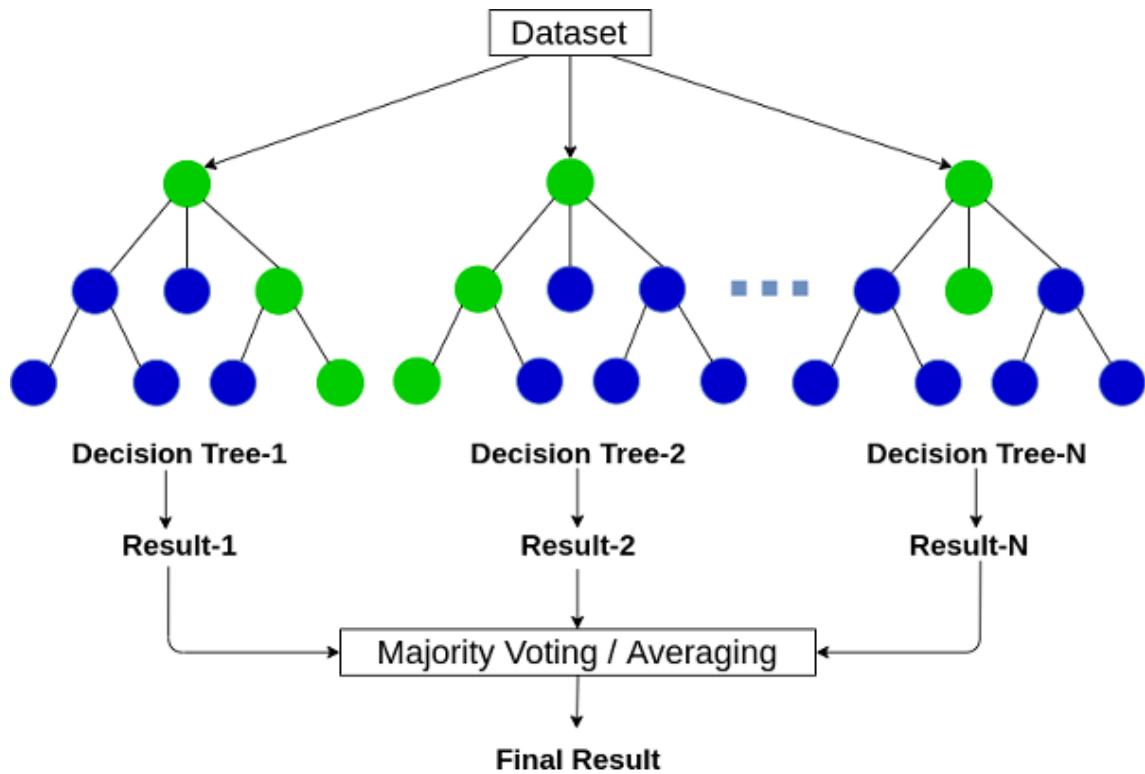


Figure 2.15: SL, Random Forest (Sharma, 2023)

is selected) or an averaging process for regression (taking the mean of the predictions) (Breiman, 2001). An example of how RF works is shown in Figure 2.15.

2.4.6.5 Gradient Boosting

Gradient Boosting is an ensemble ML technique for building predictive models, particularly for classification and regression problems. It belongs to the family of boosting algorithms, which combine the predictions of multiple “weak learners” (usually DTs) to create a stronger and more accurate model (Friedman, 2001).

The key idea behind Gradient Boosting is to sequentially train weak learners with a focus on the mistakes made by previous learners. Each new learner is trained to correct the errors of the combined ensemble up to that point. Gradient Boosting has several variants and implementations, including Gradient Boosting Machines (GBM), XGBoost (Extreme Gradient Boosting), LightGBM and CatBoost.

2.4.7 Classifiers Performance Evaluation

In ML, classifier performance evaluation is of paramount importance due to several critical reasons. It facilitates the selection of the most suitable classifier among a set of candidate models. Also, it provides a quantitative measure of how well a classifier is performing, which is crucial to assessing the quality and reliability of the model's predictions. Performance evaluation aids in tuning hyper-parameters to optimise a model's performance. It allows researchers and practitioners to fine-tune their models for better results.

ML models often underpin critical decision-making processes, such as medical diagnoses or financial risk assessments. Evaluating classifier performance ensures that these decisions are based on accurate and trustworthy information. Classifier performance evaluation is a fundamental aspect of ML. It ensures that models are not only accurate but also reliable, interpretable, and suited to their intended applications. Performance metrics provide valuable insights into a model's strengths and weaknesses, guiding further development and optimisation efforts. This practice is supported by a rich body of literature in the ML field (Japkowicz and Shah, 2011).

2.4.7.1 Confusion Matrix

The confusion matrix is an essential tool in the evaluation of ML models, particularly for classification tasks. It provides a structured and detailed summary of a model's performance by breaking down its predictions into different categories and comparing them to the actual ground truth (labels). This matrix is especially relevant in binary classification problems, but it can also be extended to multi-class problems. The confusion matrix is typically presented in a tabular format as depicted in Table 2.3.

According to (Sokolova and Lapalme, 2009), classification correctness can be assessed by calculating the following four components, which collectively form a confusion matrix that corresponds to binary classification scenarios.

- **True Positives (TP):** These are cases where the model correctly predicted the positive class, aligning with the actual positive instances. For example, this could

be when the model correctly identifies patients with a disease.

- **True Negatives (TN):** These are cases where the model correctly predicted the negative class, matching the actual negative instances. For example, when the model accurately identifies individuals without a disease.
- **False Positives (FP):** These are cases where the model incorrectly predicted the positive class when the actual class is negative. They are also known as Type I errors. An example is when a spam email is incorrectly classified as legitimate.
- **False Negatives (FN):** These arise when the model incorrectly predicts the negative class when the actual class is positive. They are also known as Type II errors. For instance, the model fails to detect a disease when it is present.

Table 2.3: ML Binary Classification, Confusion Matrix

		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	Positive	True Positive (TP)	False Positive (FP) (Type 1 Error)
	Negative	False Negative (FN) (Type 2 Error)	True Negative (TN)

Certainly, the primary objective of a predictive model is to maximise the True predictions (TP and TN) and minimise the False ones (FP and FN), which signify accurate and correct predictions.

2.4.7.2 Performance Metrics

The performance of a classification model can be assessed using various metrics derived from the confusion matrix. Throughout the research conducted in this

thesis, the following performance measures were used to evaluate the developed classifiers (Sokolova and Lapalme, 2009):

Accuracy (ACC): measures the overall correctness of predictions and is defined by Formula (2.10).

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.10)$$

Precision (PREC or P): quantifies the accuracy of positive predictions and is defined by Formula (2.11).

$$P = \frac{TP}{TP + FP} \quad (2.11)$$

Recall (REC or R) or Sensitivity: assesses the model's ability to identify all positive instances correctly and is defined by Formula (2.12).

$$R = \frac{TP}{TP + FN} \quad (2.12)$$

F-score (F): is the harmonic mean of precision and recall, providing a balanced measure. F-score was first introduced by Chinchor in 1992 for evaluation tasks in the field of information extraction technology (Chinchor, 1992) and is given by Formula (2.13).

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R} \quad (2.13)$$

where β determines the weighting of precision (P) and recall (R) in the F-score. When $\beta = 1$, it is the standard F1-score, which balances precision and recall equally. Thus, F1-score can be simply given by Formula (2.14).

$$\text{F1} = \frac{2 \times P \times R}{P + R} \quad (2.14)$$

Since Precision and Recall consider only parts of the confusion matrix, they are considered local performance metrics while accuracy and F-score are global ones as they consider all the components of the confusion matrix.

Accuracy is a commonly used evaluation metric as it is simple and intuitive but may not be the best choice for imbalanced datasets or when false positives and false negatives have different impacts (i.e., cost). Conversely, the F-score becomes particularly valuable in scenarios where A) the costs of false positives and false negatives are not the same, as seen in medical applications like mammogram evaluations for tumour detection, and B) there is a class imbalance, such as when only a small percentage of apples on trees are unripe (e.g., 10%). In such cases, a classifier will automatically categorise unripe fruit as ripe, resulting in a high accuracy rate (90%) but rendering the classifier unsuitable for practical use (Wood, 2019).

2.5 Natural Language Processing (NLP)

This section aims to provide background information on key concepts related to NLP and text analysis, which are relevant to the research presented in this thesis. The following sections include explanations of basic concepts such as n -gram, text metrics like TF-IDF, and Topic Modelling. These concepts play a crucial role in understanding and analysing textual data sources and resemble a fundamental component of many NLP and text mining tasks.

2.5.1 n -gram

n -grams are a fundamental concept in NLP and computational linguistics. They are used to analyse and model the relationships between words or characters in a given text or corpus and are widely used in NLP research and applications. n -grams are particularly useful for tasks like language modelling, text analysis, Information Retrieval (IR), Machine Learning (ML) and various NLP applications. Language models like N -gram models and Markov models utilise n -gram to predict the probability of the next word in a sequence. An n -gram is a sequence of n adjacent words, and here are some examples.

- **Unigram** (1-gram) is an individual word or character in the text. Unigrams

represent the simplest form of n -grams and are useful for basic frequency analysis. For instance, consider the text: “cyber security attack”; the unigrams are: “cyber”, “security”, and “attack”.

- **Bigram** (2-gram) consists of two consecutive words or characters. Bigrams provide information about word co-occurrences and can be used for language modelling. For the text above, the bigrams are: “cyber security” and “security attack”.
- **Trigram** (3-gram) is a sequence of three consecutive words or characters. Trigrams offer more context and information than bigrams and are often used in language modelling tasks. For the same example, there is only one 3-gram: “cyber security attack”.

2.5.2 Text Metrics

This section presents a list of textual metrics that have been utilised throughout this thesis. These metrics have been employed for a range of purposes, demonstrating their versatility and significance in our study. Firstly, these metrics have played a crucial role in ranking text tokens, enabling us to identify and prioritise the most relevant and informative terms for taxonomy construction. As discussed in Chapter 4, the process of constructing taxonomies heavily relies on the accurate identification of key terms, and these textual metrics have been quite useful in this regard.

Furthermore, the text metrics presented in the following sections have been instrumental in our text classification tasks, particularly during the feature selection stage of the created ML classifiers, as explored in Chapters 5 and 6. By incorporating these textual metrics into our feature selection stage, we have been able to extract and utilise the most discriminative and representative features, enhancing the effectiveness and accuracy of these classifiers.

2.5.2.1 Term Frequency (TF)

Term Frequency (TF) is a simple measurement of how a term (word) appears within a given text document. TF focuses on the local representation of a term within a specific document, giving weight to the frequency of occurrence. It is a document-specific metric that indicates how often a term appears in relation to the total number of terms in that document. TF can be useful for identifying the most frequent words within a document, but it does not provide information about the relative importance of a term across a collection of documents.

It is common practice to normalise term frequency to account for the potential bias introduced by varying document lengths. This is done by dividing the frequency of a term by the total number of terms in the document. Doing so mitigates the impact of document length on term importance. This normalisation ensures that the term frequency reflects the relative importance of a term within a document, regardless of its length. $\text{TF}(t, d)$ is the normalised frequency for the term t in a given document d and can be calculated using Formula (2.15).

$$\text{TF}(t, d) = \frac{\text{count}(t, d)}{\sum_{w \in d} \text{count}(w, d)} \quad (2.15)$$

2.5.2.2 Document Frequency (DF)

Document Frequency (DF) refers to the number of documents – in a corpus (i.e., a set of documents) – that contain a specific term. It is a measure used in IR and text mining to assess the significance or relevance of a term within a collection of documents. DF of a term indicates how widely it appears across a given corpus and can be used to determine the term’s rarity or commonality. A high DF suggests that a term is present in many documents and may be less informative or discriminating, while a low DF indicates that a term is relatively rare and potentially more distinctive.

DF for a term t in a set of documents D can be defined by Formula (2.16).

$$\text{DF}(t, D) = \frac{N_t}{N} \quad (2.16)$$

where N_t is the number of documents that term t appeared at least once, and N is the total number of documents.

2.5.2.3 Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) measures the global significance of a term in a corpus. It takes into account the distribution of the term across all documents in the collection. IDF assigns higher weights to terms that are less frequent in the entire corpus but occur in specific documents, which helps identify terms that are discriminative or unique to specific documents, making these terms more valuable for distinguishing between different documents.

IDF measures whether a term is common or rare in a set of documents D (i.e., a text corpus) (Nettleton, 2014). For term t , IDF_t is given by Formula (2.17).

$$\text{IDF}(t, D) = \log\left(\frac{N}{1 + N_t}\right) \quad (2.17)$$

where N is the total number of documents, and N_t is the number of documents that cover the term t .

2.5.2.4 Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF serves as a critical metric utilised in IR and ML domains. Its core purpose lies in assessing the importance of textual elements, such as words, phrases, or lemmas, within an individual document relative to a larger document collection, often referred to as a corpus (Simha, 2021).

TF and IDF are fundamental measures in text analysis, particularly in IR contexts. TF measures the frequency of a term within a document, while IDF gauges the significance of a term across a collection of documents. The TF-IDF score combines a term's local relevance within a document (TF) with its global importance

across the entire corpus (IDF) by multiplying these values for each term in a document (Stecanella, 2019). This composite measure identifies terms that are both frequently occurring within a document and relatively rare in the entire collection.

Furthermore, TF-IDF finds practical application in tasks like keyword extraction as TF-IDF scores can be employed as features for ML classifiers (Lippmann et al., 2017; Aslan, Sağlam, and Li, 2018). It is defined with respect to a set of “documents” or text corpora. For a given term t and a specific “document” d , $\text{TF-IDF}(t, d)$ represents the product of the term frequency $\text{TF}(t, d)$ for that term t in document d and the inverse document frequency IDF within the considered set of documents. To compute the TF-IDF matrix for a term t across a set of documents D , the same relationship is applied to the TF matrix, which contains term frequencies for term t in each document d from the set of documents D (Stecanella, 2019). This calculation is illustrated by Formula (2.18).

$$\text{TF-IDF}_D(t, d) = \text{TF}_D(t, d) \times \text{IDF}(t, D) \quad (2.18)$$

2.5.2.5 Weirdness Score

Weirdness score, in the context of text classification, refers to a measure used to assess and quantify the level of abnormality or distinctiveness exhibited by a given text or document when compared to a reference corpus. In work done by (Ahmad et al., 2000), they argued that the distribution of items (i.e., terms) would differ between a special corpus s and a general one g . Thus, they introduced the weirdness score, which is defined by Formula (2.19).

$$\text{Weirdness}(w) = \frac{\text{TF}(w, s)}{\text{TF}(w, g)} \quad (2.19)$$

where $\text{TF}(w, D)$ is the normalised frequency of word w in a given domain (i.e., a “class” of documents) D and can be calculated using Formula (2.20).

$$\text{TF}(w, D) = \frac{\text{count}(w, D)}{\sum_{w_i \in D} \text{count}(w_i, D)} \quad (2.20)$$

which is the division of the number of times word w appeared in domain D by the total number of words w_i in that domain. In this case, D can be either s or g .

2.5.2.6 Prototypical Words

The concept of prototypical words has been widely studied and applied in various research domains, including NLP, ML, and text mining. For instance, in text classification tasks, prototypical words are used as informative features to distinguish between different classes or categories. They are often identified through techniques such as feature selection, where the words with the highest discriminative power are selected based on their frequency, mutual information, or other statistical measures.

The selection of prototypical keywords is beneficial for classification tasks as they capture the characteristic lexical expressions commonly associated with users in a particular class. Pasca and Durme used a probabilistic model in their study (Pasca and Durme, 2007) to extract the prototypical words automatically, which was also adopted by others (Pennacchiotti and Popescu, 2011b; Aslan, Sağlam, and Li, 2018). Suppose we have n classes, and S_i are the seed users belonging to class c_i . Each word w will be assigned a proto score for each class using Formula (2.21).

$$\text{proto}(w, c_i) = \frac{|w, S_i|}{\sum_{j=1}^n |w, S_j|} \quad (2.21)$$

where $|w, S_i|$ is the total number the word w was issued across all users S_i in class c_i . The denominator cannot be zero for a given word if it was found at least once in any account's timeline. To get a higher value for the proto score of a word, it should appear only in one domain, not more, see Formula (2.21). The user u score for a particular prototypical word w_p can be given using Formula (2.22).

$$\text{f_proto_wp}(u) = \frac{|u, w_p|}{\sum_{w \in W_u} |u, w|} \quad (2.22)$$

where $|u, w_p|$ is the frequency of w_p in the user u timeline, and W_u is the set of all words found in that timeline.

2.5.3 Text Readability Scores

2.5.3.1 Lexical Diversity (LD)

Lexical Diversity (LD), often referred to as lexical richness, is a measure used to assess the variety of unique words in a text or a corpus. It indicates how many distinct words are used relative to the total number of words. A higher LD suggests a broader and richer vocabulary, while a lower LD indicates repetitive language usage. LD is a valuable metric in various fields, including linguistics, literature analysis, and text mining (Malvern et al., 2004).

LD can be calculated in different ways, but one common measure is called the “type-token ratio” (TTR), which assesses the proportion of unique words (i.e., types) to the total number of words (i.e., tokens) in a given text or dataset. TTR is given by Formula (2.23).

$$\text{TTR} = \frac{\text{number of word types}}{\text{number of word tokens}} \quad (2.23)$$

2.5.3.2 SMOG Grading

The SMOG grading stands for Simple Measure of Gobbledygook, which is a readability score designed to estimate the number of years of education a person needs to be able to understand a piece of text (McLaughlin, 1969). For example, an SMOG score of 10 would indicate that the text is readable by an average 10th grader.

The formula counts the number of polysyllabic words (i.e., words with three or more syllables) in a sample of text and is given by Formula (2.24).

$$\text{SMOG} = 3.1291 + 1.0430 \times \sqrt{\left(\frac{\text{Polysyllabic word count} \times 30}{\text{Sentence count}} \right)} \quad (2.24)$$

2.5.3.3 Flesch–Kincaid Grade Level (F-K)

The Flesch-Kincaid (F-K) readability formula, also known as the Flesch-Kincaid Grade Level Formula, is a readability test designed to measure the readability of

English text. It is based on the idea that simpler text is easier to understand. The formula calculates a grade level equivalent to the US education system, indicating what level of education is required to understand a given piece of text (Kincaid et al., 1975). F-K is given by Formula (2.25).

$$\text{F-K} = 0.39 \left(\frac{\text{words}}{\text{sentences}} \right) + 11.8 \left(\frac{\text{syllables}}{\text{words}} \right) - 15.59 \quad (2.25)$$

where: “words” is the number of words in the text, “sentences” is the number of sentences in the text, and “syllables” is the number of syllables in the text.

The output of this formula is a number, typically between 0 and 100. The resulting number is then mapped to a U.S. grade level, such as “grade 8” or “grade 12” indicating the minimum education level a person should have to understand the text easily. This formula is widely used in education, publishing, and content creation to assess the complexity and readability of texts. The lower the grade level, the more accessible the text is to a general audience.

2.5.4 Topic Modelling

Topic modelling is quite useful for doing a quantitative analysis of textual sources. We used Latent Dirichlet Allocation (LDA) algorithm (Blei, Ng, and Jordan, 2003) in our topical analysis, one of the most widely used topic modelling algorithms in the literature (Aslan, Li, et al., 2020; N. Pattnaik, Li, and Nurse, 2023). LDA is an unsupervised method for clustering N documents into k categories (i.e. topics). The interesting thing about LDA is that assigning a document to a topic is probabilistic, where each document is assigned to each topic with a probability, and the sum of all these probabilities is 1.0 per document (Kigerl, 2018).

LDA works in iterations to do two estimations: the distribution of words (i.e., tokens) into topics and the distribution of topics over documents (Blei, 2012). Thus, it requires two essential parameters to work, which are k , the number of topics, and r , the maximum number of iterations.

Chapter 3

Related Work

“If I have seen further, it is by standing on the shoulders of giants.”

- Isaac Newton

3.1 Cyber Security Taxonomies and Ontologies

3.1.1 Automatic Taxonomy and Ontology Building

AUTOMATIC and semi-automatic processing of information for building taxonomies and ontologies has been an active topic in different research fields. For instance, one popular technique, Formal Conceptual Analysis (FCA) (Priss, 2006), has been widely used to automatically construct a formal ontology from a given set of objects and their properties (Cimiano, Hotho, and Staab, 2004). NLP and ML techniques have also been widely used to automate taxonomy and ontology building, especially based on natural language texts (H. Yang and Callan, 2009). Such techniques are less used in building cyber security taxonomies and ontologies, as reviewed in the next two sections.

3.1.2 Selected Taxonomies in Cyber Security

Critical Infrastructure (CI) protection is one of the cyber security sub-domains where researchers have proposed frameworks for building taxonomies. Luijff and Nieuwenhuijs proposed a generic threat taxonomy for CI that is made up of 325 nodes (Luijff and Nieuwenhuijs, 2008). They built an extensible taxonomy to support adding more elements to their taxonomy as they did not develop their taxonomy from scratch but relied on existing threat databases instead. Also, Y. Jiang et al. proposed a domain-specific language for security in CI (Y. Jiang et al., 2018), where they created a simple taxonomy for CI and cyber components.

Some studies focused on building taxonomies for cyber-physical threats and attacks. Heartfield et al. built a taxonomy for the cyber security threats that affect smart homes (Heartfield et al., 2018). In their taxonomy, they considered the impact on both the system and users. On the other hand, Loukas et al. worked on vehicle security attacks, and they created a taxonomy for the characteristics of Intrusion Detection Systems (IDS) for different types of vehicles (Loukas et al., 2019). Sedjelmaci and Senouci also worked on taxonomy for vehicles but aerial ones in their study (Sedjelmaci and Senouci, 2018), where they examined the current detection schemes for aerial vehicle security and then classified them into a small taxonomy.

In a bid to analyse the propagation of large-scale cyber attacks, researchers have formulated several taxonomies. These taxonomies are intentionally developed to enhance comprehension and facilitate comparisons of such attacks, ultimately aiding in their prevention (Mohsen et al., 2022). Radmand et al. focused on creating a taxonomy for the cyber security attacks on wireless sensor networks. They studied many possible attacks on such networks. Their final taxonomy was limited to classes about attack categories (Internal or External) and whether they are Active or Passive (Radmand et al., 2010).

To assist in identifying and defending against cyber attacks, Simmons et al. proposed a cyber attack taxonomy named **AVOIDIT**. They employ five primary classifiers to categorise the characteristics of an attack: Attack Vector, Attack Tar-

get, Operational Impact, Informational Impact, and Defense. They claimed that classification by Defense is particularly valuable as it offers network administrators guidance on how to mitigate or rectify an attack by offering crucial attack information (Simmons et al., 2009). AVOIDIT is presented in a tree structure to carefully categorise common vulnerabilities employed in launching cyber attacks.

Syafrizal, Selamat, and Zakaria introduced **AVOIDITALS** cyber attack taxonomy, which was derived from both the AVOIDIT (Simmons et al., 2009) and Treadstone71 (Treadstone71, 2014) taxonomies. Their taxonomy encompasses 8 domains, 105 sub-domains, 142 sub-sub-domains, and 90 additional sub-sub-domains. This comprehensive taxonomy serves as a valuable tool for administrators, aiding in identifying cyber attack patterns that frequently occur on digital infrastructure. Furthermore, it offers the best prevention method to minimise the impact of such attacks (Syafrizal, Selamat, and Zakaria, 2021). They followed a 4-phase methodology: firstly, identifying existing cyber attack taxonomies; secondly, determining domains and subdomains related to cyber attacks; thirdly, classifying them; and finally, constructing the enhanced cyber attack taxonomy.

Mohsen et al. expanded the AVOIDIT taxonomy by introducing additional elements related to defence strategies, scale, and more. They conducted a comparative analysis between their newly proposed taxonomy and established state-of-the-art taxonomies. Furthermore, they analysed 174 significant cyber security attacks using their taxonomy. In addition, they introduced a web-based tool they developed, enabling researchers to explore existing cyber attack datasets and contribute new ones (Mohsen et al., 2022).

We cannot talk about cyber attacks without mentioning cyber harms. Agrafiotis et al. identified various types of harm and created a taxonomy of cyber harms encountered by organisations. Their taxonomy comprises five broad themes of harm: physical or digital, economic, psychological, reputational, and social/societal harm. For each of these themes, they presented several cyber harms that can result from cyber attacks (Agrafiotis et al., 2018). They analysed real-world case studies to offer

initial insights into the interconnections among these various harm types and the potential propagation of cyber harm in general. This supports their argument for the necessity of analytical tools to address organisational cyber harm based on a taxonomy such as the one they developed.

CTI is a sub-domain in which many ontologies have been developed to facilitate information sharing among different organisations and computer systems. Such ontologies are mostly represented as a common data format such as IODEF (Incident Object Description and Exchange Format) (Danyliw, 2016), STIX (Structured Threat Information eXpression) (Open, 2019) and OpenIOC (Open Indicators Of Compromise) (Gibb, 2013). Burger et al. created a taxonomy model to analyse and classify existing threat-sharing technologies using a neutral framework, to identify their limitations, and to explain their differences (Burger et al., 2014). On the other hand, Mavroeidis and Bromander surveyed existing CTI taxonomies and ontologies at the time they did their study (Mavroeidis and Bromander, 2017) and concluded that none of them is readily available to be used within CTI due to lack of expressiveness. They also suggested some actions in order to address this problem. Elnagdy, Qiu, and Gai built a knowledge structure (i.e., a mini taxonomy) of cyber insurance for practitioners in this specific industry (Elnagdy, Qiu, and Gai, 2016).

One of the most comprehensive taxonomies in the cyber security domain was proposed by Canbek, Sagiroglu, and Baykal in their study (Canbek, Sagiroglu, and Baykal, 2016), where they focused on the mobile security domain and built a large taxonomy covering different concepts. Supporting their taxonomy with two sub-taxonomies for mobile malware and mobile malware analysis, they proposed an overall hierarchy with over 1,300 nodes.

While it is challenging to quantify cyber security strength, Bhol, Mohanty, and P. K. Pattnaik argue that one can assess security efforts by employing cyber security metrics. These quantifiable measures are valuable for monitoring the progress of a specific process and evaluating its effectiveness. In their study (Bhol, Mohanty, and P. K. Pattnaik, 2021), they presented a taxonomy of cyber security metrics which

categorises cyber security into five primary domains: Vulnerabilities, Protection Mechanisms, Threats, Users, and Situational Encounters. Each of these metrics is further subdivided. Additionally, their taxonomy provides tools under a Multi-Criteria Decision-Making approach for assessing the strength of cyber security.

Another interesting study is the development of a comprehensive taxonomy for cyber security known as the European Cyber Security Taxonomy. This taxonomy aims to standardise cyber security terminologies and definitions, thereby facilitating the categorisation of cyber security competencies (Nai Fovino et al., 2019). The development methodology consists of several steps: 1) defining the scope of the selected subject, 2) identifying and selecting data sources, 3) collecting terms and concepts, (4) clustering concepts, and 5) reconciling the content and structure of the taxonomy. The authors considered several existing clustering approaches in the cyber security domain, and also international standards and reference documents related to cyber security. This meticulous process yielded a taxonomy comprising three dimensions, 18 categories, and 188 characteristics.

3.1.3 Selected Ontologies in Cyber Security

Quite a number of studies about building or using ontologies for the cyber security domain exist in the literature. However, most of them were created for a particular application (such as detecting vulnerabilities) and ignored several vital concepts in cyber security. Here, we review some typical work on this topic.

Elahi, Yu, and Zannone proposed a design for an ontology about the security concepts related to vulnerabilities in software. Their main goal was to integrate the captured knowledge into the security system requirements. Although their work was centred on vulnerabilities and software requirements, their modelling inspired us about taxonomy building (Elahi, Yu, and Zannone, 2009).

Razzaq et al. proposed a method based on semantic web techniques to detect attacks on web applications by analysing users' requests as such requests cover rich attack information. Then, they created ontology models for attacks and communi-

cation protocols (Razzaq et al., 2014).

Wang and Guo proposed an ontology for vulnerabilities and populated it using the description of some common vulnerabilities taken from the NVD (National Vulnerability Database) (Wang and Guo, 2009). Although their ontology modelled several entities related to vulnerabilities, they did not use it for exploring or finding new vulnerabilities.

Zamfira and Ciocarlie proposed a method for creating an ontology that can be used for detecting cyber security attacks. The ontology they built focuses on cyber operations and conceptualises different data needed in the processes. They also tested the use of the ontology with a prototype web firewall and showed that their ontology did help. They worked on implementing a cyber-defence using the techniques of the Semantic Web and tried to make their proposed ontology a comprehensive model in the context of cyber defence (Zamfira and Ciocarlie, 2018).

Maybe the most similar study to the work reported in Chapter 4 is (Costa et al., 2016), in which Costa et al. set an ontology for modelling insider threat attacks. The most interesting part – for this study – is their semi-automated approach to developing their ontology. They collected sources related to insider threat cases and used NLP tools to parse the extracted text sentences automatically. Then, they used human analysts to determine the meaning of each sentence for building the ontology. Their work was only on cyber indicators related to insider threats, which differs from our work on taxonomy building - presented in Chapter 4 - on several aspects: (1) Their output is an ontology while ours is a taxonomy. (2) Our focus is the cyber security domain as a whole, not just insider threats. (3) We use NLP and n -gram ranking, and the human expert is only involved in the taxonomy creation stage. Yet another related work is (Georgescu and Smeureanu, 2017), where Georgescu and Smeureanu set a theoretical framework for enhancing the cyber security field using ontology and other semantic web techniques. Their focus was on identifying black hat hackers by analysing the internet communication channels they usually use.

The explored studies about taxonomies and ontologies were focused on a sub-

field of cyber security or for a specific purpose rather than a general taxonomy. Also, the vast majority of these studies lack the utilisation of automatic text processing tools such as NLP and IR, and as a result, their taxonomies/ontologies were manually built. Moreover, these studies rely solely on expert knowledge to create their taxonomies/ontologies, unlike our data-driven human-machine teaming based process that we used to create a general cyber security taxonomy, see Chapter 4.

3.2 Detecting Cyber Security Related Accounts

For social media analytics research, there is a general need to automatically detect a certain community or type of account on a specific OSN platform. Classifying users on OSNs has been conducted through a variety of methods (Pennacchiotti and Popescu, 2011a). Some people worked on detecting the political orientation (Colleoni, Rozza, and Arvidsson, 2014) or party affiliation (Pennacchiotti and Popescu, 2011b) from the posts made by users on social media, while others researched detecting gender, age (Peersman, Daelemans, and Van Vaerenbergh, 2011), personality (Liu et al., 2016), income (Preoțiuc-Pietro et al., 2015), and many other attributes related to social media users.

There has been a range of related work on the automatic classification of OSN accounts for cyber security purposes. For instance, since spammer accounts are used by cyber criminals and hackers to perform a wide range of attacks, many of the studies conducted in this field have been around spam/spammer detection (Benvenuto et al., 2010; Krithiga and Ilavarasan, 2019). Similarly, due to the role of social bots and fake accounts in spreading misinformation/disinformation on OSNs, the detection of such accounts has become a hot topic in recent years (Abulaish and Fazil, 2020; Cao et al., 2012).

Monitoring activities of cyber security related accounts on OSN can provide interesting insights about about cyber criminals and cyber security professionals. There has been however relatively little work on automatic detection of those accounts.

Lee et al. developed a social media threat intelligence system called Sec-Buzzer, which includes a semi-automated component for adding new cyber security experts on OSNs by combining a number of mechanisms: mentions of active known accounts, a “topic relevance” score defined by the number of relevant cyber security topics, and manual confirmation by the Sec-Buzzer manager (Lee et al., 2017). The first fully automated classifier of this kind we are aware of is (Aslan, Sağlam, and Li, 2018), in which Aslan, Sağlam, and Li identified multiple candidate feature sets and tested several ML models to develop a classifier for detecting cyber security related accounts. Their best-performing model was Random Forest, which achieved accuracy above 97% using different feature sets. The dataset they used is relatively small (424 accounts) and was constructed in an ad hoc manner (e.g., cyber security related accounts were selected from an ad hoc public list, and all account labels were assigned by a single cyber security expert).

Yet another interesting work is (Jones, Nurse, and Li, 2020), where Jones, Nurse, and Li developed a classifier for detecting Twitter accounts affiliated with the well-known hacktivist group “Anonymous” to reconstruct a network of such accounts for studying their activities over time. Their best classifier was based on the Random Forest model and achieved an F1-score of 94%. Their classifier relies on ad hoc features manually identified for Anonymous accounts and the used dataset was collected based on a small number of (five) seed accounts, so it cannot be directly generalised to other types of cyber security related OSN accounts.

The aforementioned studies point to a clear research gap in using ML classifiers for detecting cyber security accounts on OSNs, especially those for detecting different sub-communities such as individual experts, hacking communities, and cyber security academia. In addition, there is a lack of public datasets for supporting the development and bench-marking of such classifiers. The datasets of (Aslan, Sağlam, and Li, 2018; Jones, Nurse, and Li, 2020) were neither publicly available to other researchers nor systematic enough, so they cannot be easily generalised. We aim to fill these gaps, including more systematically constructed datasets.

3.3 Studying Cyber Security Communities

With the enormous content created by OSN users daily, researchers have access to a massive and wide range of individuals (Andreotta et al., 2019). Different types of users can be found on OSNs, such as individuals, businesses, organisations and communities, hacktivists, and cyber criminals (Nouh and Nurse, 2015). To the best of our knowledge, there has been no previous work on studying cyber security researchers using a data-driven approach based on OSN data.

A lot of work has been done on studying cyber criminal groups on OSNs. For example, Aslan, Li, et al. studied a list of 100 defacer accounts by analysing their activities, social structure, clusters, and public discussions on Twitter (Aslan, Li, et al., 2020). While Kigerl studied the comments of 30,469 users from three carding forums in his study (Kigerl, 2018) by applying a clustering technique based on topic modelling. In another study about cyber criminals, Tavabi et al. built and analysed a large corpus of messages across 80 deep and dark web forums to identify the discussion topics and to examine their patterns (Tavabi et al., 2019).

Moreover, several other researchers studied activist and hacktivist groups on OSNs. For instance, Jones, Nurse, and Li analysed the presence of the Anonymous group on Twitter (Jones, Nurse, and Li, 2020). They built an ML classifier and identified over 20k accounts from the Anonymous group. Then, the key players were identified using SNA and centrality measures. By applying topic modelling, the main topics were found and used to study similarities between the key accounts. Another interesting work was done by Nouh and Nurse, where they studied a Facebook Activist group of 274 users with 670 posts. They created several graphs representing the users' friendships and interactions through the replies on the collected posts. Using SNA and different centrality measures, they analysed these graphs and identified the influential users. Also, sub-communities were discovered and studied. Then, they used sentiment analysis to study how user sentiment affected the group and investigated trust relations using link analysis techniques (Nouh and Nurse, 2015).

A few studies related to analysing non-expert users on OSNs were found. For

example, N. Pattnaik, Li, and Nurse conducted a large-scale analysis using cyber security and privacy discussions of non-experts on Twitter. The researchers developed two ML classifiers, one for detecting non-expert users and the other for detecting tweets related to cyber security and privacy. They used topic modelling to find the top topics discussed by non-experts. Using sentiment analysis, they discovered a general negative sentiment from non-experts about such topics (N. Pattnaik, Li, and Nurse, 2023). Another interesting study was conducted by Saura, Palacios-Marqués, and Ribeiro-Soriano, where they studied cyber security related issues discussed by home users on Twitter using a large dataset of 938k tweets. They used sentiment analysis, topic modelling, and mutual information to find these issues and study their effects on user privacy (Saura, Palacios-Marqués, and Ribeiro-Soriano, 2021).

The work of Zeng et al. is quite interesting and related to our community analysis of cyber security experts on OSNs (presented in Chapter 6), where they introduced a cyber security community detection framework designed for OSNs. They created a machine learning classifier to detect security related accounts on Twitter with good performance, reaching 95% for accuracy. Their classifier was used to identify the security related accounts in the ego networks of selected seed accounts. Then, they built the social graphs of security related accounts in each ego network. A pruning strategy was adopted to eliminate weak connections between accounts based on edge features. After that, they applied an unsupervised overlapping community detection model to uncover potential communities (Zeng et al., 2023).

The literature encompasses various studies examining different types of cyber security groups and their communities on OSNs, ranging from cyber criminals and hackers to activists, non-experts, and general cyber security accounts. However, to the best of our knowledge, the specific realm of online communities comprising cyber security experts remains relatively unexplored. Studying the activities and communities of these experts on social media holds the potential to yield valuable insights into their community dynamics, communication patterns, discussions, influence, and more. This research gap serves as the focal point addressed in Chapter 6.

Chapter 4

Building General Cyber Security Taxonomy

*“We are drowning in information and starving
for knowledge.”*

- Rutherford D. Roger

4.1 Introduction

TAXONOMIES and ontologies are handy tools in many application domains, such as knowledge systematisation and automatic reasoning. In the cyber security field, many researchers have proposed such taxonomies and ontologies, most of which were built based on manual work. Some researchers proposed the use of computing tools to automate the building process, but mainly on very narrow sub-areas of cyber security. Thus, there is a lack of “general” cyber security taxonomies and ontologies, possibly due to the difficulties of manually curating keywords and concepts for such a diverse, interdisciplinary and dynamically evolving field.

This chapter presents a new human-machine teaming-based process to build taxonomies, which allows human experts to work with automated NLP and IR tools

to co-develop a taxonomy from a set of relevant textual documents. The proposed process could be generalised to support non-textual documents and to build (more complicated) ontologies as well. Using the cyber security domain as an example, we demonstrate how the proposed taxonomy-building process has allowed us to build a “general” cyber security taxonomy covering a wide range of data-driven keywords (topics) with a reasonable amount of human effort.

4.1.1 Cyber Security Taxonomy

Taxonomies and ontologies are both useful knowledge representation tools for systematically and structurally conceptualising human knowledge about objects (or things) and concepts in many domains, especially in sciences, engineering, business and education (June, 2010). The two words “taxonomy” and “ontology” have very similar meanings, but the latter has a more theoretical flavour and requires typically more advanced components such as relations between concepts, therefore allowing formal reasoning about the meanings of sentences (New Idea Engineering Inc., 2018).

The field of cyber security encompasses various disciplines and undergoes constant evolution. It is not surprising that many researchers and practitioners have attempted to build and use taxonomies and ontologies to better organise the knowledge from different sub-areas of the broad subject. See Section 3.1 for a brief overview of some related work on cyber security taxonomies and ontologies. Although there has been a lot of work on taxonomy and ontology building, there is a general lack of more automated processes for building taxonomies and ontologies. In addition, more general taxonomies and ontologies covering the whole subject are rare, possibly due to the more demanding human effort required, the complexity of putting everything together and the constant effort to keep such taxonomies and ontologies up to date.

4.1.2 The Necessity of a “General” Taxonomy

First of all, by “general” we do not mean full or extensive, but we mean a taxonomy that is not specialised or exclusive to a sub-domain, topic, or area in the cyber security domain like the taxonomy examples that were mentioned in the literature review, see Section 3.1.2. Specialised cyber security taxonomies offer precision, relevance, depth, customisation, and interoperability. Also, they provide a tailored framework for specific domains or areas, enabling focused research, analysis, and decision-making, while fostering collaboration and information exchange within targeted communities.

The decision to build a “general” cyber security taxonomy was motivated by several considerations. Firstly, the cyber security landscape is vast and constantly evolving, encompassing a wide range of threats, vulnerabilities, and actors. A general taxonomy provides a broad framework that can accommodate the diverse nature of cyber security phenomena, including, for example, Network Security, Application Security, Cloud Security, Data Security, Identity and Access Management (IAM), Socio-technical Security, Security Operations (SecOps), Incident Response and Forensics, Ethical Hacking, Security Compliance and Governance, and many other emerging sub-domains.

Second, we plan to create a taxonomy that covers cyber security as a topic, so it can be used in the next steps of our research, in the intended ML classifiers (Chapter 5) and the cyber security expert communities analysis (Chapter 6). As highlighted in the literature review, previous studies have explored various types of cyber security groups on OSNs, including cyber criminals, hacktivists, activists, and non-experts. However, there remained a significant gap in understanding the online communities of cyber security experts specifically. By opting for a general cyber security taxonomy, we aimed to address this research gap comprehensively and provide insights into the activities and interactions of cyber security experts across the different sub-domains and areas within the cyber security domain. Moreover, cyber security experts represent a diverse group encompassing researchers, practi-

tioners, innovators, vendors, and more. Focusing on a specific sub-domain or area within the cyber security domain would have limited the scope of our analysis and potentially overlooked important insights from other areas. Studying the broader landscape of cyber security experts' activities, discussions and communities allowed us to capture a more holistic view of this domain, including emerging threats, best practices, collaboration opportunities and much more.

Additionally, a general cyber security taxonomy serves as a foundational framework that can be adapted and extended to suit specific use cases and domains. While it may not cover every niche area in detail, it provides a common language and structure for organising cyber security concepts and knowledge. Furthermore, the choice to build a general cyber security taxonomy does not preclude the development of more specialised taxonomies or ontologies for specific areas or groups. Instead, it provides a starting point from which more targeted taxonomies can be derived. By establishing a general framework first, we lay the groundwork for future research and development efforts in cyber security taxonomy construction.

In terms of specific use cases and limitations, it is essential to acknowledge that a general cyber security taxonomy may not be exhaustive or fully applicable to every context. However, its utility lies in its flexibility and maintainability, as well as its potential to facilitate knowledge sharing and collaboration across diverse cyber security communities.

4.2 Methodology

In this chapter, we propose to apply the new concept of human-machine teaming (UK Ministry of Defence, 2018) for taxonomy-building in order to reduce the human effort involved in the building process and to make it easier to create taxonomies and maintain them. The proposed process includes three stages: A) the data collection phase for preparing a large set of textual documents of interest and also documents in other areas, B) the text analysis phase for processing the

textual documents to produce a list of relevant terms (keywords); C) the taxonomy-building phase for creating and refining the taxonomy based on a defined structure and assignment of all terms from Stage B to the structure. Stage A involves a manual selection of textual documents done by the human analyst and the automatic processing of collected data to form a properly formatted dataset ready for the next stage. Stage B is heavily automated using NLP and IR tools, but the human analyst controls the final selection of terms. Stage C is mostly done manually based on the human analyst's expert knowledge but can be facilitated by an automated tool for visualising the taxonomy. Taking cyber security as an example domain, we demonstrated how the proposed process has been used to build a general cyber security taxonomy with a reasonable amount of human effort and a large set of textual documents processed by automated NLP and IR tools.

While our methodology indeed leverages computational tools and automated techniques, it is important to underscore the integral role of human expertise and input throughout the taxonomy-building process. Firstly, human involvement is paramount in defining the initial scope and objectives of the taxonomy construction, including defining the key concepts, top classes, sub-domains, and categories within the cyber security domain. Additionally, human experts play a crucial role in curating and validating the data sources used for the corpora creation, ensuring their relevance and comprehensiveness.

Furthermore, human judgment is indispensable in refining and validating the results obtained through automated processes across the different stages. This includes reviewing and interpreting the output at each step, identifying any discrepancies or inaccuracies, and iteratively refining the taxonomy based on expert insights and domain knowledge. In essence, while our approach incorporates machine-driven techniques to expedite certain aspects of the taxonomy-building process, it is fundamentally guided and enriched by human expertise, intuition, and oversight. The synergy between human and machine capabilities fosters a collaborative and iterative approach that harnesses the strengths of both to achieve better results and

reduce the time and effort needed by the human expert to ensure better, representative and robust taxonomy development in the complex and rapidly evolving domain of cyber security.

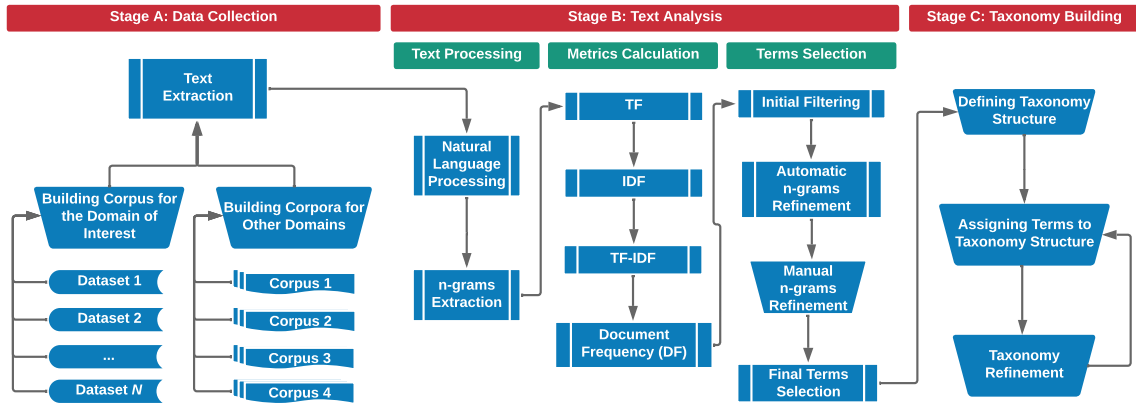


Figure 4.1: Human-machine teaming-based methodology for taxonomy-building

The high-level overview of our proposed taxonomy-building process is illustrated in Figure 4.1. The process starts with selecting the needed textual sources. This is mostly done by humans, but their job is limited to the identification and collection of documents so that it can be done more easily. After that, the text analysis phase is executed to produce a list of relevant terms (mostly automated). Finally, the process ends with a relatively light human-driven process of building the taxonomy. The three steps of the process are outlined in the following sections.

4.2.1 Stage A: Data Collection

The main goal of this stage is to prepare a properly formatted set of textual documents for analysis in Stage B. Different textual datasets are needed here so that the built taxonomy can cover more useful terms that can help separate the domain of interest (e.g., cyber security) from other domains. This stage requires human effort to identify relevant documents from different domains and to aggregate all selected documents together, using automated tools, into a format ready for Stage B.

4.2.2 Stage B: Text Analysis

The purpose of this stage is to produce a number of relevant terms to be assigned later to a taxonomy structure defined in Stage C. Stage B consists of three major steps. The first one is about processing the prepared corpora to extract the n -grams. The second step is calculating some useful metrics, including TF, IDF, TF-IDF and DF scores, which will be explained later in the next sections using cyber security as an example domain. The last step is about selecting relevant terms from all n -grams based on the calculated metrics, which include three sub-steps: initial filtering, n -gram refinement, and final term selection based on TF-IDF ranking. Stage B is largely automated using NLP and IR tools. However, the last step needs to involve the human analyst to empirically determine some parameters and exclude irrelevant terms based on the analyst's domain knowledge.

4.2.3 Stage C: taxonomy-building

In this stage, we build the taxonomy from the relevant term list produced earlier in Stage B. Stage C consists of the following steps. In the first step, the initial structure of the taxonomy is defined, especially the top-level and the second-level classes¹ and other components that are known before inspecting all the relevant terms. The definition of the initial structure can be done by the human analyst based on their domain knowledge, independently of the relevant terms, but can be informed by what key concepts can be used to cover all those terms. The second step is assigning the relevant terms produced in Stage B to the taxonomy structure. In the final step, the whole taxonomy is refined based on any issues identified from the term assignment step, including necessary adjustments of the basic structure due to re-assignments of some terms. This stage is mostly done by the human analyst, although some software tools could be developed to facilitate the term assignment. For example, an automated recommendation system can be used to suggest where

¹Taxonomy classes/subclasses, also called domains/sub-domains.

a term should be mapped. Also, creating a visualisation of the taxonomy would aid the human analyst during terms assignment and taxonomy refinement steps.

The whole process can be repeated in full, or partially, to keep the created taxonomy up to date. For example, after the taxonomy is created, the text processing step in Stage B can be run again to produce more candidate n -grams to enrich the taxonomy. Moreover, some new documents can be added and processed to update the taxonomy to reflect changes from relevant topics. Only new terms or outdated terms need processing for such updates of the taxonomy, so its maintenance can be relatively light.

As a whole, the process combines the work of humans and machines well to co-create the taxonomy. A vital feature of the process is that the human analyst does not need to arbitrarily define a list of keywords, which is often the hardest, the most time-consuming, error-prone and “random” part of any taxonomy-building process. Instead, the analyst can work with a list of automatically produced terms. Such a human-machine teaming process not only helps reduce human effort but also increases the accuracy of the built taxonomy.

Note that the proposed methodology can be used for any domain or purpose, not just the cyber security domain. This is because the approach is data-driven, i.e., the candidate terms are automatically harvested from a given dataset. Domain knowledge experts are still needed to select relevant documents and process the automatically produced candidate terms, but the most time-consuming and arbitrary part of the work – defining what terms to use – is largely automated. Therefore, by using a different dataset containing documents from a different domain, one can build a taxonomy for that domain.

In the following three sections, we will use cyber security as an illustrative example domain to showcase the application of the suggested process to create a general cyber security taxonomy. Each section will delve into a specific stage of the process, providing a focused examination of its implementation and outcomes.

4.3 Data Collection

For building the cyber security taxonomy, we collected five datasets (corpora) of textual documents. The first dataset consists of cyber security related sources, while the others cover documents from four selected non-cyber security domains. The reason for collecting non-cyber security documents is to eliminate common terms across all domains and, therefore, not indicative of the cyber security domain. This will be accomplished mainly by evaluating the TF-IDF (explained in Section 2.5.2.4) scores of n -grams, to identify good candidate terms for the cyber security domain.

We tried to ensure a good degree of distinctiveness between the text corpora across different domains. Firstly, during the data collection phase, careful attention was paid to creating each corpus from sources specifically relevant to their respective domains. For instance, the News corpus was sourced from reputable news outlets covering a broad range of topics, excluding those primarily focused on cyber security. Similarly, texts for the English Literature corpus were selected from literary works, letters, linguistics, novels, ..etc which are unrelated to cyber security topics. The Law corpus contained government articles and debates from the Canadian parliament. Also, the topics covered in the Science corpus are far from any cyber security concepts. Furthermore, to mitigate the risk of inadvertent inclusion of cyber security related content in non-relevant corpora, careful screening and filtering processes were applied. This involved manual review and validation of the collected texts to ensure that they predominantly pertained to the designated domain and were free from cyber security references.

While the possibility of minor overlap or contamination between corpora cannot be entirely ruled out, the research endeavoured to minimise such occurrences. However, ensuring a complete separation is not necessary as we have some tolerance in the later stages of the taxonomy-building process. See Section 4.4.3.1 for more details about how we achieved this using the IDF measure during the Initial Filtering step, where we allowed a candidate n -gram to appear in at most two corpora.

4.3.1 Building Cyber Security Dataset

Since our main objective is to build a general taxonomy for cyber security, we needed first to create a representative corpus for cyber security using human-written textual documents. For this purpose, different sources were collected to cover more diverse n -grams that are commonly used by different cyber security related people such as professionals, academics and hackers. The created textual corpus consisted of documents selected from the following four representative types of data sources.

1. **Professional reports** related cyber security and issued by well-known organisations such as ENISA (European Network and Information Security Agency)² and UK's NCSC (National Cyber Security Centre)³. These reports correspond to cyber security professionals associated mostly with government and industry.
2. **Academic papers** written by cyber security researchers. Some papers in this category were collected by searching Google Scholar using broad keywords e.g., "cyber security" and "information security". Other papers were cyber security papers already known to us. These papers correspond to people related to academia.
3. **OSN data**, which was a collection of Twitter timelines of cyber security related Twitter accounts detected by a trained machine learning classifier reported in (Aslan, Sağlam, and Li, 2018). Each account's timeline was a text file that resulted from merging all the tweets that were retrieved for this Twitter account. The maximum number of tweets that could be collected per account was 3,250, as this was a limitation set by the Twitter API.
4. **Underground forum posts** that contained discussions took place in underground forums used by hackers and cyber criminals. For this data source, we used data from the Cybercrime Centre at the University of Cambridge (Pastrana et al., 2018). This data source was used to gain insights into the terms that are usually used among hackers and cyber criminals.

²<https://www.enisa.europa.eu/>

³<https://www.ncsc.gov.uk/>

The number of documents and their formats that were used in the cyber security corpus are listed in Table 4.1, including the numbers of tokens and word counts.

Table 4.1: Data sources statistics in the cyber security corpus

Data Source	Format	Documents	Tokens	Words
Professional reports	PDF	117	1,850,804	736,280
Academic papers	PDF	385	5,465,389	2,001,726
OSN data	TXT	219	10,635,807	3,532,320
Underground forums	HTML	69	10,554,781	3,448,526

4.3.2 Building Non-Cyber Security Corpora

For non-cyber security documents, we used four corpora corresponding to the following domains: News, Law, General Science and English Literature. Each corpus is big enough to be considered representative of its domain. See Table 4.2 for statistics about these textual corpora, from which we can see the different corpora have 2-10m words and their sizes are comparably rich.

Table 4.2: Statistics of all five corpora used

Corpus	Documents	Tokens	Words
Cyber Security	790	28,506,781	9,718,852
English Literature	3,321	24,581,859	7,544,295
Law	635	19,474,191	6,422,816
News	4,561	8,536,780	3,346,196
Science	5,149	5,553,724	2,282,068

4.3.3 Text Extraction

The processing pipeline started with reading the sources, after which the texts were extracted. For each file type, we used a different file parser. For example, the webpages parser removes all HTML tags (i.e., characters between < and >) and extracts the remaining text. Moreover, the parser of PDF files removes the meta-data fields and URLs. Then, it extracts the plain text. For the Twitter timeline of an account, the parser converted it into one text file by concatenating only the plain text of the tweets and removing all other data fields.

4.4 Text Analysis

As previously mentioned, the text analysis stage comprises three primary steps: Text Processing, n-grams Metrics Calculation, and Term Selection. Each of these steps consists of several tasks, which are outlined in detail below. These tasks collectively contribute to the comprehensive analysis of textual data by processing and transforming the text, calculating relevant metrics, and selecting informative terms for subsequent analysis.

4.4.1 Text Processing

In this step, and before applying the NLP tools, several pre-processing steps were applied to reduce the number of tokens for the later steps. We removed URLs, emails, independent numeric strings and punctuation symbols. Other strings related to Twitter data, such as retweet indicator “RT”, @usernames and hashtag symbol “#” were also removed.

4.4.1.1 Natural Language Processing

An NLP tool takes the raw text as input and returns the annotated text as output. We used several annotators from the Stanford CoreNLP (Manning et al., 2014) library (Stanford NLP Group, 2019) to first tokenise the text, split tokens into

sentences, assign part of speech (POS) tags to each token and then to apply lemmatisation for each token to obtain the original root without any suffixes or prefixes. After that, we removed the stop words and applied several rules to eliminate words that were less useful for our purpose, e.g., words that were too short, too long or non-English.

4.4.1.2 n-grams Extraction

Following the NLP processing, we extracted n -grams with n ranging from 1 to 5. Initially, only unigrams and bigrams were considered. However, since we were interested in building a general taxonomy of the cyber security domain, we had to ensure that our methodology covers as many relevant concepts as possible. Also, we noticed that for this domain, many valid terms are long n -grams (e.g., “Access Control Policy”, “Cyber Threat Information Sharing”, “National Cyber Security Awareness Month”). Thus, trigrams, fourgrams and fivegrams were considered as well. This reflects how the proposed process can easily adapt to refine the taxonomy.

Table 4.3: Statistics of extracted n -grams for each corpus

Corpus	1-grams	2-grams	3-grams	4-grams	5-grams	Total
Cyber Security	260,737	978,997	547,868	250,389	150,080	2,188,071
English Literature	280,380	1,119,025	470,754	143,686	46,662	2,060,507
Law	60,477	414,052	208,730	61,662	16,371	761,292
News	121,513	511,277	226,911	75,255	24,871	959,827
Science	95,618	350,101	180,699	56,999	16,950	700,367

Statistics about all the corpora are presented in Table 4.3, which shows the number of n -grams of each size (1 to 5) for each corpus. The table clearly shows that longer terms are used more often in the cyber security domain than in others.

4.4.2 N-grams Metrics Calculation

At the end of the last step, we ended up with over 6.7 million n -grams, including 2.2 million from the cyber security corpus, which were too many to work with during the manual assessment in Stage C of the taxonomy-building process. Thus, we needed to filter them down to a more manageable size. To this end, we calculated a number of metrics for each n -gram, which are then used to filter and select a smaller number of n -grams as valid terms.

4.4.2.1 Term Frequency (TF)

Term Frequency is a simple measurement of how a term or word appears within a given document. The term frequency for word w in document d is $TF_{w,d}$, which is defined as the count of w in d .

4.4.2.2 Inverse Document Frequency (IDF)

According to TF, all terms are treated equally in terms of importance. However, some terms are useless and not informative despite the high TF score they can have e.g., “the”, “of”, “is”, “that”, ..etc. Thus, we must consider how rare a term can be across several documents. For this purpose, we used the IDF measure, which was explained in Section 2.5.2.3, and it is calculated using Formula (2.17). In our case, the total number of “documents” N used in the IDF formula is 5 as we have 5 corpora in total.

4.4.2.3 Term Frequency-Inverse Document Frequency (TF-IDF)

We used the TF-IDF scores to rank all the n -grams. The ranked list was used in the next step to filter and select the top n -grams with higher TF-IDF values as candidates for building the taxonomy. We explained the TF-IDF in Section 2.5.2.4, and it was given in Formula (2.18), which is the product of TF by IDF.

4.4.2.4 Document Frequency (DF)

Some terms are assigned a relatively high TF-IDF value, although they appear in a very small number of documents (i.e., sources) in the cyber security corpus. Those highly document-specific terms are more likely unrelated to the cyber security domain; otherwise, they should have been more widely used. To exclude such terms, we also calculated each n -gram's DF, which is the number of documents in the corpus of interest (e.g., cyber security) this n -gram appears at least once.

DF was explained in Section 2.5.2.2, and defined by Formula (2.16). It is important to mention that DF was considered within each corpus only. Thus, the “documents” in this context means the textual sources that formed a corpus, while for IDF, the documents are the corpora (i.e., each corpus is a document).

4.4.3 Terms Selection

We aimed to streamline the selection process by focusing on a condensed set of the most significant and representative n -grams within the cyber security domain. This approach aimed to minimise the manual effort required during the taxonomy-building stage where assigning terms and refining taxonomy require a lot of manual work. Thus, we implemented a series of filtering and refinement procedures to achieve this objective, as outlined in the subsequent sections.

4.4.3.1 Initial Filtering

Two filtering procedures were applied in this step to reduce the number of candidate n -grams. First, we utilised the text corpora created earlier to select the n -grams that are more unique to the cyber security domain using the IDF score, where we only select the n -grams that appeared either in the cyber security corpus alone or in two corpora and one of them is the cyber security corpus. Since we have five domains, that means we have five possible values for the IDF score, ranging from 0 to 0.7 where the lowest value (0) corresponds to an n -gram that appeared in all corpora

and hence it is very common and less unique or related to any particular domain, e.g., “time”, “data”, “security”, “network”, “computer”, ..etc. On the other hand, the maximum is 0.7, which corresponds to an n -gram that appeared only in one domain, e.g. “malware”, “cybersecurity”, “inforsec”, “zero-day”, ..etc. IDF score of 0.4 corresponds to n -grams that appear in no more than two corpora. By considering these n -grams as well, we ensured that n -grams such as “blog”, “phishing”, “bitcoin”, “antivirus”, ..etc were not excluded because they appear in at most one corpus beside the cyber security corpus.

Second, the IDF condition alone is not enough as we still have to deal with a large number of n -gram. Also, we are interested in the n -grams that are not just related to the cyber security domain but at the same time popular in this domain. Thus, we added a second condition utilising the Document Frequency (DF) measure, where we select the n -grams that appeared at least in five documents (i.e., sources).

Table 4.4: Statistics of the initial filtering step

IDF		All Domains	Cyber Security Domain	
Score	Domains	Ngrams	Ngrams	Ngrams (DF \geq 5)
0	5	101,095	20,219	14,855
0.1	4	145,580	34,139	12,429
0.22	3	229,710	56,440	10,003
0.4	2	565,754	159,073	15,428
0.7	1	5,627,925	1,918,200	47,094

Consequently, the initial filtering will select candidate n -grams that satisfy the following two conditions:

1. $IDF_w > \log(5/2) = 0.4$, i.e. the n -gram should appear in no more than two corpora; otherwise, it is not unique enough for the cyber security domain.
2. $DF_w \geq 5$, i.e. the n -gram should appear in at least 5 cyber security documents.

Table 4.4 shows the n -gram statistics for these filtering steps. As a result of the initial filtering, we reduced the n -grams to around 62k ones.

4.4.3.2 Automatic Refinement of n -grams

The automatic n -gram refinement step is necessary to help reduce irrelevant terms automatically without the need for a human expert. It can be applied to all the extracted n -grams without any additional effort, unlike any manual refinements. We applied three automatic sub-processes, as illustrated below. It is worth mentioning that this refinement step can actually appear anywhere after extracting the n -gram.

A) Remove Spam n -grams

This was needed because part of the used cyber security corpus contained underground forum posts, which had a lot of spam texts, links, advertisements for selling different products (especially medicines and drugs), and other less useful texts. To eliminate those “spam” n -grams, We followed a simple approach that proved very effective: we identified the most used names of products that usually appear after the word “buy” and created a blacklist for those names. Then we eliminated any n -gram containing at least one word from the blacklist. This removed more than 95% of those terms. We did not need an extensive advertisement terms removal mechanism as the remaining 5% had lower TF-IDF values anyway and did not appear among the top extracted n -grams.

B) Remove n -grams covered by other longer n -grams

To eliminate redundant n -grams that are *completely* covered by other ones, we applied what we named as n -gram “coverage rules”. By “covered”, we mean that an n -gram t of size S is a sub “word sequence” of another n -gram t' of size $S' > S$, and the former appears only as part of the latter. For example, if a unigram “formalising” appears only when “formalising security”, then the latter is the concept that should be captured. See Table 4.5 for some examples. Since we extracted n -grams from sizes one to five, we considered four different types of coverage rules as follows: 1) a unigram covered by a bigram, 2) a bigram covered by a trigram, 3) a trigram

covered by a four-gram, and 4) a four-gram covered by a five-gram.

When applied correctly, coverage rules can help reduce the number of n -gram candidates for future processing. However, in some cases, an n -gram completely covered by another one may not be redundant as it can bear a broader semantic meaning than the latter, e.g., in a corpus “cyber” may accidentally appear only with ‘cyber security’, but “cyber” clearly should be kept as a standalone n -gram since it has a richer semantic meaning. To this end, the coverage rules should not be used alone, but its results can always be manually checked to avoid mistakes, i.e., “good” n -grams got wrongly eliminated.

Table 4.5: Examples of n -grams eliminated by the coverage rules

Case	1-gram	2-gram	3-gram	4-gram	5-gram
2-gram covered by a 3-gram	default	windows			
3-gram covered by a 4-gram	default	windows	kernel		
4-gram covered by a 5-gram	default	windows	kernel	debugging	
Accepted 5-gram	default	windows	kernel	debugging	setting

C) Remove n -grams with low TF/DF scores

For each n -gram size (1 to 5), we set an empirically determined threshold for TF and a size-specific threshold for DF to eliminate further some n -grams that do not appear frequently enough.

4.4.3.3 Manual Refinement of n -grams

After the automatic refinement step, we sorted all remaining n -grams by their TF-IDF scores. Then we selected the top 1,000 unigrams, the top 1,500 bigrams, the top 1,000 trigrams, the top 500 fourgrams and the top 500 fivegrams, which formed a set of 4k n -grams as candidate terms for further processing. These candidate terms were then examined manually to remove irrelevant terms, correct wrongly extracted terms, and merge terms that represent the same thing.

4.4.3.4 Final Terms Selection

Following the manual refinement step, we got a list of around 2k terms, which we used in the taxonomy-building stage. The final terms in the list were informative and related to the cyber security domain.

4.5 Taxonomy Building

This is the third stage of the proposed taxonomy-building process, where the taxonomy is created and refined. The stage consists of the following steps.

4.5.1 Defining Taxonomy Structure

We needed to define an initial basic – not necessarily comprehensive – structure for the cyber security taxonomy with a sufficient level of detail to facilitate the term assignment in the next step. To this end, we studied existing cyber security taxonomies to get some insights into how we could design the initial structure of our taxonomy. Although the terms used in the taxonomy were selected following a data-driven process, we had to define at least the top-level classes manually and sometimes subclasses at a lower level to set the basic structure. The initial structure is not supposed to be fixed and will be further adjusted during the term assignment step, which may suggest a better way to refine the classes and subclasses.

Upon manually checking the final selected terms and the existing taxonomies and ontologies in the literature, we created 9 top-level classes, each one representing an important aspect of the cyber security domain. Then, we added another special top-level class called “Sub-domain” which corresponds to any sub-domain in the cyber security domain. Thus, we started building the taxonomy from top to bottom using 10 top-level classes which will be described below. It is worth mentioning that the number of the top-level classes is not fixed and it may vary from one taxonomy to another and based on the human expert who is building the taxonomy. Figure 4.2 represents a visualisation of the high-level structure of the initial cyber

security taxonomy, showing the main classes and subclasses created before the term allocation phase. Class node colours were used for illustration purposes only.

4.5.1.1 Class: “Individual”

Human users are the main actors in the cyber security domain. Therefore, a more detailed view should be provided about the different roles that cyber security related people can have. Some of the subclasses under this class are: “End User”, “Expert”, “Academic”, “Hacker”, “Cybercriminal”, “Activist” and “Journalist”.

4.5.1.2 Class: “Party”

This is a main class added to represent different types of human gatherings and organisations that play a role in the cyber security domain. Some of the subclasses are: “Research Centre”, “Educational Institute”, “Government”, “Critical Infrastructure”, “NGO”, “Business”, and “Group”. The “Business” subclass has more subclasses about different types of business entities and the “Group” subclass was added to represent groups of people or organisations.

4.5.1.3 Class: “Event”

This is a main class covering all the activities and events that take place in the cyber security domain. An event can be attended by an “Individual” or any of its subclasses. Some of the subclasses under “Event” are: “Conference”, “Expo”, “Workshop”, “Awareness Event” and “Training Event”.

4.5.1.4 Class: “Vulnerability”

This main class covers vulnerabilities that can be exploited by attackers (ISO, 2018). Three subclasses were created under this class: “Dataset”, which refers to existing vulnerabilities databases such as CVE (Common Vulnerabilities and Exposures)⁴

⁴<https://cve.mitre.org/>

and CWE (Common Weakness Enumeration)⁵; “Software”, which covers different categories of software vulnerabilities, e.g., “OS (Operating System)”, “Application” or “Web Server”, and “Hardware”, which covers hardware-related vulnerabilities.

4.5.1.5 Class: “Threat”

This main class is about “potential causes of an unwanted incident, which may result in harm to a system or organisation” (ISO, 2018). According to the nature of a “Threat”, these subclasses were created: “Criminal”, “Technical”, “Business”, “Legal”, and “Other”.

4.5.1.6 Class: “Attack”

This main class is about “attempts to destroy, expose, alter, disable, steal or gain unauthorised access to or make unauthorised use of an asset” (ISO, 2018). This main class covers a wide range of cyber security attacks mapped to the following subclasses: “Physical Attack”, “Software Attack”, “Network Attack”, “Social Engineering”, “Data Breach”, and “Unauthorised Access”.

4.5.1.7 Class: “Technical”

This is a main class that covers technical concepts such as “Cryptography”, “Protocol” and “Standard”. Under “Cryptography” there are “Encryption” and “Hashing” subclasses. “Encryption” contains well-known encryption algorithms (e.g., “DES”, “AES”, “Diffie-Hellman” and “RSA”). The same applies to “Hashing”, with several subclasses added beneath it to represent hashing algorithms (e.g., “BSD”, “MD5”, “SHA-1”, “SHA-256”) and related concepts such as “salting” and “rainbow table”.

⁵<https://cwe.mitre.org/>

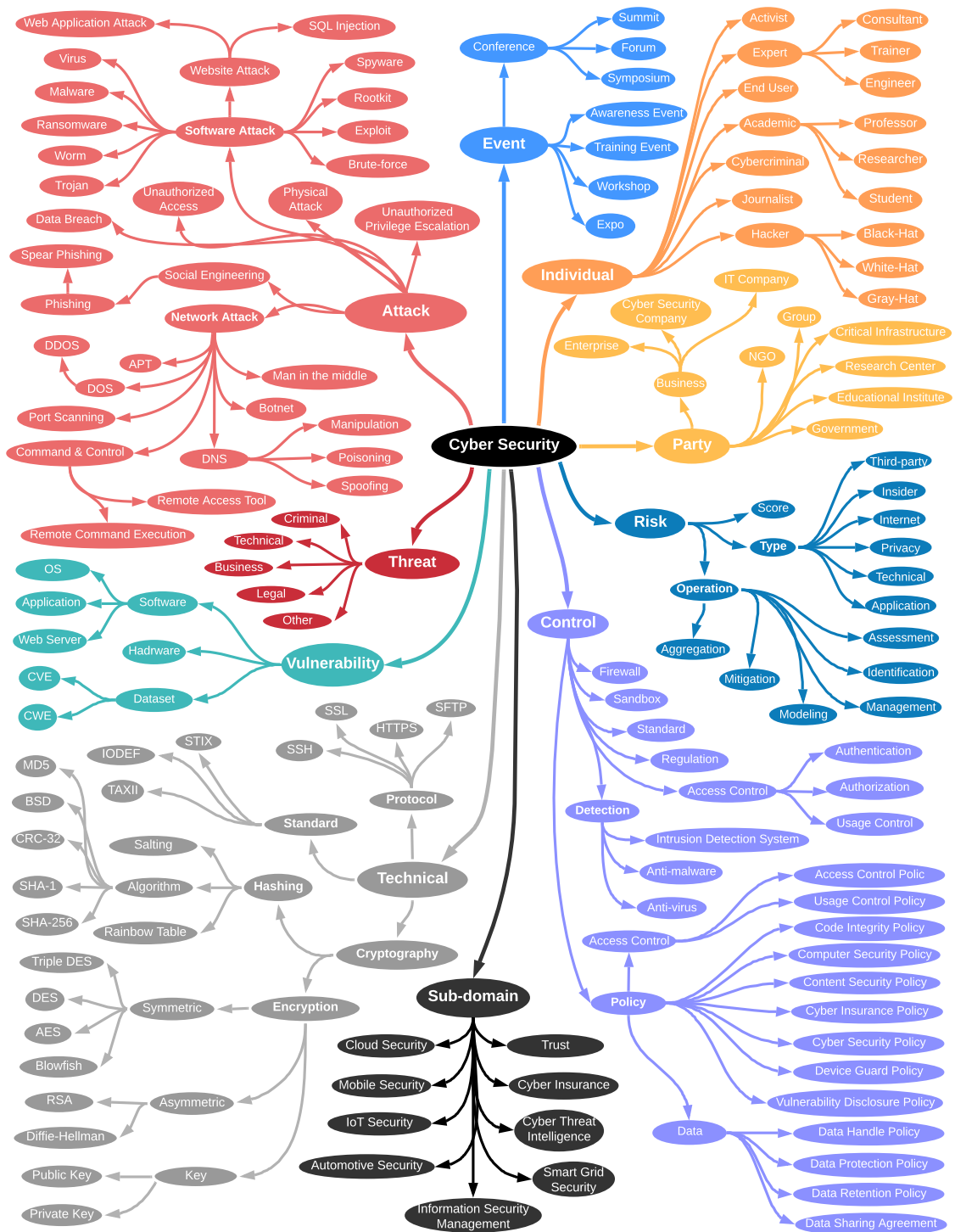


Figure 4.2: A visualisation of the initial cyber security taxonomy structure

4.5.1.8 Class: “Control”

Security “**Control**” is a main class that refers to any measure or actions that can be taken to reduce risk. Control mechanisms contain processes, policies, devices, practices, and other actions that can alter risk (ISO, 2018). Some of the “Control” subclasses are: “Firewall”, “Access Control”, “Standard”, “Policy”, “Regulation”, “Training”, “Detection”, and “Sandbox”. The “Policy” subclass contains more than 15 subclasses representing different kinds of policies such as data protection policies, access control policies and other general policies.

4.5.1.9 Class: “Risk”

This is a main class that covers concepts related to cyber risks. We defined a subclass named “Type” to reflect the nature of a risk, e.g. “Application”, “Insider”, “Internet”, “IoT”, “Privacy”, “Technical”, “Third-party”. Usually, a risk has a numeric value to quantify it. Thus, we added a subclass named “Score”. Additionally, we added the “Operation” subclass to cover all the common operations that are usually associated with cyber risks, e.g., “Aggregation”, “Assessment”, “Identification”, “Management”, “Mitigation” and “Modeling”.

4.5.1.10 Class: “Sub-domain”

This main class covers cyber security topics that can each have its sub-taxonomy, such as “Cloud Security”, “Mobile Security”, “IoT Security”, “Automotive Security”, “Smart Grid Security”, “Information Security Management”, “Trust”, “Cyber Insurance” and “Cyber Threat Intelligence”.

4.5.2 Assigning Terms to Taxonomy Structure

Each term produced in the text analysis stage was examined manually to assign it to the right class (or subclass). In some cases, a new subclass was added to accommodate a term, which means refining the taxonomy structure (to be explained

in the next subsection). During this step, any different spellings or synonyms for the same term should be considered and grouped together. Each term has a list of n -grams that represent different spellings that a term may have. For example, the term “cyber security” has a list of the following words with the same meaning: “cyber security”, “cybersecurity”, “cyber-security”. Another example is the “zero-day” attack, where the following words have the same meaning: “zero_day”, “zeroday”, “zero-day”, “0day”, and “zeroday”. Also, throughout this step, we distinguish a class (or subclass) from its members. For example, if “Ransomware” is used as a class, then ransomware attacks such as “WannaCry”, “NotPetya”, “Bad Rabbit” and “Locky” should be made members of that class.

Term	DC	TF	Idf	Tfidf	Node	Relation Type	Flag	
malware	409	19048	0.69	13313.98	Malware	Linguistically_Alike	Assigned	View Edit Unassign
cybersecurity	305	9562	0.69	6683.55	Cyber Security	Linguistically_Alike	Assigned	View Edit Unassign
phishing	288	3469	0.39	1380.45	Phishing	Linguistically_Alike	Assigned	View Edit Unassign
botnet	255	2651	0.39	1054.93	Botnet	Linguistically_Alike	Assigned	View Edit Unassign
ddos	252	2106	0.39	838.06	DDOS	Linguistically_Alike	Assigned	View Edit Unassign
ransomware	224	5321	0.39	2117.43	Ransomware	Linguistically_Alike	Assigned	View Edit Unassign
antivirus	219	1329	0.39	528.86	Anti-virus	Linguistically_Alike	Assigned	View Edit Unassign
data breach	208	2030	0.39	807.81	Data Breach	Linguistically_Alike	Assigned	View Edit Unassign
cybercrime	190	2008	0.39	799.06	Cybercrime	Linguistically_Alike	Assigned	View Edit Unassign
ddos attack	176	962	0.39	382.81	DDOS	Linguistically_Alike	Assigned	View Edit Unassign
kaspersky	172	1800	0.39	716.29	Cyber Security Company	is-member-of	Assigned	View Edit Unassign
blackhat	166	1649	0.39	656.20	Cyber Security Company	is-member-of	Assigned	View Edit Unassign

Figure 4.3: Assigning terms to taxonomy classes

Saving the list of the different n -grams each term can possibly have is an important addition and a key feature for any taxonomy built using our data-driven human-machine teaming approach for several reasons. First, the experts need to check the n -grams behind each term while assigning terms to classes. This is important and only the human expert can do it, not the machine. The human expert

needs also to revisit the n -grams list to decide if the term should be moved to another class or be merged with another term. Second, the human expert might modify or extend the n -gram list if a different keyword for an existing term in the taxonomy is later found. Third, the researchers and potential users of the taxonomy will benefit from knowing the other different keywords that a term or concept can have, especially in a diverse and fast-evolving domain list cyber security. Fourth, supporting a multi-lingual taxonomy structure can be simply achieved by having multiple n -gram lists beneath each term, each for a language. Finally, and most importantly, the taxonomy terms and their n -gram lists will be used in the ML classifiers in Chapter 5 and the analytical study in Chapter 6.

To streamline the process of creating the taxonomy and minimise human error, we developed a user-friendly application, see Figure 4.3. This application was designed to assist human experts in assigning terms efficiently to the appropriate categories within the taxonomy. By utilising this application, we aimed to save time for the experts and ensure accuracy in the term assignment process.

4.5.3 Taxonomy Refinement

While assigning terms to their potential classes (and subclasses), changes to the taxonomy's basic structure may be necessary. A class may need renaming or moving from one place to another to improve the semantic hierarchy of the proposed taxonomy. In addition, merging two or more classes or splitting a class could also happen. Such changes are normally done in an embedded manner as part of the terms assignment step, so these two steps often work in parallel. The refinement process can also be applied to the terms themselves because mapping terms to the taxonomy structure can change how the human analyst understands and organises them. For instance, some terms may be discarded, and some new terms can be added to classes with fewer children.

After following all the steps, we managed to build a general cyber security taxonomy. Since the cyber security taxonomy is based on a selected set of textual

documents, it is actually not “complete” and more like a baseline subset of what a “full” taxonomy can be. For instance, the automatically produced terms contain the names of some well-known cyber security experts, but many others were not included. Similarly, the names of some cyber security attacks and data breaches were captured, but many were not. Therefore, the built taxonomy should be further refined by using more textual documents, other existing taxonomies, and manually added classes and terms based on the domain knowledge of a human expert. The evolution of the cyber security domain also requires the taxonomy to be dynamically updated by re-running the proposed process with new textual documents regularly.

To some extent, the built (incomplete) taxonomy can be considered a good guideline to allow the human analyst to find ways to enrich the taxonomy further, e.g., after seeing a small number of terms for a specific concept (e.g., encryption algorithm, cyber security company, and cyber security expert, to name a few), the human analyst can systematically look for more relevant terms and consider how to refine the taxonomy’s structure and content.

4.6 Taxonomy Visualisation

4.6.1 Initial Taxonomy Structure Visualisation

An overview of the high-level structure of the initial cyber security taxonomy is shown in Figure 4.2. The diagram contains over 170 objects showing the main classes and subclasses of the taxonomy before going through the terms assignment step. The high-level structure was repeatedly refined while terms were allocated. This visualisation provides a simple way to understand the hierarchy of the proposed taxonomy quickly.

The root node of the taxonomy is the “Cyber Security”, which acts as the foundation for all the primary classes at the top level. These main classes are distinguishable by their larger node size and the bold styling of their internal text. Each main class is interconnected with its respective subclasses, forming a hierarchical struc-

ture. The subclasses themselves may branch out into further subclasses, creating a layered hierarchy within the taxonomy.

To enhance readability and facilitate ease of navigation and comprehension, each main branch, extending from a top-level class to its individual branches, was assigned a distinct colour. This colour-coded system aids in quickly identifying and locating all associated classes and subclasses within the taxonomy.

4.6.2 Final Taxonomy Structure Visualisation

While engaged in the process of term assignment and refining the taxonomy structure, we recognised the need for an interactive real-time visualisation tool to facilitate the actions performed by human experts. To this end, we developed a comprehensive visualisation interface that dynamically represents the evolving taxonomy.

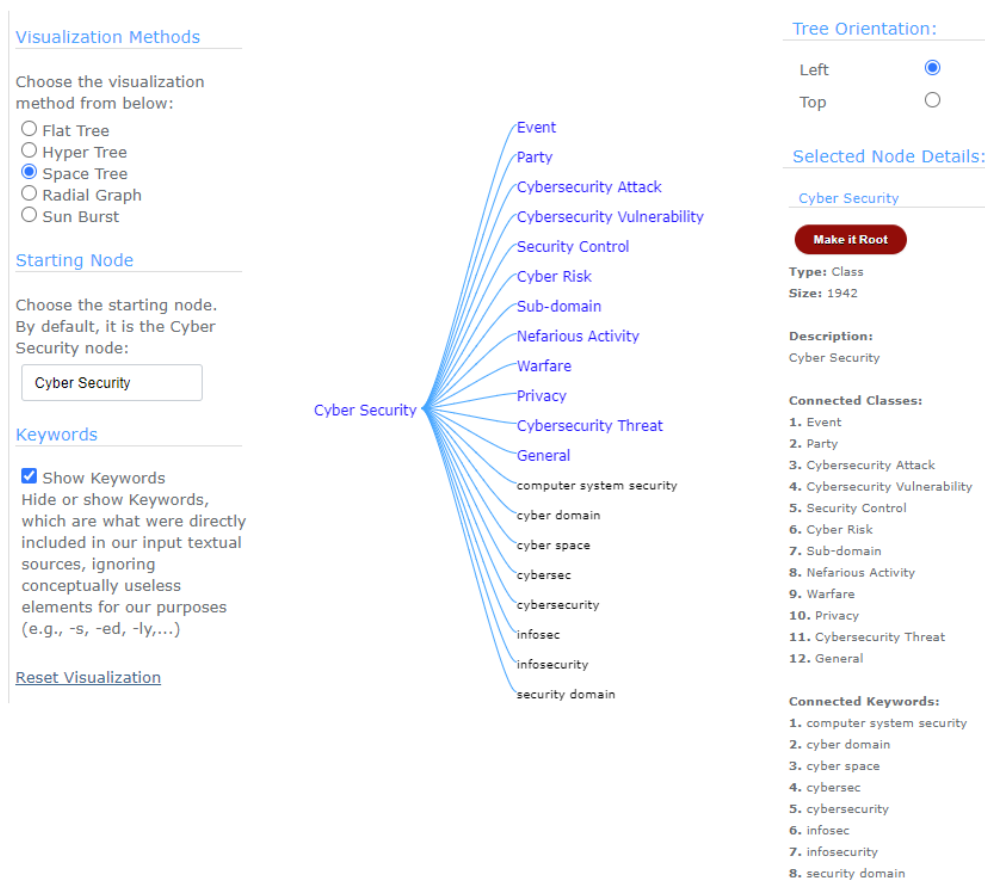


Figure 4.4: Visualisation interface: different visualisation methods and features

Five distinct visualisation methods were employed, each offering a unique perspective of the taxonomy: Flat Tree (Figure 4.5), Sun Burst (Figure 4.6), Space Tree (Figure 4.7), Radial Graph (Figure 4.8), and Hyper Tree (Figure 4.9). These methods afford users the ability to swiftly comprehend both the local and overarching structure of the taxonomy. For an enhanced viewing experience, we recommend exploring the interactive visualisations on the taxonomy webpage, see Appendix A.

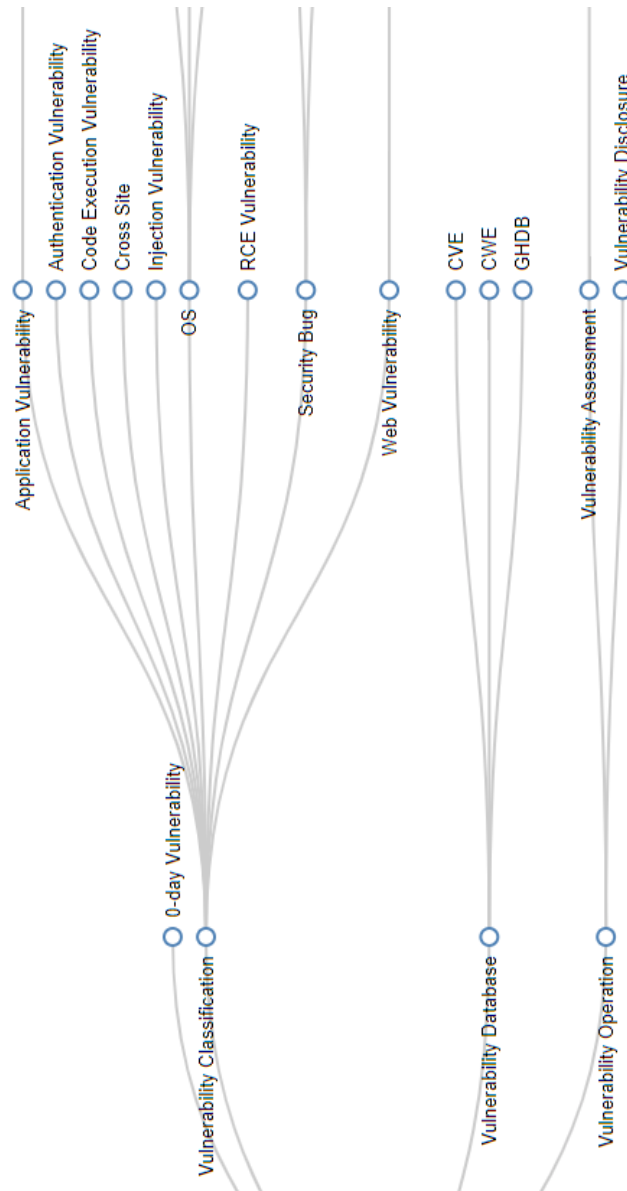


Figure 4.5: Visualisation [Flat Tree], showing the “Vulnerability” branch

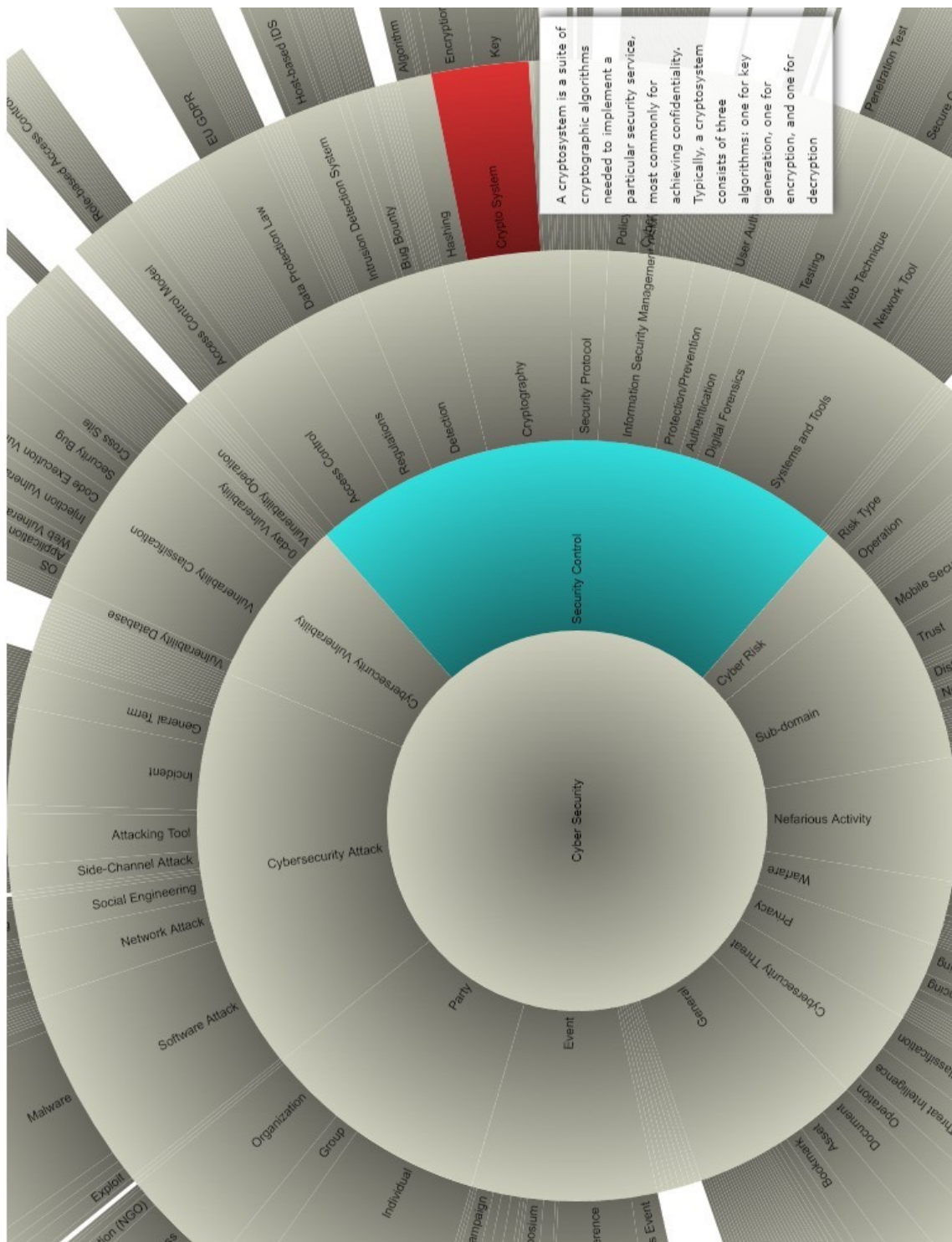


Figure 4.6: Visualisation [Sun Burst], showing the “Security Control” branch

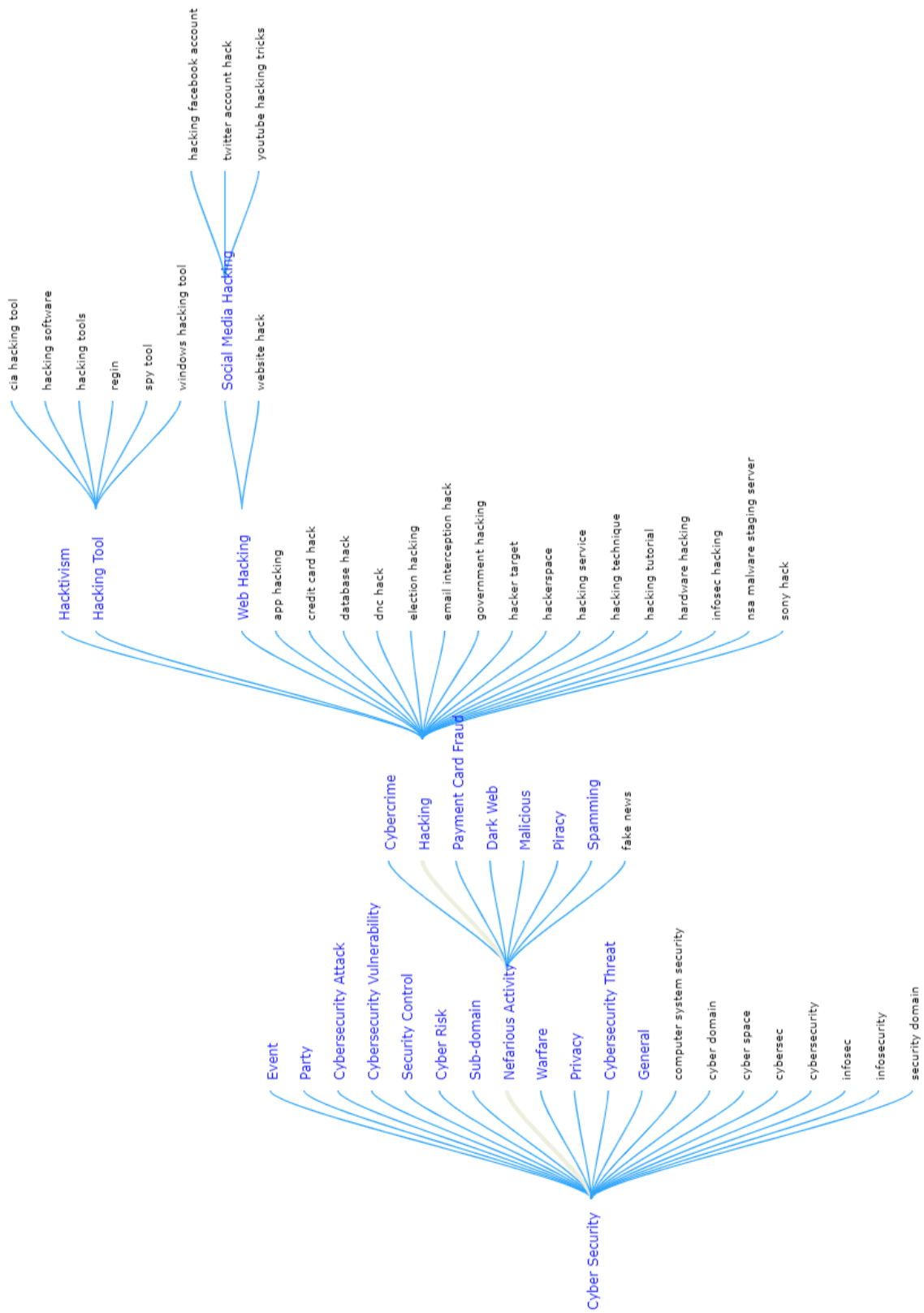


Figure 4.7: Visualisation [Space Tree], showing the “Hacking” branch

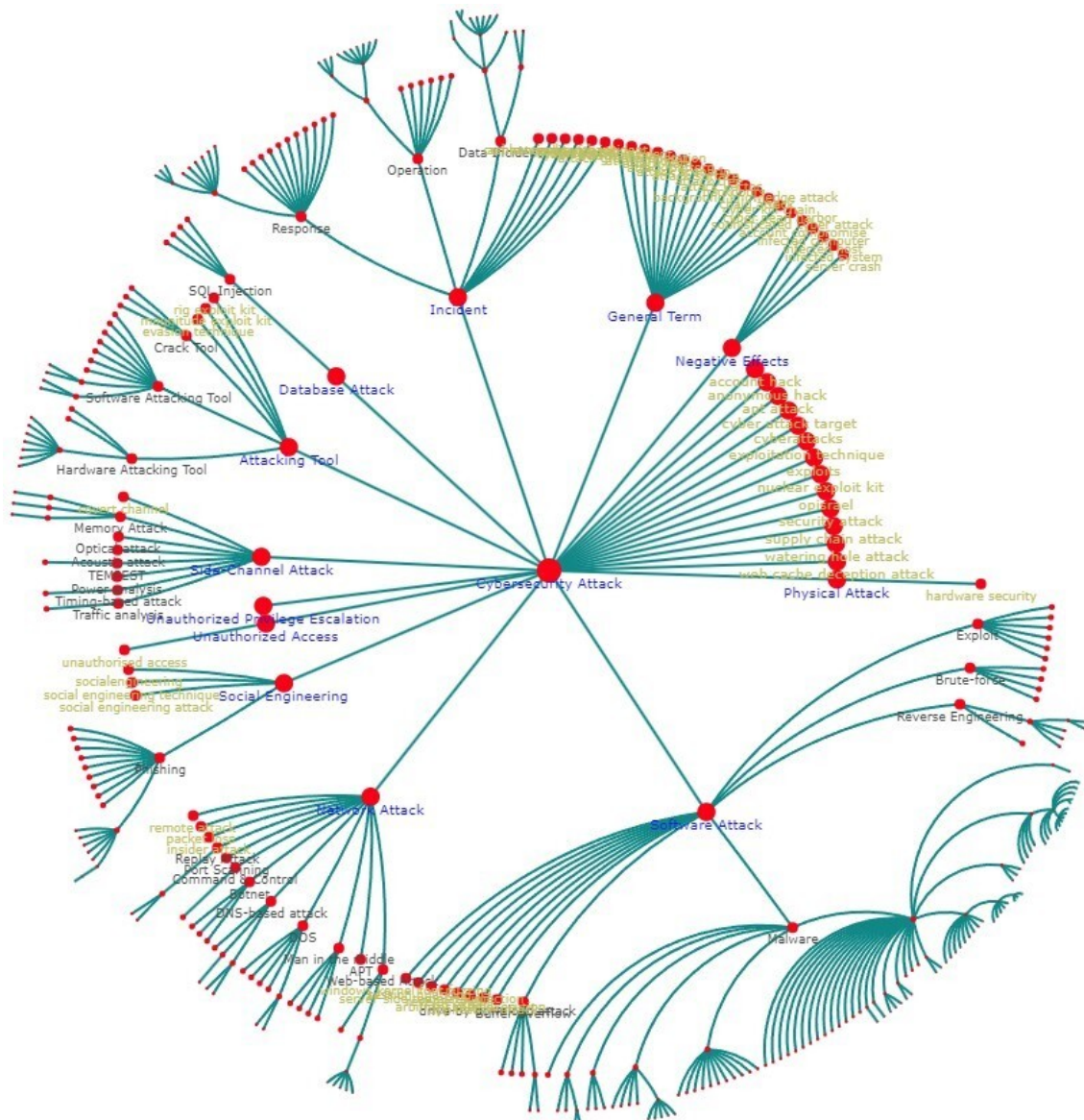


Figure 4.9: Visualisation [Hyper Tree], showing the “Attack” branch

4.7 Potential Applications

The created general cyber security taxonomy has various notable applications where it can be utilised. First of all, the taxonomy can be used to capture cyber security related discussions on OSNs. This can be achieved by analysing OSN feeds like tweets to determine if those tweets are related to the cyber security domain and then

identify the topics and concepts that are discussed. This also can help in building monitoring applications for OSNs with security purposes e.g., monitoring the spread of new malware on the internet and its impact on people and organisations. Of course, a more advanced version of such a monitoring system can be built using topic modelling.

Using the cyber security taxonomy, we can analyse Twitter account timelines to determine if these accounts are related to the cyber security domain. Also, we can determine which cyber security related concepts they are interested in. Such analysis can help understand human behaviour on OSNs for security purposes, e.g., cyber security awareness campaigns. Moreover, the taxonomy can be used to select a set of keywords as features for a machine learning classifier to automatically classify cyber security related people into different classes, which can provide helpful information about cyber security activities, such as impending or fresh attacks and first responses.

The taxonomy can further be used to analyse cyber security related textual sources in a semantic way, leading to better systematic analysis for such sources. One interesting application would be connecting such semantic analysis with eye-tracking data to understand how human users understand cyber security related documents such as privacy policies and security warnings. Also, the taxonomy can be used for cyber threat intelligence, where it can be employed to classify and analyse cyber threats, providing a structured framework for organising and understanding different types of threats. Last but not least, during incident response activities, the taxonomy can assist in categorising and prioritising incidents based on their nature and severity, enabling more efficient and targeted response efforts.

4.8 Taxonomy Validation

In taxonomies, the entities that can potentially be evaluated include (i) the process of developing the taxonomy, (ii) the outcome products of this process, which are the taxonomy itself and any other artefacts like user documentation, and (iii) the

resources required to create or utilise the taxonomy (Unterkalmsteiner and Adbeen, 2023). While there exist several guidelines and studies for developing taxonomies, there is a lack of actionable criteria to compare taxonomies. Also, we found that there is less agreement in the literature on the quality attributes that make a taxonomy “good” or effective (Usman et al., 2017; Szopinski, Schoormann, and Kundisch, 2020; Unterkalmsteiner and Adbeen, 2023). For example, (Ralph, 2019) created methodological guidelines for process theories and taxonomies in the Software Engineering field. He considered good taxonomies according to three characteristics: 1) the class structure allows for the differentiation between instances, (2) relevant properties of an instance can be deduced from its class membership, and (3) the taxonomy serves the purpose for which it was built.

Szopinski, Schoormann, and Kundisch reviewed 54 publications focused on developing and evaluating taxonomies within information systems research. Through this review, they identified 43 distinct quality attributes for taxonomies, such as Usefulness, Comprehensiveness, Conciseness, Explanatory, Extensibility, Robustness, and Applicability (Szopinski, Schoormann, and Kundisch, 2020). Also, Usman et al. conducted a review of 270 studies within the field of software engineering, focusing on the development and proposal of taxonomies. They found that 66% of the taxonomies underwent validation through methods such as illustration, case study, experiment, expert opinion, or survey (Usman et al., 2017). The aim of this validation was to showcase utility, which aligns with the quality attributes of Usefulness and Applicability identified by (Szopinski, Schoormann, and Kundisch, 2020). Some of the quality attributes used by (Usman et al., 2017) are Applicability, Comprehensiveness, Reliability, Usefulness, Orthogonality (Mutual exclusiveness), Conciseness, Robustness, and Extensibility.

Validating the output taxonomy, which was the direct application of the presented methodology in this chapter, is not a straightforward task. Thus, we decided to compare it with another “diverse” or general taxonomy in the cyber security domain as we could not arrange an external review by independent researchers or

industry professionals. For this comparison, we chose the European Cybersecurity Taxonomy (Nai Fovino et al., 2019) which seemed to be the best match since it aims at aligning definitions and terminologies to support the classification of cyber security competencies. As for the quality attributes used in the comparison, we decided to choose the most commonly used attributes for evaluating taxonomies found in the literature and the ones that were considered objectives and “internal” as reported in (Unterkalmsteiner and Adbeen, 2023). Unterkalmsteiner and Adbeen consider internal attributes the ones that can be measured solely by examining the product itself, whereas external attributes can only be observed in relation to the product’s behaviour within a specific environment.

Based on the aforementioned studies, we chose the following quality attributes: Comprehensiveness, Robustness, Conciseness, Extensibility, Explanatory, and Mutual exclusiveness. The detailed comparison between our taxonomy and the taxonomy built by (Nai Fovino et al., 2019)⁶ was already done and reported by (Unterkalmsteiner and Adbeen, 2023). However, we provided more details as there were missing data in their comparison.

- **Comprehensiveness:** Our taxonomy was built for a purpose which is to classify cyber security related discussions in OSNs. We set a new methodology focusing on further automation by boosting the human-machine teaming. Three researchers worked on it with two from the same university with a cyber security background and the third one from a different university with a software engineering background. As for (Nai Fovino et al., 2019), the authors established also use cases for their taxonomy, which is classifying cyber security competencies of organisations within the EU. Their team involved seven researchers from different universities in the EU. However, both taxonomies had no involvement of industry professionals.
- **Robustness:** To report robustness, we used the measurement proposed by (Unterkalmsteiner and Adbeen, 2023). Our taxonomy scored 0.57 compared to only

⁶Downloaded from (European Commission, Joint Research Centre (JRC), 2021)

0.31 for (Nai Fovino et al., 2019) taxonomy. This is due to the relatively large number of categories (425) and characteristics (1529) created in our taxonomy compared to 18 categories and 188 characteristics in (Nai Fovino et al., 2019).

- **Conciseness:** As explained by (Nickerson, Varshney, and Muntermann, 2013), robustness and conciseness can both be measured by the dimensions, categories and characteristics of a taxonomy. However, conciseness decreases when robustness increases. We used the conciseness measure proposed by (Prat, Comyn-Wattiau, and Akoka, 2015), where our taxonomy scored 0.14 while the other taxonomy, i.e. (Nai Fovino et al., 2019), scored 0.19. We acknowledge that a higher number of categories and characteristics depends on the goal or intended use case(s), and is not necessarily always good. Nickerson, Varshney, and Muntermann suggested that an extensive classification scheme, with numerous dimensions and characteristics, has the potential to overwhelm the cognitive capacity of the researcher, thereby making it challenging to comprehend and apply effectively.
- **Extensibility:** (Nai Fovino et al., 2019) did not explain how to change their taxonomy. For our taxonomy, the proposed methodology which is characterised by two important features, the human-machine teaming and the data-driven, supports changing taxonomy at any time. This was explained in Section 4.5.3.
- **Explanatory:** The taxonomy created by (Nai Fovino et al., 2019) contained description for their dimensions and categories. For our taxonomy, we added descriptions only for the main (top-level) categories, as we had 425 categories in total. We plan to add descriptions for all categories in future releases.
- **Mutual exclusiveness:** This was considered from the beginning in the design of our taxonomy. However, we did not enforce it to give the flexibility to the researchers and practitioners to classify an entity under more than one category. This was needed clearly for the “Individual” main category as the same person may fit into two categories. We are planning to further restructure the taxonomy to eliminate

such cases. As for the taxonomy created by (Nai Fovino et al., 2019), no data was found regarding mutual exclusiveness.

As a conclusion for the comparison, if we are interested in Robustness, Extensibility and Mutual exclusiveness attributes, then our taxonomy excels and should be the one to use, while if Conciseness is more important, then the taxonomy created by (Nai Fovino et al., 2019) is better.

Although the suggested methodology presented in this chapter aids human experts and reduces the effort needed, the taxonomy structure itself and how extracted terms are assigned to taxonomy classes is a subjective task determined by the expert executing the task. In other words, two experts with the same set of input textual sources and extracted terms can produce utterly different output taxonomies. After all, a taxonomy is a representation of knowledge in a given domain as seen by domain experts. At the same time, we acknowledge the need for an external review of our taxonomy, especially from industry professionals, as the feedback from such reviews can potentially lead to a huge improvement which can increase the reliability and applicability of this taxonomy.

Finally, while there may be valid criticisms of building a “general” cyber security taxonomy, we believe it is a valuable tool for organising and understanding the complex landscape of cyber security threats and activities. Its broad applicability and flexibility make it a useful resource for researchers, practitioners, and policymakers in the cyber security field.

4.9 Conclusion

In this chapter, we presented a human-machine teaming-based process to build taxonomies, starting from a given set of textual documents, followed by mostly automated processing done by NLP and IR tools, which produce a list of relevant terms to be assigned to a defined taxonomy structure. The key feature of the proposed process is the higher level of automation, which helps reduce human effort

and make the selection of relevant terms more data-driven (less subjective). The process also allows any built taxonomy to be maintained more easily. An example is given to show how a general taxonomy was constructed using this process for the cyber security domain. The example taxonomy was built with a reasonable amount of human effort and a large number of candidate terms automatically collected from multiple data sources, which can be extremely time-consuming and more error-prone if done solely by humans.

Chapter 5

Automatic Detection of Cyber Security Accounts on OSNs

“The ability to perceive or think differently is more important than the knowledge gained.”

- David Bohm

MANY cyber security experts, organisations and cyber criminals are all active users of OSNs. Given their diverse presence, the ability to detect cyber security related accounts on OSNs and monitor their activities holds significant value for various purposes. One key application is in the domain of cyber threat intelligence, where by identifying and tracking cyber security related accounts on OSNs, valuable insights can be gained into potential threats, emerging trends, and malicious activities. This information can contribute to the development of proactive measures to detect and prevent cyber attacks and mitigate online harm to OSNs.

Furthermore, monitoring the activities of cyber security related accounts on OSNs enables the evaluation of the effectiveness of cyber security awareness activities conducted on social media platforms. By analysing the engagement, reach, and impact of such accounts, organisations and researchers can assess the efficacy of their awareness campaigns, identify areas for improvement, and tailor their strategies to

disseminate cyber security information to a wider audience effectively. Overall, the detection and monitoring of cyber security related accounts on OSNs offer significant benefits in terms of enhancing cyber threat intelligence, safeguarding against cyber attacks, mitigating online harms, and evaluating the effectiveness of cyber security awareness initiatives.

In this chapter, we report our work on developing a number of machine learning based classifiers for detecting cyber security related accounts on Twitter, including a baseline classifier for detecting cyber security related accounts in general, and three sub-classifiers for detecting three subsets of cyber security related accounts, individuals, hackers, and academia, respectively. To train and test the classifiers, we followed a more systemic approach (based on a cyber security taxonomy, real-time sampling of tweets, and crowdsourcing) to construct a labelled dataset of cyber security related accounts with multiple tags assigned to each account. We considered a richer set of features for each classifier than those used in past studies. Among the five ML models tested, the Random Forest model achieved the best performance, 93% for the baseline classifier, and 88-91% for the other three sub-classifiers. We also studied feature reduction of the baseline classifier and showed that we can already achieve the same performance using just six features.

5.1 Introduction

The rise in the development of internet-based technologies creates new vulnerabilities every day. Hackers are always up to date with these technologies and consistently exploit their vulnerabilities to use them either for their benefit (black-hat hackers) or to patch the security holes (white-hat hackers). Other types of hackers, which are hacktivists, on the other hand, hack for a cause. In any case, the severity and impact of these types of attacks are experiencing a rapid escalation. Consequently, cyber security experts are persistently engaged in proactive measures to prevent such attacks and effectively mitigate the resulting damage. They are constantly work-

ing towards implementing robust security measures, developing advanced defence mechanisms, and staying updated with emerging cyber threats.

OSNs have become part of everyday life for many people. As internet experts, both cyber security experts and cyber criminals are among the active users of OSNs. Cyber security professionals use OSNs for different purposes, such as knowledge exchange, cyber security awareness, and offering help to people and organisations on cyber security matters. On the other hand, cyber criminals often utilise OSNs to reach out to victims, boast about their past and new “achievements”, and even talk about their future attack plans.

The activities of cyber security professionals and criminals have been found to be a good source of information for many purposes, such as cyber threat intelligence and understanding behaviours of cyber criminals and related groups (Lippmann et al., 2017; Jones, Nurse, and Li, 2020; Aslan, Li, et al., 2020). Furthermore, studying cyber security “related” accounts on OSNs is vital to understanding the different networks of those accounts, e.g. hackers, experts..etc. Thus, finding – and continuously monitoring – cyber security related accounts on OSNs requires automatic detection of such accounts.

Furthermore, it would be advantageous to determine whether a cyber security account belongs to an individual or a group/organisation. This distinction would enable a more advanced analysis of these accounts, such as studying the personality traits of individuals, which is only applicable and valuable in the case of an individual account rather than a group account managed by multiple individuals. Additionally, a crucial and significant next step would involve detecting whether an account exhibits characteristics of a hacker or belongs to the academic sector. For instance, being able to predict whether an account demonstrates hacker-like behaviour could potentially provide insights into the malicious activities associated with that account. Similarly, predicting whether an account belongs to academia, industry, or any other domain can be correlated with their behaviour and the content they create or share.

There has been some recent research about the use of machine learning to detect whether a Twitter account is cyber security related or not (Aslan, Sağlam, and Li, 2018) or to detect if an account belongs to a specific hacktivist group, e.g., Anonymous (Jones, Nurse, and Li, 2020). However, such studies are still limited in their generalisability and validation of performance. In addition, there is a lack of more general-purpose sub-classifiers that can classify different sub-groups of cyber security related accounts, e.g., cyber security individuals (vs. groups and organisations), hackers in general (both people and groups), researchers and research organisations, etc. Developing sub-classifiers enhances the precision of monitoring different sub-groups, allowing for more targeted observation and behavioural analysis.

5.2 Methodology

The study presented in this chapter addresses various challenges associated with classifying cyber security related accounts on Twitter, where it is based on a three-staged methodology: a more systematic data collection process, crowdsourcing-based labelling experiment, and the development of machine learning based classifiers. Following similar work of other researchers, especially (Aslan, Sağlam, and Li, 2018), we designed a general methodology for detecting a specific type of accounts on Twitter based on textual data, which can be easily applied to any other OSNs.

The methodology consists of three main stages, as illustrated in Figure 5.1. The first stage, Stage A, is Data Collection, where raw (i.e., unlabelled) data about a reasonable number of accounts is collected from the target OSN platform. Moving to the second stage, Stage B, where the focus is on Dataset Construction. Here, we employ a crowdsourcing-based approach to enlist the expertise of cyber security experts in assigning ground truth labels to the OSN accounts selected during Stage A. This crowdsourcing process ensures reliable and accurate labelling, providing a solid foundation for the next stage.

The third stage, Stage C, centres around Building Classifiers utilising the labelled

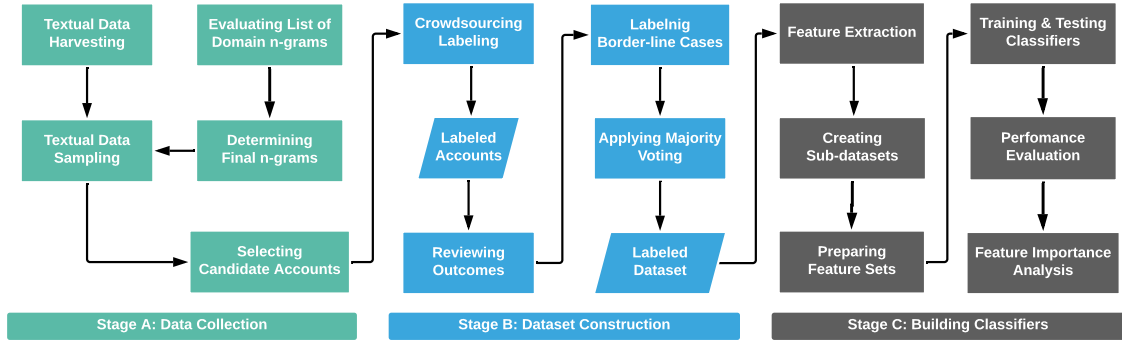


Figure 5.1: Proposed methodology for detecting cyber security accounts

dataset from Stage B. We followed the standard steps for developing supervised machine learning classifiers: feature extraction, training and testing of classifiers, performance evaluation, and finally, feature importance analysis to reduce the dimensionality of the features used. After feature extraction, two additional steps are incorporated: creating sub-databases for multiple classifiers and preparing various candidate feature sets to identify the optimal ones for each classifier.

By following this methodology, we aim to effectively detect the targeted account type(s) by utilising the expertise of cyber security professionals and leveraging ML techniques. The process ensures robust classification and facilitates the application on other OSNs. Throughout this chapter, we focused on Twitter as an example OSN platform because we observed more active cyber security related accounts on this platform. The raw data – collected in this case – includes user profile data and timeline (both are user-generated textual data). The following sections will give details of the three stages, respectively.

5.3 Data Collection

In this section, we comprehensively describe the data collection process undertaken to acquire the necessary data for our research. Additionally, we outline the method that was employed to do the tweet sampling and then select the candidate accounts that were used in the labelling experiment and the subsequent steps.

5.3.1 Harvesting Tweets

Researchers have used various methods to collect data from Twitter. Some studies manually identified a seed set of Twitter accounts related to their research focus, e.g., (Jones, Nurse, and Li, 2020), while others utilised public lists created by Twitter users with relevance to the target area of interest, e.g., (Aslan, Sağlam, and Li, 2018). In the work of (Pennacchiotti and Popescu, 2011b), they constructed a dataset by leveraging public Twitter directories such as Twellow and WeFollow, which provided access to well-labelled Twitter accounts. Researchers need to consider the pros and cons of each data collection method and select the approach that aligns with their research objectives and ensures the inclusion of diverse and representative data. By employing a comprehensive and thoughtful data collection strategy, researchers can enhance the validity and reliability of their studies conducted on Twitter.

Many studies in the literature have utilised the Twitter Search API¹ to search for tweets and users based on a curated list of relevant keywords, e.g., (Preoțiuc-Pietro et al., 2015). While such an approach offers convenience, it may also introduce potential limitation(s). The reliance on keyword-based searching can lead to the inclusion of less representative samples, which can impact the overall generalisability of the research findings. Considering the limitations of the Twitter Search API, we decided to use the Twitter Sampling API to collect a set of tweets that include cyber security related discussions, from which we can further identify cyber security related Twitter accounts.

Table 5.1: Statistics of Tweets Sampling step

Step	Count	Step	Count
All tweets	477,863,647	Unique tweets	1,140,228
Filtered tweets	1,146,588	Retweets tweets	602,126
Duplicates tweets	6360	Original tweets	538,102

¹<https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview>

We created a data harvesting tool that connects to the Twitter Sampling API endpoint and consumes tweets. According to Twitter², the Sampling API returns a random sample, percentage $\sim 1\%$, of all the public tweets on Twitter in real-time. We collected 478 million tweets over 16 months from January 2019 to April 2020, see Table 5.1 for data collection statistics.

5.3.2 Sampling Cyber Security Related Tweets

This step aims to select a subset of tweets from the harvested data collection in the previous step. Those tweets should be cyber security related tweets and thus more likely to be tweeted – or retweeted – by cyber security experts or at least cyber security related users. To achieve this, we must filter the tweets using a list of cyber security related terms (i.e. n -grams).

The required list was extracted from the general cyber security taxonomy we built in Chapter 5 and reported in (Mahaini, Li, and Belen-Sağlam, 2019). The taxonomy contains over 1,900 terms with their different wordings, e.g., (“Cyber Security”, “Cybersecurity”) and (“0day”, “zero day”). Thus, we compiled a list of 2,236 n -grams in total. Subsequently, we employed this list to search through our collection of tweets, specifically targeting the presence of these n -grams. As a result, we obtained search result statistics for each n -gram search. Then, we manually reviewed each n -gram aided by these search statistics, which enabled us to reduce the list of n -grams to just 795 by excluding n -grams that are duplicates, loosely cyber security related, implicit by other n -grams or have poor search results, see Table 5.2 for statistics about the selected n -grams. The final reduced list is quite important not just for this step but also will be used later in the features extraction step, see Section 5.5.

Using the final n -gram list, we applied a filtering procedure to the initial tweet data collection. The filtering aimed to narrow down the dataset to include only tweets that contained at least one cyber security n -gram. As a result of this filtering

²<https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams>

Table 5.2: Statistics of the n -grams list review process

Status	Reason	Count
Excluded	Low Results Tweets	824
Excluded	n -gram is implicit by another n -gram	290
Excluded	Duplicates n -grams	170
Excluded	Loosely Cyber Security Related n -grams	101
Excluded	Manually Eliminated n -grams	47
Excluded	Outliers	9
Accepted		795

process, the total number of tweets was significantly reduced to 1.1 million, which accounted for approximately 0.23% of the whole data collection.

Subsequently, we focused on extracting original tweets that were not retweets, as our main interest was analysing the content generated directly by the Twitter accounts themselves. To achieve this, we removed retweets from the dataset, which constituted 53% of the remaining tweets. Consequently, we retained only the original tweets, amounting to 47% of the filtered dataset. After completing these filtering steps, our dataset consisted of 538,102 tweets, which accounted for approximately 0.11% of the entire initial collection of harvested tweets. Further details and statistics regarding the tweet sampling process can be found in Table 5.1.

5.3.3 Selecting Candidate Accounts

We aimed to generate a good-sized labelled dataset to serve as a reliable ground truth for developing the required classifiers. This dataset would not only facilitate the construction of accurate classifiers but also provide valuable insights into the characteristics that define a Twitter account as cyber security related, as perceived by the participants in our labelling experiment (refer to Section 5.4.3). By leveraging the labelled dataset, we aimed to enhance our understanding of the key attributes

that contribute to the classification of a Twitter account as cyber security related. This dataset would serve as a valuable resource for training and evaluating our classifiers, allowing us to understand the nuances and characteristics associated with cyber security related accounts on Twitter.

We identified the accounts that posted the “original” tweets from the Tweets Sampling step, resulting in 57,018 unique accounts. After removing the suspended and deleted accounts (which were 9%), we ended up with 51,868 accounts from which to select a subset. Since the selected accounts would be manually labelled by cyber security experts and due to difficulties in the recruitment process, we set our target to obtain a labelled dataset between 1,300 to 1,400 accounts that were more likely to be labelled as cyber security related accounts in the labelling experiment. That means we needed to select 2.5% out of 51,868 Twitter accounts. Therefore, we set a criterion to filter the accounts first and then select the needed number of accounts. To achieve this, we introduced two simple measures to each Twitter account:

- TiD (the number of **T**weets in the **D**ata collection, i.e., the original tweets).
- KiD (the number of **K**eywords in the **D**escription field of the Twitter account).

The keywords list that was used here is the same one we used previously in the Tweets Sampling, refer to Section 5.3.2. Then, we set two criteria as follows:

- A) Twitter accounts that had at least two tweets in the original tweets data collection (i.e., $TiD \geq 2$), which means they had mentioned at least one cyber security keyword in their captured tweets.
- B) Twitter accounts that had at least one cyber security keyword (i.e., $KiD \geq 1$) mentioned in their profile description.

To determine the best criterion to use, we formed three test groups, each consisting of 100 accounts. The first group represents only criteria A alone, while the second one represents criteria B, and the last one, C, represents the intersection between A and B criteria. Each group was then manually labelled by checking

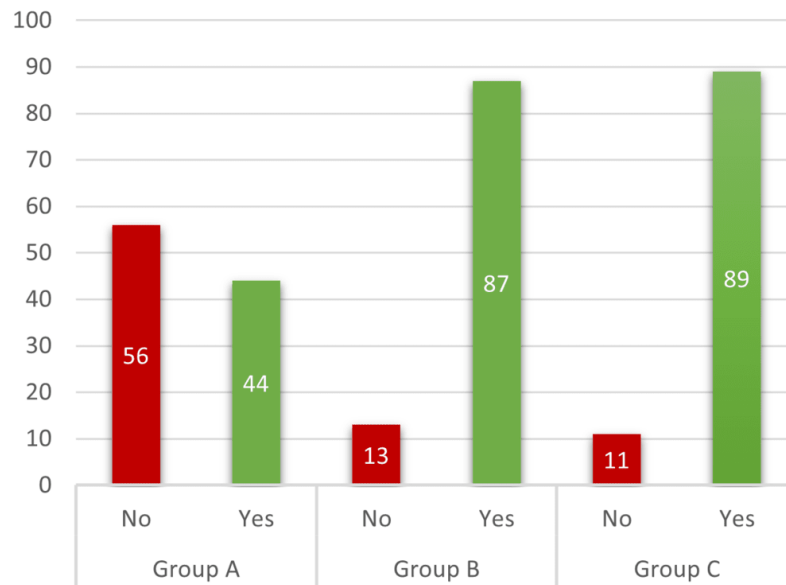


Figure 5.2: Accounts Selection Verification - Results

whether each Twitter account was cyber security related. The results of the three groups were 44, 87, and 89, respectively as shown in Figure 5.2. That means out of the three groups, Group C yielded the highest number of cyber security related accounts. Based on these results, we confirmed that a selection criterion based on the intersection of A and B criteria together would be better than A or B alone. After that, the 51,868 accounts were filtered and randomly selected 1,300 accounts to be used for the crowdsourcing labelling experiment.

The criteria employed served as a means to filter through the accounts, with criterion A identifying users who exhibited slightly higher levels of tweeting activity compared to others, while criterion B flagged those who included cyber security keywords in their account descriptions. The combined application of both criteria proved to be a more effective filtration method than selecting accounts randomly, thereby ensuring the relevance and validity of the subsequent labelling experiment.

5.4 Dataset Construction

In this stage of our methodology, we conducted a crowdsourcing-based labelling experiment, which involved recruiting a group of cyber security experts as human labellers. The primary objective of this experiment was two-fold:

1) **Construction of a labelled dataset:**

The participating cyber security experts were tasked with labelling a dataset of Twitter accounts, specifically focusing on identifying the cyber security related accounts. Additionally, they were asked to assign other relevant tags, such as Individual, Hacker, Academia, etc, to provide further context and categorisation for the accounts. This process resulted in the creation of a labelled dataset, which served as the ground truth for training and evaluating the ML classifiers.

2) **Gathering insights through a questionnaire:**

Alongside the labelling tasks, the participating experts were presented with a short questionnaire which was split into three parts. The first and second parts are shown to the participant before the labelling tasks, while the third appears afterwards. The questionnaire structure was as follows (see Appendix B for the actual questions):

Part 1, Basic Demographics: This part contains five questions about the participants' demographics: gender, age bracket, employment status, and education. Also, we added a question about their experience in the cyber security domain.

Part 2, Cyber Security Experts on OSNs: This part of the questionnaire aimed to gather valuable information regarding the experts' perspectives on the definition of "cyber security expert" and their observations of the behaviour exhibited by cyber security related Twitter accounts. The responses collected through this part of the questionnaire provided useful insights and guidance for defining better features that could enhance the performance of the intended ML classifiers.

Part 3, Labelling Process Feedback: The purpose of this part was to collect feedback about the labelling process and to validate the findings from the second part of the questionnaire, as the participant fills the second part before labelling the Twitter accounts and fills the third part afterwards.

By leveraging the knowledge and expertise of cyber security experts, the crowdsourcing based labelling experiment facilitated the creation of the needed labelled dataset(s) to train the ML classifiers that we wanted to build. Moreover, the questionnaire responses, especially the free-text answers, added a qualitative dimension to the process, enabling the incorporation of expert insights into the feature engineering process for the subsequent development of the intended classifiers.

5.4.1 Crowdsourcing Labelling

5.4.1.1 Crowdsourcing Labelling Application Implementation

We thoroughly explored existing crowdsourcing tools in our quest for an efficient and user-friendly labelling experience for the potential participants. However, none of the available tools or platforms fully met our specific requirements. We sought a solution that would allow for dynamic task allocation and monitoring while considering potential challenges, such as Twitter accounts being deleted or suspended before they could be labelled. Additionally, we aimed to provide participants with a labelling interface that was easy to navigate and offer on-screen controls to expedite the labelling process.

Given these unique requirements and the need to customise labelling controls to align with our specific objectives, utilising an existing crowdsourcing platform proved pointless. Therefore, we decided to develop a custom web-based Crowdsourcing Labelling Application (CLA) that would cater specifically to our needs. This approach allowed us to design a labelling interface that prioritised efficiency, ease of use, and flexibility. By building a tailored solution, we ensured that the labelling experience was optimised for our specific research goals and accommodated any potential obstacles that could arise during the labelling process.

The application was designed in collaboration with local cyber security experts and potential participants, taking into consideration their feedback and expertise. Additionally, we conducted extensive discussion groups involving over 20 local cyber security academics to gather valuable insights and ensure the system's effectiveness.

Twitter Labelling Tasks Open Tutorial


Tasks List | **Information Sheet: @emahayni**

Google Search

➊ Search Google for this Twitter account:
 Search


Twitter

➋ Check Twitter account's profile below or visit it by clicking here: [emahayni](#)


Avatar		Joined	September 02, 2009
Name	Emad Mahayni	Screen Name	emahayni
Followers	209	Friends	486
Statuses	96	Favourites	37
Listed	4	Location	Damascus
Website	https://uk.linkedin.com/in/emahayni		
Description	Early Stage Researcher working on #Cybersecurity and Social Networks.		

➌ Review Twitter account's timeline:

Tweets by @emahayni

 **Emad Mahayni**
@emahayni

Change your password and your underwear :)
[uk.businessinsider.com/facebook-secu...](#)



Facebook just announced it was hacked, and almost 50 million users hav...
It's not yet clear who was behind the attack that affected 50 million Facebook

Figure 5.3: Labelling Interface, Tasks List (left) & Information Sheet (centre)

To maintain the accuracy and quality of the labelling results, we implemented a

majority voting mechanism. Each account in the dataset was assigned to three different participants, and a consensus was reached through majority voting. This approach ensured that the labels assigned to the accounts were robust and reliable.

5.4.1.2 Crowdsourcing Labelling Application Overview

The system interface is divided into four main sections for an efficient and user-friendly labelling experience:

I) **Top** section: This section provides buttons that give participants access to the participant information sheet (PIS) and a tutorial page. The PIS provides participants with detailed information about the study, including its purpose, procedures, and any potential risks or benefits. The tutorial page serves as a guide to help participants navigate through the labelling process efficiently and effortlessly.

II) **Left** section: In this section, participants can view a list of all the accounts assigned to them for labelling. Initially, all the tasks are marked in orange. As participants complete the labelling for an account, its colour changes from orange to green. This list provides an overview of the progress, indicating the number of both finished and unfinished tasks. Participants can click on any task to revisit it and make any necessary edits to the labels or answers if they want to.

III) **Centre** section: This section displays the details of the Twitter profile for the current labelling task, including the account's timeline. Participants also have the option to use Google Search for more information about their current Twitter account. See Figure 5.3 for a screenshot of this section and the left section.

IV) **Right** section: The participant will utilise this section to label the current selected Twitter account. The interface provides the necessary controls and options for assigning the relevant labels and answering any associated questions. A screenshot of this section can be found below, see Figure 5.4.

The division of the system interface into these four parts aims to streamline the labelling process, providing participants with clear and organised access to labelling tasks and account information.

Labelling Sheet: @emahayni

Account type:	<input checked="" type="radio"/> Individual <input type="radio"/> Group or Organization
Is it a <u>Twitter bot</u> ?	<input type="radio"/> Yes <input checked="" type="radio"/> No
Is it Cyber Security related?	<input checked="" type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not Sure

Can you classify it?

Academia
Hacker
Industry
Company

Government
NGO
Other

Can you say more?

Lecturer
Researcher
Student
White-Hat Hacker

Gray-Hat Hacker
Black-Hat Hacker
Expert

Professional
Consultant
Journalist
Activist

Save Changes

Figure 5.4: Crowdsourcing Labelling Application, Labelling Sheet

5.4.2 Research Ethics and GDPR

As the labelling experiment involved recruiting participants and collecting some personal data (PI), the experiment had to go through a detailed review process by the Research Ethics Advisory Group (REAG) at the University of Kent. This is necessary to ensure that all research upholds integrity and is conducted according to the appropriate ethical, legal, and professional frameworks, obligations, and standards. This includes being compliant with the General Data Protection Regulation (GDPR), providing the essential GDPR rights to participants, explaining the purpose of the experiment and what data will be collected, and ensuring confidentiality.

5.4.3 Participant Recruitment

As we aimed to label 1,300 accounts, and since we adopted a majority voting system with three votes per account, we required a total of 3,900 labelling tasks to be completed. To accomplish this, we sought to recruit approximately 100 participants. Given the nature of the labelling task, which demanded considerable time and good cyber security knowledge, we decided to introduce a £15 Amazon UK voucher as compensation for their valuable contribution. The participants should have cyber security experience to qualify for the experiment. This experience could be either theoretical or practical, making individuals such as cyber security students, researchers, educators, consultants, and other types of experts eligible to participate.

We employed various strategies to reach out to potential participants for our experiment. We sent emails about the experiment to several cyber security professional networks, such as SPRITE+³, as well as to the mailing lists of some cyber security conferences' like FOSAD⁴. Additionally, we reached out to cyber security research groups and centres in the UK and extended invitations to professionals working in the industry or non-governmental organisations (NGOs). We manually verified each candidate before recruiting them as participants in the experiment to eliminate the risk of having unqualified individuals labelling part of the dataset and to ensure accuracy and quality. Due to the COVID-19 pandemic, the recruitment process took some time, and the experiment itself ran for approximately three months. Despite the challenges, we managed to recruit a total of 100 participants, but 90 of them finished the whole experiment and completed all the labelling tasks assigned to them.

The participants had diverse demographics and professional backgrounds related to cyber security. The demographics included gender, age, employment status, and level of education. The gender distribution among the participants varied, as females were 26% compared to 74% for males, see Figure 5.5b. Also, the age range

³<https://spritehub.org>

⁴<https://sites.google.com/uniurb.it/fosad>

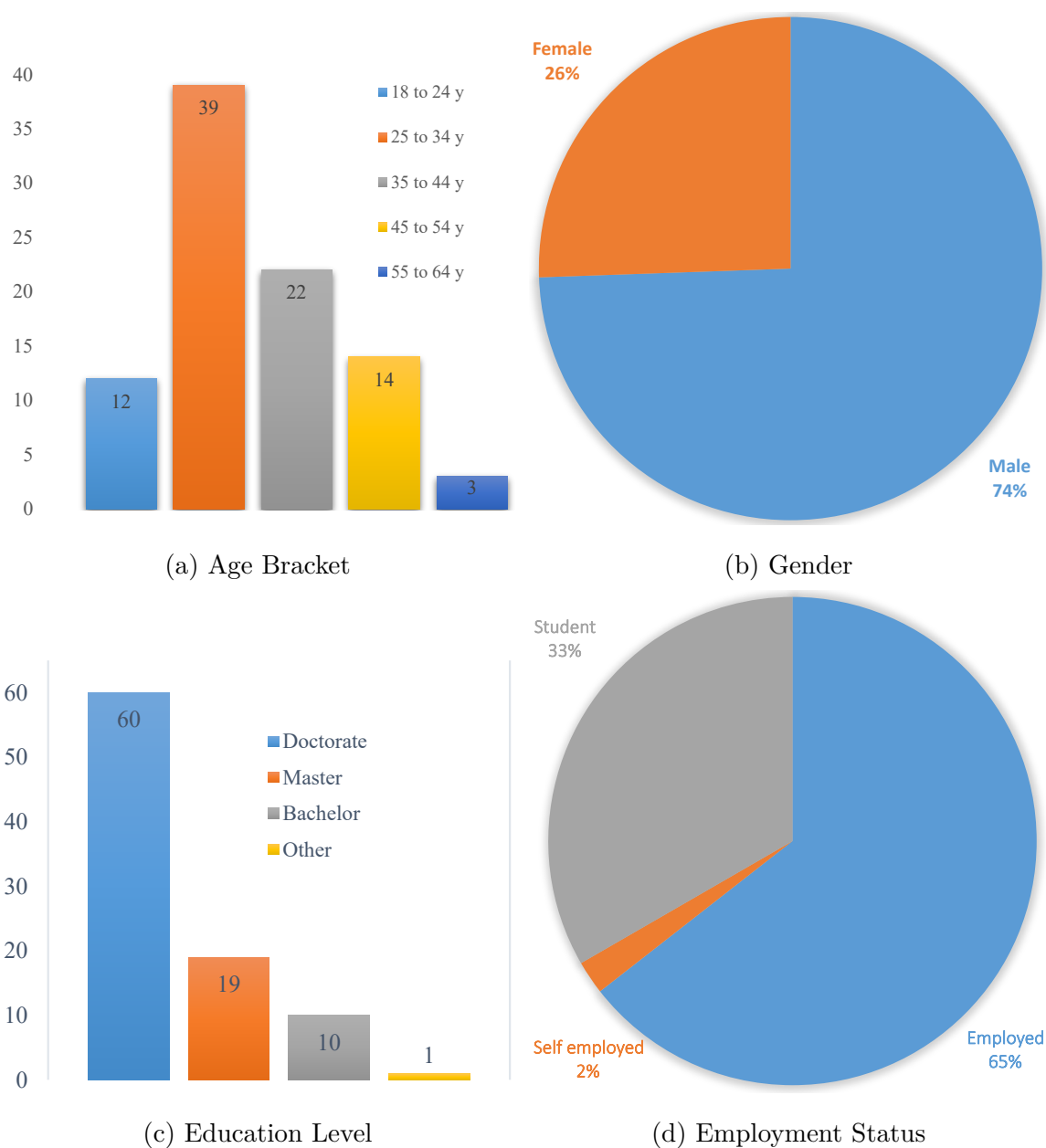


Figure 5.5: Labelling Experiment - Participant Demographics

of the participants was likely to vary, covering a wide range of age groups, see Figure 5.5a. Regarding employment status, the participants included students, self-employed and employed individuals, see Figure 5.5d. The level of education varied as well, including individuals with undergraduate degrees, postgraduate degrees, and higher academic qualifications in cyber security or related fields, see Figure 5.5c.

Answer	Count
I research cyber security	65
I study or have a degree in cyber security	51
I teach cyber security	33
I have hands-on skills that relate to cyber security	29
I work in the cyber security industry	16
I work on cyber security for the government or a not-for-profit organisation	7
I am not a professional, but interested in the latest updates about cyber security	4

Figure 5.6: Labelling Experiment - Participants Cyber Security Experience

Furthermore, the participants had varying levels of experience in cyber security, see Figure 5.6, which encompassed both theoretical knowledge and practical expertise, which was valuable to ensure the quality and accuracy of the labelling process.

5.4.4 Labelled Dataset

Due to the challenges faced in recruiting the desired 100 participants for the experiment, some borderline cases emerged during the majority voting process. These cases involved accounts with just two votes, where one participant labelled it “Related” while another labelled it “Non-Related”. Similarly, there were instances where three different votes were received for the same account due to the three available options on the labelling interface (“Related”, “Not Related” and “Not Sure”).

To address these borderline cases, we had to review them and make additional votes. After this additional review step, we were able to apply majority voting to all the accounts in the labelled dataset. As a result, a total of 987 accounts were identified as cyber security related, while 231 accounts were classified as non-related. To balance the final dataset, 756 additional non-related accounts were randomly selected from the accounts dataset after undergoing manual inspection, see Section 5.3.3. These accounts were added to the dataset, resulting in a balanced dataset with 1,974 samples evenly distributed between the cyber security related and non-related classes, with 987 samples in each class.

5.5 Feature Extraction

Since we wanted to build several classifiers, we identified a rich set of (63 types of) features grouped into five larger groups, namely, Profile (**P**), Behavioural (**B**), Content Statistics (**C**), Linguistic (**L**), and Keyword-based (**K**) features. Each feature type within these groups measures a specific aspect of the Twitter account and contributes to the overall feature set used for training and testing the ML models. For a detailed list and description of the specific features within each group, please refer to Table 5.3. Most of the feature types are simple (i.e., contain just one single feature), while others include a group of features sharing a particular attribute. We explain the features of each group in the following sections.

5.5.1 Profile Features (**P**)

These features are extracted from the profile fields of each Twitter account and divided into four categories as follows.

First, the **Screen name** category contains features corresponding and calculated from the Twitter username (`screen_name`) field. The features include the length of the screen name and the number of different types of characters present in the screen name (e.g., Alphabetic, Lowercase, Uppercase, Numerical, and Special).

Second, the **Description** category contains features derived from the account's description field. These features include the length of the description, the number of different types of characters in the description (similar to the `screen_name` category), and the number of control characters, i.e. Non-Printing Character (NPC), found in the description. The presence of NPCs in the description can be an indication, as we observed, that the account may be associated with hacker activities. Additionally, the number of words in the description and the occurrence of cyber security keywords in the description (i.e., KiD) were also calculated as features.

Third, the **Network** category contains the number of friends (i.e. following), followers and the ratio between them. Fourth, the **Miscellaneous** (Misc) category

Table 5.3: List of all extracted features and their main groups

Profile Features (P)		Behavioural Features (B)			
Screen Name	F01	Length (screen name)	F26	Count (Tweets)	
	F02	Count (Alphabetic chars)	F27	Count (Original tweets)	
	F03	Count (Lowercase chars)	F28	Count (Retweets)	
	F04	Count (Uppercase chars)	F29	Count (Replies)	
	F05	Count (Numerical chars)	Tweets	F30	Count (Tweets with mentions)
	F06	Count (Special chars)	Statistics	F31	Ratio (Original tweets to all)
Description	F07	Length (Description)	F32	Ratio (Retweets to all)	
	F08	Count (Alphabetic chars)	F33	Average (Number of mentions)	
	F09	Count (Lowercase chars)	F34	Average (Number of hashtags)	
	F10	Count (Uppercase chars)	F35	Average (Number of URLs)	
	F11	Count (Numerical chars)	F36	Count (Tweets received likes)	
	F12	Count (Special chars)	F37	Count (Tweets were retweeted)	
	F13	Count (Control chars)	F38	Count (Mentioned users)	
	F14	Count (Words)	Network	F39	Count (Replied-to users)
	F15	Count (Keywords)	F40	Count (Likes given)	
	Network	F16	Count (Friends)	F41	Count (Likes received)
F17		Count (Followers)	F42	Count (Retweets received)	
F18		Ratio (Followers/Friends)	F43	Average (Daily Tweets)	
Misc	F19	Profile Image used?	Activity	F44	Average (Weekly Tweets)
	F20	Profile Theme used?	F45	Average (Monthly Tweets)	
	F21	Location provided?	F46	Average (time between tweets)	
	F22	Count (Lists)	F47	STD (time between tweets)	
	F23	Account protected?			
	F24	URL provided?			
	F25	Account Age			
Content Features (C)		Linguistic Features (L)			
Cyber Security Keywords Statistics	F48	Count (Keywords)	LIWC	F57	Measures ($L_i, i \in [1, 93]$)
	F49	Count (Keywords) [no retweets]	Keyword-based Features (K)		
	F50	Count (Unique Keywords)	F58	Weirdness Score	
	F51	Count (Unique Keywords) [no retweets]	F59	Prototypical Words	
	F52	Count (Tweets with keywords)	Keywords	F60	TF-IDF Score
Readability & Diversity	F53	Ratio (Tweets with keywords)	Frequencies	F61	Document Frequency (DF)
	F54	Flesch-Kincaid Score	F62	Hybrid Metric DF-IDF	
	F55	SMOG Index	F63	Hybrid Metric DF-TFIDF	
	F56	Lexical Diversity			

encompasses additional features related to other profile fields that do not fit into the previously defined categories, such as the use of profile image, profile theme, URL

field, and location. Also, these features contain the number of user's created lists, account age (i.e., the number of years since the account was first created), and a feature to indicate if the account is protected.

5.5.2 Behavioural Features (B)

These features cover three aspects: statistics about the account's tweets, the interaction between Twitter accounts, and the account's general activity patterns on Twitter. We divided these features into three categories: **Tweets Statistics**, **Network**, and **Activity**. Each category covers a different aspect of the account's behaviour and interaction with other accounts. The **Tweets Statistics** category contains measures about the user's tweets such as the number of tweets, original tweets (i.e. non-retweets), retweets, replies and tweets with mentions. Also, average and ratio measures were calculated for tweets with mentions, hashtags or URLs.

The purpose of the **Network** features in this group is to represent the interaction between accounts on Twitter in a few simple measures. For example, an account can post a tweet, retweet a tweet, like a tweet, comment on a tweet, reply to another account, or mention another account in a tweet or comment. Each of these actions can be seen as an interaction between two Twitter accounts. Finally, the **Activity** category presents some measures to reflect how much a user is active on Twitter in terms of tweeting and retweeting tweets.

5.5.3 Content Statistics Features (C)

These features were extracted from the timeline's content. For each Twitter account, we retrieved up to 3,250 tweets from its timeline as permitted by Twitter API at the time we carried out our study. The Content features contain two categories. First, **Keywords Statistics**, which corresponds to measures calculated about the cyber security keywords found in the account's timeline. The used keyword list was obtained from the general cyber security taxonomy built earlier in Chapter 3. We distinguished between the keywords found only in the original tweets and those in

the whole timeline (original tweets and retweets). Also, we calculated the number of unique cyber security keywords that were used in tweets and retweets. Second, **Readability and Diversity** metrics, which include SMOG score (see Section 2.5.3.2), Flesch-Kincaid reading grade level (see Section 2.5.3.3), and Lexical Diversity, which has been employed in previous studies for spam detection (Aswani, Kar, and Vigneswara Ilavarasan, 2018), see Section 2.5.3.1 for more details.

5.5.4 Linguistic Features (L)

For the linguistic features, we used LIWC (Linguistic Inquiry and Word Count) features (Tausczik and Pennebaker, 2010), which are widely recognised and extensively employed features for text analysis, with applications spanning various research domains. LIWC measures have been successfully utilised to analyse posts from underground forums and hacking websites, as demonstrated by (McAlaney et al., 2020). We leveraged the power of LIWC (2015 Edition v16) to conduct an in-depth analysis of the accounts' timelines to capture and quantify diverse linguistic characteristics exhibited within each timeline. LIWC offers a rich set of predefined categories and linguistic dimensions with 93 features that facilitate the systematic analysis of textual content. Examples of LIWC categories include positive and negative emotions, cognitive processes, social words, and various linguistic styles.

5.5.5 Keyword-based Features (K)

The keyword-based feature group comprises several sub-groups, each consisting of features defined by a specific keyword and its corresponding metric calculated from the account's timeline. Certain metrics require text corpora reflecting the domain of interest. To address this, we created two text corpora using the labelled dataset. The first corpus was generated by combining the Twitter timelines of cyber security related accounts, while the second corpus was formed by applying the same process to non-cyber security related accounts. To ensure consistency in the analysis, we performed pre-processing steps on each timeline, removing stop words, URLs, email

addresses, punctuation marks, screen_names, and other Twitter-related symbols such as “RT”, “#”, and “@”. Then, we extracted unigrams and bigrams along with their frequencies from the pre-processed text on a per-timeline and per-corpus basis.

Next, we calculated the keyword metrics and scores as detailed below and used each of them to rank all candidate keywords, from which we selected the top k from each domain to form a list of $k \times 2$ keywords per each metric. Any frequencies used in such metrics were normalised per timeline to allow comparison across accounts. We chose $k = 100$ as a reasonable number of top-ranked keywords, considering a number of factors, such as the size of our dataset, the number of features from other groups, and the need to reduce the total number of features for our classifiers.

5.5.5.1 Weirdness Score

Weirdness score, explained in Section 2.5.2.5, was used for keywords-based features. One important thing to highlight is that Weirdness score method assumes there are two corpora, a special one s and a general one g , which is in our case the textual corpus corresponding to the timelines of the cyber security related accounts, and the other corpus corresponding to the timelines of the non cyber security accounts.

5.5.5.2 Prototypical Words

Prototypical words, explained in Section 2.5.2.6, refer to those words that are considered typical or representative examples of a category or concept. These words are often used as reference points when people think about or discuss a specific category. In our work, it is worth mentioning that before calculating the proto scores, the words with a frequency below six or less than three characters long were eliminated, and we selected the top k words from each class.

5.5.5.3 Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF, explained in Section 2.5.2.4, is commonly used in IR and text classification tasks. We used the TF-IDF metric to extract informative keywords to use their TF

scores as features for the ML classifiers. TF-IDF was defined using Formula (2.18), which is the product of the term frequency (TF), see Formula (2.15), by the inverse document frequency (IDF), given in Formula (2.17). However, for our classification tasks, the number of “documents” (i.e. corpora) is two as we have two classes.

5.5.5.4 User Count (UC)

Some keywords selected by TF-IDF or Prototypical metrics got a relatively high score, although they appeared only in about 25% of all user timelines (i.e., documents⁵) in both corpora. Thus, a supplementary metric was introduced, the user count (UC) score of a keyword, which is the number of users (i.e., documents/timelines) in the domain corpus that this keyword appeared in at least once (Y. Yang and Pedersen, 1997). The UC for a word is given by Formula (5.1).

$$\text{UC}(w, d) = \frac{U(w, d)}{U(d)} \quad (5.1)$$

where $U(w, d)$ is the number of users from the domain d (i.e., class) that cover word w , and $U(d)$ is the total number of users in d .

For this metric, we noticed that only unigrams were found in the top k terms we selected from each domain, as it is impossible for a bigram to have a higher frequency than any of its two unigrams. Also, it is worthwhile to mention that when we selected the top 100 words from each domain, there were 34 overlapping keywords. Thus, we chose $k = 121$ so we could get 200 unique keywords at the end.

5.5.5.5 Hybrid Metric UC-IDF

To expand the feature set, we introduced two hybrid metrics. The first one was the product of UC and IDF, given in Formula (5.2).

$$\text{UC-IDF}(w, d) = \text{UC}(w, d) \times \text{IDF}(w) \quad (5.2)$$

⁵In TF-IDF, a “Document” means the class’s corpus, while here it means the Twitter account’s timeline. All timelines in a given class form its corpus.

Table 5.4: Top unigrams by each keyword metric from the cyber security corpus

Weirdness	Prototypical	TFIDF	UC	UC-IDF	UC-TFIDF
threatpost	frama	frama	security	xxe	frama
securityaffair	layer7datasolutions	layer7datasolutions	time	imperva	layer7datasolutions
eprint	cyberhoot	cyberhoot	help	smbghost	cyberhoot
trustyourinbox	cyware	cyware	data	coinhive	cyware
vulnerabilitymanagement	readcybernews	readcybernews	day	raas	readcybernews
secaas	mikejulietbravo	mikejulietbravo	check	qnap	mikejulietbravo
dretrico	opendir	opendir	attack	databreaches	opendir
avira	castlekeep	castlekeep	start	endpointsecurity	cybersecuritynew
securityawareness	cybersecuritynew	cybersecuritynew	free	cybersecuritynew	castlekeep
shibboleth	avatier	avatier	create	sqlmap	avatier
wss	pentestmag	pentestmag	team	phishy	pentestmag
emailsecurity	laukaya	laukaya	learn	lfi	laukaya
technews	rmail	rmail	user	cyberhygiene	rmail
webroot	araldi	araldi	change	email-based	araldi
bleepingcomputer	redspam	redspam	people	goznym	redspam
cwpodcast	k12cybersecure	k12cybersecure	read	casb	k12cybersecure
whackinfest	equologix	equologix	system	threatdetection	consentua
threatmodeler	consentua	consentua	support	vpns	equologix
as46606	pentestblog	pentestblog	network	agenttesla	pentestblog
cyberslide	ptblog	ptblog	service	sigred	ptblog
cryptojacking	modex	modex	share	shlayer	modex
bwapp	cybersurkshaabhiyan	cybersurkshaabhiyan	online	trendmicro	cybersurkshaabhiyan
internetsecurity	securityintelligence	securityintelligence	report	strandhogg	securityintelligence
crede	nulljob	nulljob	update	trojanize	nulljob
keywords	attacksolution	attacksolution	access	womenincybersecurity	attacksolution
cybercriminals	chimenetworking	chimenetworking	secure	vulnhub	chimenetworking
threatintelligence	onlinedanger	onlinedanger	company	cybergang	onlinedanger
cybersecuritytrain	datesecurity	datesecurity	cybersecurity	prolock	datesecurity
securityweek	inspiredbmedia	inspiredbmedia	tool	blackenergy	inspiredbmedia
hxxp	mikeechols	mikeechols	protect	crackmapexec	mikeechols
groupolicy	cyberlawconf	cyberlawconf	post	covid-19-themed	cyberlawconf
centaurus	securitymadesimple	securitymadesimple	week	rsa2019	securitymadesimple
itsecurity	netspi	netspi	email	onelogin	netspi
proba	amag	amag	issue	sqlinjection	security
gartneriam	giwebint	giwebint	include	ourmine	amag
ransomwareprotection	employmentlawnew	employmentlawnew	video	nerc	time
giacomo	imperva	imperva	news	cyberdefence	help
franking	attackdefense	attackdefense	account	zscaler	imperva
egregor	extremehacking	extremehacking	code	webdav	data
zimperium	hardwear	hardwear	business	internet-facing	day
bugbountytip	attacksolutions	attacksolutions	talk	network-based	check
anti-phishing	netsparker	netsparker	app	redirector	attack
intezer	hackerscharity	hackerscharity	cyber	neutrino	start

where $UC(w, d)$ is the User Count score for the word w in the domain d , and $IDF(w)$ is the Inverse Document Frequency of the word w across both domains. UC-IDF aims to capture both the local importance of a term within a class corpus and its global significance across the class corpora.

5.5.5.6 Hybrid Metric UC-TFIDF

The second hybrid metric was the lambda of UC and TF-IDF, shown in Formula (5.3).

$$UC\text{-TFIDF}(w, d) = ((1 - \alpha) \times TFIDF_d(w)) + \alpha \times UC(w, d) \quad (5.3)$$

where α is a constant between zero and one ($0 < \alpha < 1$). Setting $\alpha = 0$ means that we ignore the UC part, while setting ($\alpha = 1$) will ignore the TF-IDF part. In our experiments, we found that 0.2 is the best value of α .

The results and performance of the used keywords metrics can be found in Section 5.7. To see the differences between these metrics in terms of the selected keywords, we listed the top unigrams produced by applying each keyword metric (as a ranking method) in Table 5.4. It is worth mentioning that we considered both unigrams and bigrams for the keyword features.

5.6 Classification Tasks & ML Models

Using the labelled datasets, we developed multiple ML classifiers during this study. The first classifier, named the Baseline classifier (Task 1), was created using the entire labelled dataset. Additionally, three sub-classifiers (Tasks 2,3,4) were created, each focusing on a specific subset extracted from the labelled dataset based on the additional assigned tags obtained from the labelling experiment.

Throughout the practical experiments, we employed the labelled datasets to train and evaluate various ML models across the four classification tasks. The

objective was to determine the optimal model and feature set combination. We selected popular and widely used models for similar classification tasks, including Decision Trees (DTs), Random Forests (RF) with 100 estimators/trees, Support Vector Machine (SVM), and Logistic Regression (LR). For SVM, we explored two kernel functions: the Radial Basis Function (RBF) kernel and the Linear kernel. These kernels provide different approaches for separating and classifying data points. The RBF kernel is known for its ability to handle non-linear relationships between features, while the Linear kernel assumes a linear decision boundary. For more details about these ML models, refer to Section 2.4.6.

5.6.1 Detecting Cyber Security Related Accounts

This is the first classification task that corresponds to the baseline classifier. This classifier was designed to determine whether a Twitter account is related to cyber security or not. It aims to establish a foundation for the following three classification tasks (2, 3, 4). By utilising the predefined feature sets and ML algorithms, the baseline classifier aims to accurately classify Twitter accounts based on their relevance to the cyber security domain.

5.6.2 Detecting Cyber Security Individual Accounts

This classification task aims to detect the cyber security accounts belonging to individuals and those representing non-individuals such as groups, companies, NGOs, etc. This classifier should be used after the baseline classifier. Thus, all the accounts in the Individual sub-dataset must be cyber security related accounts. In the dataset, we have 542 samples labelled as Individual accounts and 448 as non-individual ones.

5.6.3 Detecting Hacker Related Accounts

We wanted to create a classifier that can detect if a Twitter account is affiliated with a hacker (whether an individual or a group) or acting like a hacker. The

labelling experiment interface has several tags corresponding to the different types of hackers, e.g., white-hat, grey-hat, and black-hat hackers. Also, there is a general tag “Hacker” to make it easier for participants when they are unsure about the hacker type. We got 166 accounts labelled as Hackers, and we added randomly, yet manually checked, another 166 general cyber security accounts from the main dataset to make the Hacker sub-dataset balanced.

5.6.4 Detecting Cyber Security Academia Related Accounts

The classifier aims to identify whether a cyber security related account belongs to someone (or a group) in academia, such as students, lecturers, or researchers. We used specific tags in the labelling experiment, including “Student”, “Lecturer”, “Researcher”, and a general tag “Academia”. We obtained 129 accounts labelled as Academia accounts. In order to create a balanced dataset, we randomly selected an additional 129 general cyber security related accounts from the main dataset.

5.7 Experimental Results

We used Scikit-Learn (Pedregosa et al., 2011), a widely adopted machine learning library written in Python⁶, for our experiments. To ensure robust evaluation, we employed 5-fold stratified cross-validation for training and testing the ML models we used. The results were reported using four performance metrics: Accuracy, F1-score, Precision and Recall. These metrics provide insights into the classifiers’ overall effectiveness, their ability to correctly classify instances, and their balance between true positives and false positives, see Section 2.4.7.2 for more on performance metrics.

The experimental results for all the classifiers are reported in Table 5.5. Each row in the table corresponds to a specific feature set for a particular classifier. The #F column indicates the total number of features in each feature set, while the #S column represents the number of samples used for training and testing the

⁶<https://scikit-learn.org/>

classifier. The keyword-based feature sets are denoted with a prefix **K** in their names, highlighting their specific nature. We only showed the best-performing sub-groups of keyword-based features for some classifiers to save space.

It is important to note that in some cases, the number of samples may be smaller than the total dataset size of 1,974. This is because not all samples possess the required features for a specific feature set. This applies to both the baseline classifier (Task 1) and the subsequent sub-classifiers (Tasks 2-4). Overall, these performance metrics and experimental results allowed us to assess the effectiveness of the different feature sets and classifiers in accurately detecting and classifying cyber security related accounts on Twitter and the other related sub-groups.

5.7.1 Classifier Results

5.7.1.1 Baseline Classifier

For the baseline classifier, we used different feature sets based on the feature groups we described in Section 5.5. We experimented with many feature sets resulting from either one feature group or a mixture of two or more feature groups. We wanted to observe the impact of selecting different groups of feature sets on the results.

The best performance achieved reported by F1-score using the Behavioural features is 77%, while the figure is 86% for Profile features and 88% for Linguistic features. For the Content Statistics features, the best performance achieved is 93%, which is surprisingly good considering it is a small feature group with just nine features. As for the keyword-based features, we noticed that the prototypical keywords method was not quite good, with a score of 68% using the SVM (RBF kernel) model and the same for the TF-IDF method using the same model. The weirdness scored 82% by F1-score using Random Forest, and then the UC-IDF score method scored 85% using also Random Forest. The best performance was achieved using the UC feature set with 93%, followed by the UC-TFIDF feature set with 90%. When all keyword-based features are added together, i.e. (K_ALL), the performance is also 93%, which is likely due to the UC feature set existence.

Table 5.5: Overall experimental results for the four classifiers

Feature set	#F	#S	Decision Tree				Random Forest				Logistic Regression				SVM (Linear)				SVM (RBF)			
			Acc	F1	Prc	Rec	Acc	F1	Prc	Rec	Acc	F1	Prc	Rec	Acc	F1	Prc	Rec	Acc	F1	Prc	Rec
Baseline	P	25	1974	0.78	0.78	0.79	0.77	0.85	0.86	0.81	0.91	0.82	0.82	0.82	0.84	0.84	0.82	0.87	0.84	0.84	0.83	0.86
	B	22	1974	0.71	0.71	0.71	0.70	0.77	0.77	0.75	0.79	0.74	0.73	0.76	0.70	0.74	0.73	0.76	0.71	0.74	0.74	0.74
	C	9	1882	0.89	0.89	0.90	0.89	0.92	0.93	0.90	0.95	0.92	0.92	0.94	0.90	0.92	0.92	0.92	0.93	0.92	0.93	0.91
	PBC	56	1974	0.88	0.88	0.88	0.88	0.92	0.92	0.90	0.95	0.91	0.90	0.92	0.89	0.91	0.91	0.91	0.91	0.91	0.91	0.90
	L	93	1882	0.80	0.81	0.80	0.82	0.87	0.88	0.87	0.88	0.85	0.86	0.85	0.88	0.87	0.88	0.86	0.90	0.87	0.88	0.86
	PBCL	149	1974	0.88	0.88	0.88	0.87	0.91	0.92	0.89	0.94	0.91	0.91	0.92	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	K_WEIRD	200	1885	0.81	0.79	0.90	0.70	0.83	0.82	0.91	0.74	0.52	0.68	0.52	1.00	0.51	0.68	0.51	1.00	0.57	0.70	0.55
	K_PROTO	200	1885	0.68	0.55	1.00	0.38	0.68	0.56	1.00	0.39	0.51	0.68	0.51	1.00	0.51	0.68	0.51	1.00	0.51	0.68	0.51
	K_TFIDF	200	1885	0.66	0.51	1.00	0.34	0.66	0.51	1.00	0.35	0.51	0.68	0.51	1.00	0.51	0.68	0.51	1.00	0.52	0.68	0.52
	K_UC	199	1885	0.87	0.87	0.88	0.87	0.93	0.93	0.91	0.96	0.88	0.88	0.90	0.87	0.62	0.73	0.58	1.00	0.91	0.91	0.90
	K_UC-IDF	200	1885	0.85	0.83	1.00	0.71	0.87	0.85	1.00	0.74	0.51	0.68	0.51	1.00	0.51	0.68	0.51	1.00	0.79	0.74	0.98
	K_UC-TFIDF	196	1885	0.87	0.87	0.87	0.87	0.90	0.90	0.89	0.92	0.76	0.81	0.68	0.99	0.51	0.68	0.51	1.00	0.89	0.89	0.90
	K_ALL	903	1885	0.87	0.87	0.88	0.86	0.92	0.93	0.91	0.95	0.88	0.88	0.90	0.87	0.63	0.73	0.58	1.00	0.90	0.90	0.90
PBCLK_ALL	1052	1885	0.88	0.89	0.89	0.88	0.93	0.93	0.90	0.97	0.91	0.92	0.91	0.92	0.92	0.92	0.91	0.93	0.91	0.92	0.91	
Individual	P	25	957	0.65	0.67	0.68	0.67	0.76	0.78	0.79	0.76	0.79	0.77	0.81	0.77	0.79	0.78	0.80	0.75	0.77	0.77	
	B	22	957	0.76	0.78	0.78	0.78	0.82	0.83	0.85	0.81	0.80	0.81	0.82	0.81	0.80	0.81	0.83	0.81	0.81	0.86	
	C	9	937	0.70	0.73	0.72	0.74	0.78	0.79	0.81	0.77	0.79	0.81	0.78	0.84	0.79	0.81	0.79	0.83	0.80	0.82	
	PBC	56	957	0.77	0.79	0.78	0.80	0.85	0.85	0.89	0.82	0.85	0.87	0.86	0.87	0.85	0.86	0.87	0.86	0.85	0.86	
	L	93	937	0.83	0.84	0.84	0.84	0.88	0.89	0.92	0.86	0.86	0.86	0.92	0.82	0.85	0.85	0.91	0.81	0.85	0.85	
	PBCL	149	957	0.82	0.84	0.83	0.84	0.89	0.90	0.92	0.87	0.89	0.89	0.91	0.88	0.89	0.89	0.91	0.88	0.87	0.88	
	K_UC	129	939	0.74	0.76	0.75	0.77	0.82	0.83	0.88	0.78	0.54	0.70	0.54	0.98	0.54	0.70	0.54	1.00	0.80	0.81	
	K_UC-IDF	200	939	0.80	0.77	1.00	0.63	0.84	0.83	1.00	0.71	0.54	0.70	0.54	1.00	0.54	0.70	0.54	1.00	0.63	0.74	
	K_UC-TFIDF	152	939	0.71	0.73	0.72	0.74	0.81	0.82	0.86	0.79	0.54	0.70	0.54	0.98	0.54	0.70	0.54	1.00	0.81	0.82	
Hacker	P	25	317	0.53	0.53	0.55	0.53	0.62	0.62	0.63	0.68	0.68	0.69	0.68	0.67	0.66	0.68	0.65	0.66	0.66	0.67	
	B	22	317	0.53	0.54	0.53	0.57	0.60	0.61	0.60	0.62	0.64	0.67	0.62	0.72	0.62	0.66	0.60	0.73	0.62	0.65	
	C	9	313	0.54	0.54	0.54	0.55	0.55	0.54	0.55	0.54	0.65	0.67	0.63	0.72	0.61	0.67	0.58	0.79	0.64	0.67	
	PBC	56	317	0.56	0.59	0.56	0.63	0.61	0.63	0.61	0.65	0.66	0.66	0.65	0.67	0.66	0.67	0.66	0.68	0.66	0.68	
	L	93	313	0.61	0.60	0.61	0.59	0.68	0.68	0.68	0.69	0.66	0.66	0.66	0.66	0.65	0.66	0.64	0.68	0.67	0.66	
	PBCL	149	317	0.57	0.57	0.58	0.57	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.70	0.69	0.69	0.68	0.70	0.65	0.64	
	K_UC-IDF	200	314	0.80	0.81	0.82	0.86	0.88	0.88	0.89	0.89	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.56	0.23	0.89	
Academia	P	25	249	0.50	0.46	0.50	0.43	0.48	0.44	0.47	0.42	0.44	0.40	0.44	0.37	0.44	0.38	0.43	0.37	0.43	0.33	
	B	22	249	0.59	0.60	0.59	0.64	0.63	0.64	0.62	0.67	0.60	0.60	0.61	0.60	0.64	0.65	0.64	0.68	0.63	0.64	
	C	9	243	0.58	0.57	0.57	0.58	0.60	0.60	0.60	0.61	0.65	0.66	0.62	0.70	0.64	0.68	0.60	0.77	0.66	0.66	
	PBC	56	249	0.57	0.57	0.56	0.59	0.64	0.65	0.63	0.68	0.63	0.63	0.64	0.63	0.61	0.62	0.61	0.63	0.62	0.62	
	L	93	243	0.56	0.52	0.55	0.50	0.67	0.66	0.66	0.67	0.63	0.62	0.62	0.62	0.66	0.66	0.64	0.70	0.65	0.66	
	PBCL	149	249	0.61	0.63	0.61	0.64	0.63	0.64	0.63	0.66	0.63	0.62	0.63	0.62	0.64	0.64	0.64	0.64	0.65	0.64	
	K_UC-IDF	200	245	0.79	0.83	0.71	1.00	0.89	0.91	0.83	1.00	0.51	0.00	0.00	0.51	0.00	0.00	0.58	0.00	0.00	0.58	

For the mixed feature sets, we tried several combinations of all feature groups, e.g. the PBC feature set represents P, B, and C groups combined, and PBCLK_ALL means all features. As for the results, we got 92% for both PBC and PBCL. The results showed that such combined feature sets generally performed well, with an F1-score between 92-93%. However, knowing that C features alone can already achieve an F1-score of 93%, we consider such combined feature sets unnecessary.

Moreover, we created a combined feature set “K_ALL” that contains all keywords across all text metrics. The combined list contained only 903 keywords, not the expected 1200, due to the overlap between keyword lists. As reported in Table 5.5, the RF model scored 93% for K_ALL, which is the same result as the K_UC. Comparing the features count for K_UC and K_ALL, which were 199 and 903 respectively, suggests that adding more features does not always mean better results.

In terms of the five ML models, looking at the general patterns shown in Table 5.5, we can see that the Random Forest model is the only model achieving the best performance across all feature sets. The other four models also achieved good performance for many feature sets but did not perform very well for some other feature sets. As an overall conclusion, we recommend using the Random Forest model and the Content Statistics features for the baseline classifier.

5.7.1.2 Individual Sub-Classifier

For the “Individual” sub-classifier, we examined several feature sets as presented in Table 5.5. Among all feature groups, the Linguistic features (L) performed the best with an F1-score of 89%. The best performance, however, was achieved by a combined feature set PBCL, with an F1-score of 90%. The best-performing model is Random Forest across all feature sets.

5.7.1.3 Hacker Sub-Classifier

For the “Hacker” sub-classifier, we found that non-keyword feature sets did not perform well, with the highest F1-score being just 69% for the PBCL feature set. Whereas for the keyword-based features, we found features based on the UC-IDF metric gave a much better F1-score of 88% using the Random Forest model, which was the best-performing model. The Decision Tree model came second, scoring 81% for the F1-score. Note that we only put the results of one keyword feature set which is UC-IDF to save space in Table 5.5.

Looking at the Logistic Regression and SVM (Linear kernel) model columns

from the results in Table 5.5, we can see that the F1-score is 0 for the UC-IDF feature set used in the Hacker and Academia classification tasks. Upon investigating the confusion matrix in each iteration of the cross-validation training process, we noticed that both the true positives (TPs) and the false positives (FPs) were 0. Consequently, precision, recall and F1-scores are all 0 as well. This means the classifier predicted all the samples as negative cases.

5.7.1.4 Academia Sub-Classifier

In the case of the “Academia” sub-classifier, we experimented with all the created feature sets, and the results were presented in Table 5.5. Among the feature sets evaluated, the keyword-based features calculated using the UC-IDF metric exhibited the highest performance, achieving an F1-score of 91%. This performance is comparable to the “Hacker” sub-classifier. Like the previous sub-classifiers, the Random Forest model consistently outperformed other models across all feature sets, demonstrating its effectiveness in classifying cyber security academia related accounts.

5.7.2 Features Importance

We aimed to identify the most influential features in the classification tasks to optimise the classifiers by focusing on a smaller set of key features while maintaining or improving performance. This approach allows for better interpretation and understanding of the results obtained from various feature sets in each classification task. Additionally, reducing the number of features can lead to significant time savings in feature calculation, as well as shorter training and testing times.

The feature importance in this study was determined using the χ^2 feature selection method, which is a statistical measure commonly used for feature selection in text classification tasks (Y. Yang and Pedersen, 1997). By applying this method, we could rank the features based on their relevance and contribution to the classification task. We conducted feature analysis for two specific cases: the PBCLK feature set for the Baseline classifier and the L feature set for the Individual sub-classifier.

These cases were selected to gain insights into the importance and contribution of specific features in their respective classification tasks. However, further analyses and experiments can be conducted to explore the feature importance in other feature sets and classification tasks, allowing for a comprehensive understanding of the key features driving the classification performance.

5.7.2.1 Baseline Classifier with PBCLK Feature Set

In the PBCLK feature set, the top 20 ranked features were identified, and their rankings are illustrated in Figure 5.7. Notably, the first four features were derived from the Content Statistics features, which are based on aggregated statistics of cyber security related keywords. This finding is in line with expectations, as cyber security related keywords are crucial in distinguishing relevant accounts.

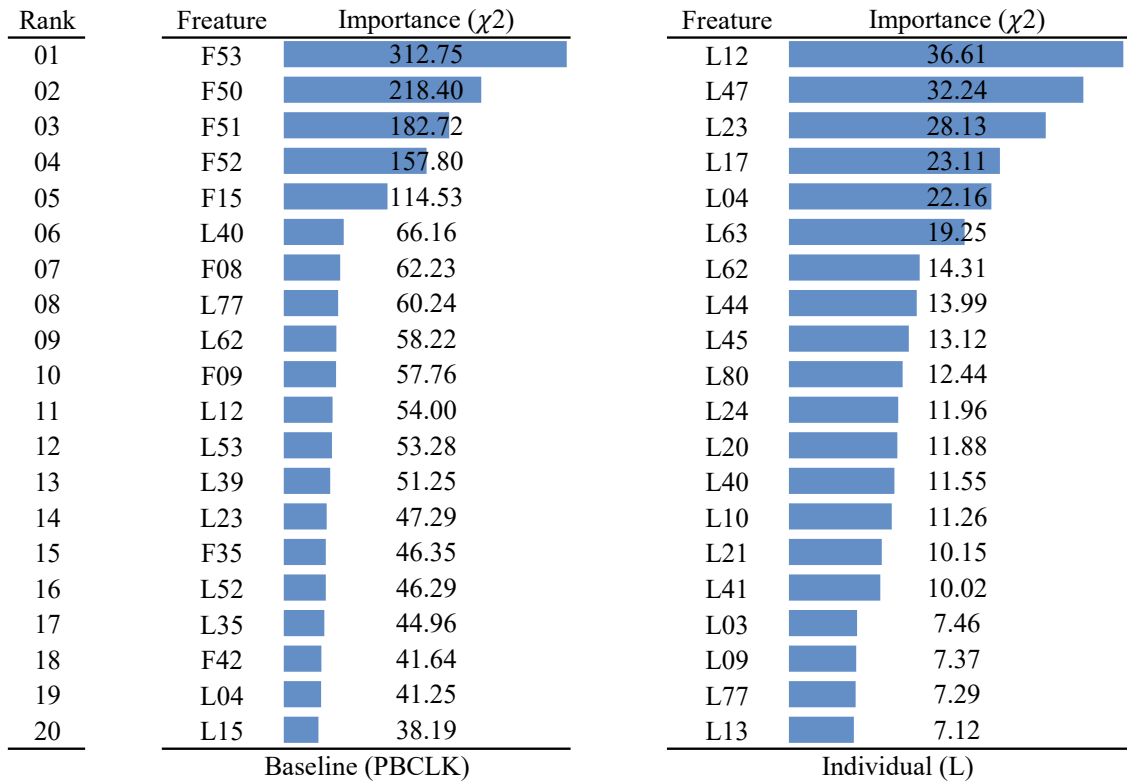


Figure 5.7: The top 20 features of two classifiers, ranked by χ^2 significance values

For instance, F53 ranked first, which is the ratio of tweets containing cyber

security related keywords to all tweets in an account timeline. Similarly, F50, ranked second, captures the count of unique cyber security related keywords found in an account’s timeline. Another notable feature, F15, ranked fifth, indicates the number of cyber security related keywords found in the profile description field.

5.7.2.2 Individual Classifier with L Feature Set

The feature importance analysis was also conducted for the L feature set used in the “Individual” sub-classifier. Figure 5.7 displays the top 20 features ranked by the χ^2 statistic. Feature L_i means the i^{th} feature in the F57 feature group (i.e., LIWC features). Interestingly, the highest-ranked feature was L12, corresponding to the “i” variable from the LIWC tool. This feature captures the frequency of words associated with the first person singular pronoun, such as “I”, “me”, “my” and “mine”. The prominence of this feature suggests that the usage of these personal pronouns may be a good indicator of Individual accounts. Notably, utilising its built-in impurity-based feature selection algorithm, the Random Forest model consistently selected this feature as the most important when training and testing the L feature set for the “Individual” classification task.

5.7.3 Features Reduction

While the baseline classifier demonstrated optimal performance with the C feature set which contains only nine features, this is not the case for other sub-classifiers. The varying feature importance as discussed in the previous section motivated us to explore the possibility of reducing the number of features for all classifiers to create lightweight versions of the classifiers without compromising their performance.

By reducing the number of features, several benefits can be achieved. Firstly, computational resources can be conserved as the training and testing time for the classifiers would be reduced. This is particularly valuable when dealing with large datasets or when deploying the classifiers in real-time applications where efficiency is crucial. Secondly, a smaller feature set can enhance the interpretability of the

classifiers. With fewer features, it would become easier to understand the underlying factors that contribute to the classification decisions. This can aid in identifying meaningful patterns and insights from the classifier predictions.

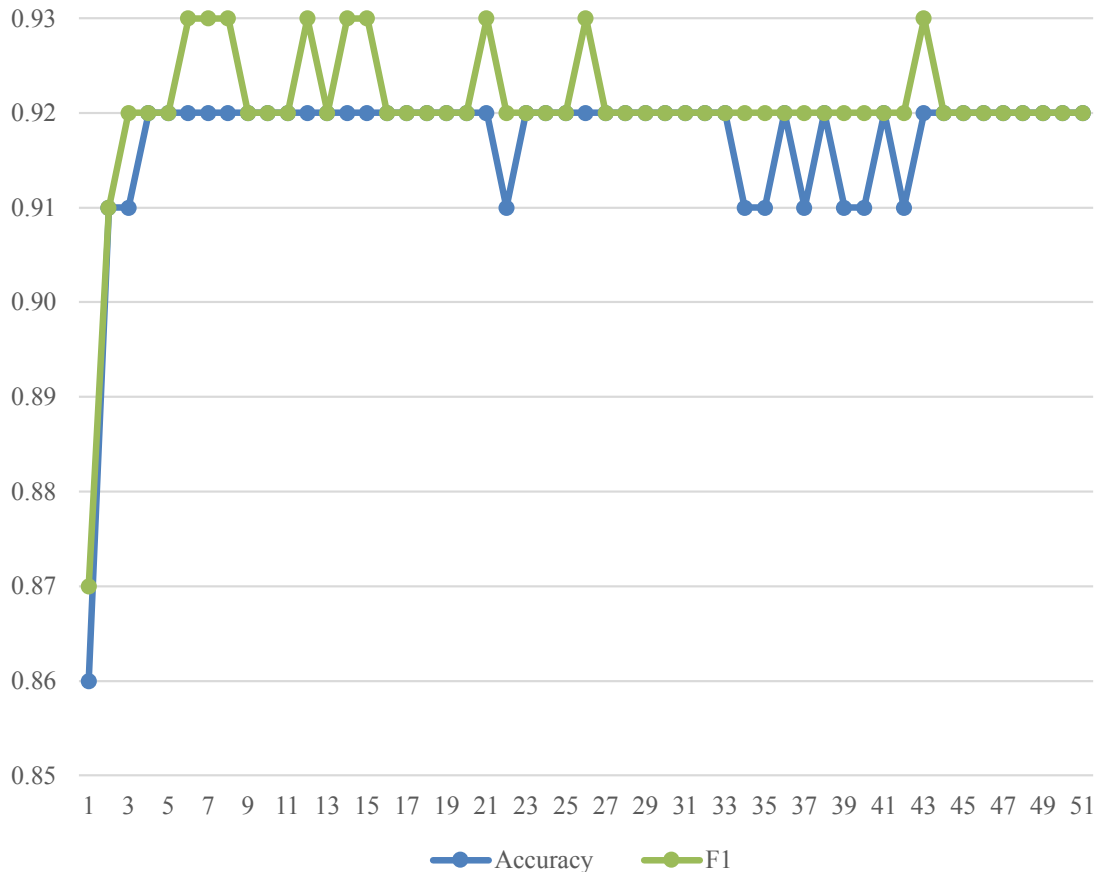


Figure 5.8: Feature reduction analysis for the baseline classifier

Using the χ^2 feature selection algorithm again, we tried to identify the smallest feature set from the most complete feature set PBCLK for the baseline classifier. We selected the top m features with the highest χ^2 scores based on the feature importance scores, then trained the Random Forest classifier again to see its performance. We evaluated how accuracy and F1 changed by adding one feature each time (until the top 51 features), and the results were reported in Figure 5.8.

The model achieved an impressive F1-score of 93% with just six features, surpassing the previous best-performing feature set C by three features. Interestingly, after

including the top six features, the classifier's performance reaches a saturation point, suggesting that additional features do not contribute to its effectiveness. Reducing the features to only six important ones, the efficiency of the baseline classifier greatly improved without compromising its performance. This approach of identifying the minimum important features can also be applied to the other classifiers.

5.7.4 Comparison to Past Studies in the Literature

5.7.4.1 Past Study 1

The work reported in (Aslan, Sağlam, and Li, 2018) was the most relevant study in the literature for our baseline classifier. Aslan, Sağlam, and Li achieved an impressive F1-score of over 97% using the RF model with various feature sets. However, it is important to note that the performance difference between their work and ours is likely attributed to the differences in the datasets used. Their dataset was smaller and potentially biased due to a more ad hoc construction approach.

To further investigate the difference in performance, we contacted the authors and obtained their dataset. We conducted a comparison between our baseline classifier and their classifier on both their dataset and ours. Surprisingly, our baseline classifier outperformed their classifier across the board. During our analysis, we identified some inaccuracies in their reported results, particularly concerning the behavioural features, which led to lower performance figures for their classifiers. We reported our findings to the authors and together confirmed the discrepancies in their results. Our performance comparison with Aslan, Sağlam, and Li's work was based on their corrected performance figures, ensuring a fair and accurate assessment.

Besides the better performance of our baseline classifier, our work exceeds (Aslan, Sağlam, and Li, 2018) with several advancements and improvements as follows.

- 1). First, we have expanded upon their classifier by developing three additional sub-classifiers, thereby providing a more comprehensive analysis of the cyber security accounts. This allows for a finer-grained classification and a deeper understanding of the different categories of users in the cyber security domain.

2). Our dataset is significantly larger, approximately 4.5 times the size of their dataset. This increased dataset size enhances our data’s representativeness and improves our findings’ generalisability. We employed a systematic approach to collect the data, ensuring a diverse and extensive coverage of cyber security accounts.

3). An important distinction lies in the labelling process. While Aslan, Sağlam, and Li relied on a single expert to label their dataset, we used a crowdsourcing-based approach, leveraging the expertise of multiple cyber security professionals. This multi-expert labelling process helps mitigate potential biases and provides a more robust and reliable dataset annotation.

4). Finally, our work achieved a higher level of granularity in feature selection and analysis. We conducted a thorough investigation of different feature sets and performed a feature importance analysis to identify the most relevant features for each classifier. This enables us to create more efficient and lightweight classifiers without sacrificing performance.

In summary, our work surpasses (Aslan, Sağlam, and Li, 2018) in terms of dataset size, representativeness, labelling process, and the development of multiple sub-classifiers. These advancements contribute to a more comprehensive understanding of cyber security accounts and improve the overall efficiency of our classifiers.

5.7.4.2 Past Study 2

In (Jones, Nurse, and Li, 2020), the authors focused on developing a machine learning classifier specifically tailored to detect Twitter accounts affiliated with the Anonymous group. Their research achieved notable results, with their best-performing model, Random Forest, achieving an F1-score of 94%. It is important to note that their classifier is designed for a specific target group, namely the Anonymous group, which may have unique characteristics and behaviour.

In comparison, our Hacker classifier is much more general, not a particular group of hackers. As a result, our classifier’s performance (88% F1-score) may appear lower when directly compared to their classifier. However, it is essential to consider the

difference in scope and target detection. Furthermore, Jones, Nurse, and Li's work relies on ad hoc features specifically designed for identifying Anonymous-affiliated accounts. These features may not be directly transferable to detecting general hackers or hacking groups. In contrast, our approach utilises a more comprehensive set of features that can capture a broader range of hackers and hacking groups, making it more applicable to a wider range of scenarios.

Overall, while Jones, Nurse, and Li achieved impressive results in detecting Anonymous-affiliated accounts, our focus on general hacker detection and the use of more comprehensive features make our work more suitable for identifying a broader spectrum of hacker accounts on Twitter.

5.7.5 Crowdsourcing Effects on Overall Results

Crowdsourcing offers numerous benefits, including scalability, cost-effectiveness, and access to diverse perspectives. However, the choice of crowdsourcing and the self-selection of crowdworkers can have a significant impact on the overall results of a study, particularly in terms of data quality, reliability, and representativeness. By implementing rigorous validation recruitment and addressing potential biases, we were able to secure and ensure the credibility and robustness of our crowdsourcing labelling experiment and its findings.

Crowdsourcing usually relies on the contributions of individuals from diverse backgrounds and expertise levels. The self-selection of crowdworkers means that participants choose to engage in the task voluntarily, potentially leading to variations in the quality and accuracy of their contributions. However, for our labelling experiment, we targeted individuals with cyber security experience or knowledge and we manually verified the candidate participants before recruiting them. Thus, the overall data quality was not compromised.

Also, individual crowd contributors may possess different biases, preferences, or interpretations of the task at hand. For this point, we utilised majority voting which helped to mitigate the impact of individual biases by focusing on the most

common response or consensus among participants. By weighing the opinions of multiple contributors equally, majority voting minimises the influence of outlier responses or subjective biases, leading to more robust and reliable results. As for representativeness, the participants in the labelling experiment exhibited diverse demographics and professional backgrounds associated with cyber security. These demographics encompass various factors such as gender, age, employment status, and educational attainment.

5.8 Conclusion

This chapter encompasses our work addressing various challenges associated with classifying cyber security related accounts on Twitter. Our approach relied on a three-staged methodology, which involves a more systematic data collection process, a crowdsourcing-based labelling experiment, and the development of machine learning based classifiers. We created a dataset of labelled Twitter accounts with multiple tags, not just binary labels like in past studies, e.g., (Aslan, Sağlam, and Li, 2018; Jones, Nurse, and Li, 2020). The dataset was specifically designed to develop four ML classifiers capable of detecting cyber security related accounts and other sub-groups within these accounts (Individuals, Hackers, and Academia). To ensure the dataset's accuracy and representativeness, we employed a cyber security taxonomy and conducted a crowdsourcing-based labelling experiment.

The general cyber security taxonomy, created in Chapter 4, played an important role in developing the ML classifiers. The terms of the taxonomy were utilised as features during the construction of the classifiers. By incorporating the taxonomy into the feature sets, the classifiers could leverage the structured and organised nature of the taxonomy's knowledge to improve their accuracy and effectiveness in identifying cyber security related accounts. The use of the taxonomy in the feature sets ensured that the classifiers had a solid foundation of knowledge and context related to cyber security. This enabled them to learn and generalise patterns

effectively, resulting in robust and reliable classification performance.

Moreover, a richer set of features was identified for developing such classifiers than those used in similar previous studies. Thus, we created and experimented with 63 types of features divided into five larger groups, resulting in over 1,200 features. We trained and tested the four classifiers using the labelled dataset. Our results revealed that the Random Forest model consistently outperformed all other models across the four classifiers, yielding F1-scores ranging from 88% to 93%. Furthermore, we investigated the significance of different features and discovered that a minimal number of features (e.g., only six for the baseline classifier) were sufficient to create a lightweight classifier with the same optimal performance.

To share our work and enable other researchers to benefit from our findings, we created a dedicated webpage⁷ that provides detailed information about this study. The webpage serves as a platform for sharing our methodology, results, and insights. We understand the importance of reproducibility and collaboration in research. We are pleased to offer the anonymised feature sets and the source code of our classifiers to other researchers by contacting us. This allows for easy comparison, validation, and potential re-use of our work. By sharing these resources, we aim to contribute to the advancement of the field and encourage further exploration and improvement.

Finally, the three-staged methodology we developed in this chapter is highly versatile and can be adapted to other communities on OSNs. Researchers in different domains can utilise our approach to create bench-marking datasets and classifiers tailored to their specific communities or topics of interest.

⁷https://cyber.kent.ac.uk/research/cyber_Twitter_classifiers/

Chapter 6

Studying Cyber Security Communities on OSNs

*“Somewhere, something incredible is waiting to
be known.”*

— Carl Sagan

MANY studies in the literature have studied different types of cyber security related users and communities on OSNs, such as activists, hacktivists, hackers, cyber criminals. A few studies also covered non-expert users who discussed cyber security related topics. However, to the best of our knowledge, none has studied the activities of cyber security researchers on OSNs. This chapter fills this gap using a data-driven approach to analyse the presence of cyber security experts on OSNs, focusing on cyber security researchers. As a case study for our analysis, we chose the UK’s Academic Centres of Excellence in Cyber Security Research (ACEs-CSR) on Twitter. The presented analysis in this chapter utilised the tools we developed in Chapters 4 and 5. Thus, we used the general cyber security taxonomy and the ML classifiers along with other tools and techniques such as social network analysis, topic modelling, and sentiment analysis in this study.

Firstly, ML classifiers were used to identify the cyber security and research accounts. Then, starting from 19 seed accounts of the ACEs-CSR, a social graph

of 1,817 research related accounts, that were followers or friends of at least one ACE-CSR, was constructed. We conducted a comprehensive analysis of the data we collected: a social structural analysis of the social graph; a topic modelling analysis to identify the main topics discussed publicly by researchers in the ACEs-CSR network, an influence analysis to find the top influential actors and to study the influence distribution, and finally a sentiment analysis of how researchers perceived the ACE-CSR programme and accounts.

This study revealed several findings: 1) social structural analysis and community detection algorithms are useful in detecting sub-communities of researchers, which helps understand how they are formed and what they represent; 2) topic modelling can identify topics discussed by cyber security researchers (e.g., cyber security incidents, vulnerabilities, threats, privacy, data protection laws, cryptography, research, education, cyber conflict, and politics); and 3) influence analysis led to identifying the top influential nodes including ACE-CSR and non-ACE-CSR nodes; 4) influence distribution analysis showed that the ACEs-CSR network is a scale-free network, and finally 5) sentiment analysis showed a generally positive sentiment about the ACE-CSR programme and ACEs-CSR. This chapter shows the feasibility and usefulness of large-scale automated analyses of cyber security researchers on Twitter.

6.1 Introduction

OSNs are increasingly recognised as a significant source of information influencing the adoption of opinions, thoughts, products and services. OSNs have become an integral part of modern society, providing a platform for individuals and communities to interact, share information, and create content (Boyd and Ellison, 2007). With the popularity of OSNs among people, identifying and finding users who form different online communities has become an interesting research topic for many as studying such communities can reveal useful insights about their memberships, people's opinions, intentions and motivations of online users' activities. Such needs

have led to a wide range of SNA applications for different purposes, such as maximising the diffusing of new ideas or technologies, improving recommendations, and increasing the accuracy of expert finding tasks (Moscatto and Sperli, 2021).

The cyber security domain is becoming increasingly complex with the vast advancements in technology, computing equipment and IT infrastructure. With that, a wide range of people is involved, whether they are professionals, researchers, cyber criminals, journalists, activists, government agents, etc. Those humans can be part of a group or community in a bigger network. The application of SNA is also frequently used to study cyber security related users on OSNs, e.g., cyber criminals (Aslan, Li, et al., 2020; Tavabi et al., 2019; Kigerl, 2018), hacktivists (Jones, Nurse, and Li, 2022; Jones, Nurse, and Li, 2020), activists (Nouh and Nurse, 2015), and non-experts (N. Pattnaik, Li, and Nurse, 2023; Saura, Palacios-Marqués, and Ribeiro-Soriano, 2021). However, no past studies have investigated cyber security researchers on OSNs using a computational data-driven approach, even though many cyber security researchers and organisations are active on OSNs, and their activities can potentially have a significant influence on other users, e.g., how non-experts learn about cyber security.

Our study addresses this research gap, by studying cyber security research related users and their activities on OSNs, which could allow us to learn more about them from several aspects, such as their memberships and social structures, their connections with other users, characteristics of their members, topics they often discuss, and their perception and opinions on different cyber security related matters. A better understanding of those aspects could help us better understand how they play a role in the wider online cyber security community.

6.2 Cyber Security Communities

6.2.1 Analysing Cyber Security Communities

Studying cyber security communities can reveal a lot of information about their members, and more insights can be learned about the studied community's motivations, opinions, and intentions. For example, studying hacker groups can reveal the motivation behind a recent attack, and we might be able to predict their future cyber attacks earlier, which allows us to take the necessary measures. Furthermore, we can also analyse cyber security expert networks or research groups to extract and learn more about the latest trending topics and insights from those people.

For this study, we utilised what we have done in the previous chapters to study cyber security communities. We found a lot of work in the literature about studying hacker groups and cyber criminals communities, but to the best of our knowledge, none has studied cyber security research communities. These communities contain cyber security researchers who are professionals in the cyber security domain, and studying their communities is essential for understanding the complete picture of the interconnected cyber security communities on OSNs.

6.2.2 Case Study: ACEs-CSR Network on Twitter

As a case study, we chose to analyse the research network around the 19 Academic Centres of Excellence in Cyber Security Research (ACEs-CSR) on Twitter. ACEs-CSR are UK universities jointly recognised by NCSC (part of GCHQ) and the Engineering and Physical Sciences Research Council (EPSRC, part of UKRI – UK Research and Innovation) (NCSC, 2019). The recognition followed an assessment process where these universities had met high standards. This includes several aspects such as the number of academic cyber security staff, the commitment from the university leadership towards cyber security research, a publishing record with high impact on cyber security research, and sustained funding from different sources (NCSC, 2019). These universities are a good representative sub-

set of cyber security researchers in the UK, allowing us to test how computational data-driven analysis can be done and to have a view of the important part of the UK cyber security research community on Twitter. See Table 6.1 for a list of all the universities that are currently recognised as ACEs-CSR.

Table 6.1: The 19 UK universities recognised as ACEs-CSR

University/ACE-CSR	Representing research unit(s)
Queen’s University Belfast	Centre for Secure Information Technologies (CSIT)
University of Birmingham	Centre for Cyber Security and Privacy
University of Cambridge	Security Group
Cardiff University	Centre for Cyber Security Research (CCSR)
De Montfort University	Cyber Technology Institute
University of Edinburgh	Cyber Security, Privacy and Trust Institute
University of Kent	Institute of Cyber Security for Society (iCSS)
King’s College London	King’s Cybersecurity Centre
Lancaster University	Security Lancaster
University College London	UCL Information Security Research Group (UCL ISec)
Newcastle University	Centre for Cyber Security and Resilience
University of Oxford	Cyber Security Oxford
University of Southampton	Cyber Security Research Group
University of Surrey	Surrey Centre for Cyber Security (SCCS)
Royal Holloway, University of London	Information Security Group
	Institute for Cyber Security Innovation
Northumbria University	Northumbria Cyber Security Research Group (NCSRG)
	Cybernetwork Systems and Security (CyberNets)
University of Bristol	Bristol Cyber Security Group
	Cryptography Research group
	Trustworthy Systems Laboratory
Imperial College London	Centre for Cryptocurrency Research and Engineering (IC3RE)
	Centre for Engineering Secure Software Systems
	Institute for Security Science and Technology (ISST)
University of Warwick	WMG Cyber Security Centre
	Data Science, Systems and Security (DSSS)

ACEs-CSR was a consequence of the 2011-2016 National Cyber Security Strat-

egy of the UK and then continued in the 2016-2021 Strategy, and a new round of recognition is ongoing based on the new National Cyber Strategy 2022.

6.3 Research Questions

We found a gap in the literature about studying cyber security experts on OSNs. Thus, we wanted to explore this area, focusing on the UK ACEs-CSR network on Twitter as a case study. The main research objective is to study the cyber security researchers in the ACEs-CSR network and to see what insights can be obtained from their social structure and sub-communities on Twitter. Also, using quantitative methods such as topic modelling and sentiment analysis, we intend to analyse the topics they discussed. Thus, our research questions (RQs) for this chapter are:

- **RQ3.1:** How to identify cyber security research accounts on Twitter?
- **RQ3.2:** What is the social structure of a typical cyber security research community on Twitter, such as the one formed by ACEs-CSR and their followers?
- **RQ3.3:** How is the influence distributed in the ACEs-CSR social graph and who are the top influential actors?
- **RQ3.4:** What topics do cyber security research accounts in the ACEs-CSR network discuss online on Twitter?
- **RQ3.5:** What is the general sentiment of cyber security research accounts towards the ACE-CSR program and accounts on Twitter?

6.4 Methodology

RQs 3.1-3.3 are interconnected, with **RQ3.1** serving as the foundation for addressing the subsequent research questions. To tackle **RQ3.1**, our approach involved the use of ML classifiers. The development and evaluation of these classifiers required

us to collect Twitter data starting from a number of seed accounts corresponding to the ACEs-CSR (detailed in Section 6.5).

We employed a two-fold approach to address **RQ3.1**. First, we utilised the ML classifiers built in Chapter 5, which were specifically designed to detect cyber security related accounts and individual accounts. By applying these established classifiers to our dataset, we aimed to evaluate their effectiveness in identifying such OSN users. Second, we developed a new classifier tailored specifically to detect cyber security research related accounts on Twitter. This involved creating a custom classifier using ML techniques and leveraging a set of relevant features. The features were carefully selected and extracted to capture the distinctive characteristics and content patterns exhibited by cyber security research accounts.

Moving on to **RQ3.2**, our focus shifted towards constructing a social graph using the connections between the friends and followers of the cyber security research related accounts connected with the ACEs-CSR. This social graph served as a fundamental framework for analysing the network's social structure. Employing community detection algorithms, we identified and examined different sub-communities within the graph. This analysis provided insights into the distinct groupings and relationships within the cyber security research community. It is important to note that **RQ3.2** heavily relies on the data collected and classified in **RQ3.1**, as the social graph construction is dependent on the connections between the identified cyber security research related accounts. By studying the social structure and sub-communities, we gained a deeper understanding of the interconnections and group dynamics within the ACEs-CSR network.

For **RQ3.3**, we were interested in studying the influence within the constructed social graph that represents the ACEs-CSR network for multiple objectives. First, identify the top influencers in the network to gain insights into their impact on the cyber security research community on Twitter. Second, investigate whether influence in the network was evenly distributed among cyber security research related accounts or concentrated within a small subset of accounts. To achieve this, we used

the network centrality measures as a proxy to quantify influence, then we used the centrality scores to rank the graph nodes and study the centrality distribution.

RQ3.4 focused on conducting topic modelling analysis to gain insights into the main topics discussed within the ACEs-CSR network on Twitter. This analysis was carried out using the Latent Dirichlet Allocation (LDA) algorithm, which enabled the identification and exploration of key themes and subject matters present in the timelines of cyber security research related accounts in this network.

Finally, for **RQ3.5**, our attention turned to sentiment analysis. This analysis aimed to assess the sentiment expressed in tweets that mentioned any ACE-CSR account or discussed the ACE-CSR program. Through sentiment analysis, we quantified the overall sentiment scores within the previously identified sub-communities from **RQ3.2**. This allowed us to gauge the general sentiment towards the ACE-CSR programme and the accounts associated with it. By applying topic modelling and sentiment analysis, we gained a more comprehensive understanding of the content and sentiment within the ACEs-CSR network. These analyses provided valuable insights into the prevalent topics of discussion and the sentiment expressed towards the ACE-CSR program among different sub-communities.

6.5 Data Collection

To study the stated RQs, we needed to select the right seed accounts and then crawl their friends and followers to get the needed accounts and connections between them to construct the social graph of the cyber security research related accounts in the ACEs-CSR Twitter network. The data collection for this study was carried out in June 2022. The detailed collected data are described in the following sections.

6.5.1 Seed Accounts

We created a list of 19 Twitter accounts, each corresponding to an ACE-CSR. First, we looked at each ACE-CSR's website and manually searched into Twitter to confirm

their official Twitter account. In some cases, when no official account was identified, we chose the ACE-CSR lead’s account as the seed account of the corresponding ACE-CSR. However, there was a single case when we found neither an ACE-CSR’s official account nor its lead’s account. For this very case, we chose the account of the most well-known cyber security researcher in that ACE-CSR. For simplicity, from now on, we will use the term the representative ACE-CSR Twitter account to refer to the Twitter account we selected for that ACE-CSR.

Since our RQs are unrelated to the individuals themselves but rather to the ACEs-CSR network as a whole, the dataset was anonymised to eliminate the risk of re-identification of individual researchers. To this end, this chapter does not mention any personal information related to any account, and our results do not refer to specific individuals or ACEs-CSR. Moreover, we used the name pattern “ACE-CSR- i ” as a display name for all the ACEs-CSR accounts in any visualisation and results presented in this chapter, where i is a random number from 1 to 19, while for other accounts (i.e., non-ACE-CSR accounts), we used this name pattern “user- j ”, where j is a random number from 1 to the maximum number of nodes. These measures preserve individual researchers’ privacy and avoid comparing individuals or ACEs-CSR against each other. Note that such a treatment does not affect the reproducibility of the work presented in this chapter.

6.5.2 Friends and Followers of the Seed Accounts

To construct the social graph of the ACEs-CSR, we started from the seed accounts, which we considered the first level (Lv1). Then, we fetched their friends and followers using the Twitter API, which was then denoted as level 2 (Lv2). Then, we did the same for the accounts in Lv2, which led to nodes at level 3 (Lv3). After that, we used Lv1, Lv2, and their connections. The retrieval of Lv3, which contained almost 16 million nodes, was necessary to capture all the connections between Lv2 accounts. Finally, we got 42,028 accounts in total (19 in Lv1 and 42,009 in Lv2).

To ensure manageable and feasible processing of the social graph, we set a thresh-

Table 6.2: Statistics of accounts and levels in the initial ACE-CSR network

Level	Accounts	Connections	Description
Lv1	19	52,042	Seed set, ACE-CSR Accounts
Lv2	42,009	63,004,417	Friends & Followers of Lv1
Lv3	15,904,166		Friends & Followers of Lv2

old for the number of followers and friends to retrieve for each account in levels 2 and 3. This decision was made due to the fact that some accounts have an exceptionally large number of followers or friends, which would result in a massive graph that could be challenging to analyse. The threshold was set to 5k accounts for both cases friends and followers. This means we only fetched the first 5k accounts of friends/followers, determined by the Twitter API. The statistics of levels and their member accounts are reported in Table 6.2.

6.5.3 Account Timelines

As several ML classifiers were used in this study to filter the Twitter accounts, we needed to extract and calculate many features from the timelines of those accounts. Thus, we used the Twitter API to obtain these timelines (up to 3,250 tweets per account, adhering to the limitations imposed by the API).

6.6 Machine Learning Classifiers

As per the case study set earlier in this chapter, studying the social graph and the communities of the ACEs-CSR network on Twitter, it was essential to identify accounts that were relevant to both cyber security and research. This necessitated the use of two ML classifiers specifically tailored for this purpose. The first one was the baseline classifier which we built earlier, and the second one was the research related classifier which we built later in this chapter. Furthermore, we required a

classifier capable of distinguishing between individual accounts and non-individual accounts, encompassing various entities such as groups, organisations, governments, NGOs, and news channels. Therefore, a third classifier was incorporated into our study to address this aspect. By employing these three classifiers, we were able to gain insights into the composition and characteristics of the ACEs-CSR network.

6.6.1 Cyber Security Related and Individual Classifiers

Regarding the cyber security related and individual classifiers, we used two classifiers developed in our work described in Chapter 5. Before using these two classifiers, we re-trained and re-evaluated their performances. We extracted the required features for our data collection as described in Section 5.5. Then, the selected trained classifiers were used to predict the class of each account in the data collection (i.e., 42k accounts) according to each classification task. The prediction statistics are listed in Table 6.4. The Individual classifier was applied following the Cyber Security Baseline classifier to detect cyber security related individuals. Also, we applied the Individual classifier after the Research classifier, described in Section 6.6.2, to determine whether a research related account is for an individual (e.g., researcher).

6.6.1.1 Previous Classifiers Evaluation

Before re-using the classifiers (created in Chapter 5), we had to do a validation for them by measuring their performance in a real-life case with data that is completely different from the one these classifiers were trained with originally. Thus, we used the data we collected for our case study (i.e. the 42k Twitter accounts) for the validation procedure. By applying the classifiers to this new dataset, we aimed to assess their performance and determine their effectiveness in a different context.

6.6.1.2 Feature Extraction

We re-trained the classifiers to ensure the results were still the same and to test some additional ML models to see if we could get better results using other models.

First, we used the same labelled datasets we created in Section 5.4. Then, in the feature extraction phase, we followed the same steps in Section 5.5. We selected the best-performing feature sets for each classification task according to the previous results reported in Section 5.7. Thus, for the Baseline and Individual classifiers, we chose the following feature sets: C, L, PBC, and PBCL, while for the Academia and Hacker tasks, we chose two keyword-based feature sets: UC-IDF and UC-TFIDF.

6.6.1.3 Machine Learning Models

We re-trained the classifiers using the same original models, Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), SVM with linear kernel (SVM-L) and with RBF kernel (SVM-R). To see if we could get better results, we added two more models: Extra Trees (ET) and eXtreme Gradient Boosting (XGBoost).

6.6.1.4 Experimental Results

The training process was also done using the Scikit-Learn library with 5-fold stratified cross-validation. The training results are reported in Table 6.3. The results were the same for the first five models. As for the ET model, we noticed a similarity in performance across all the used feature sets compared to the RF model. This was expected as they are quite similar methods. In some cases, the ET model performed slightly better than the RF. The XGBoost model performed well for the Baseline classification task with the C, PBC and PBCL feature sets where the F1-score was 91%, similar to the RF and ET models. However, XGBoost was slightly ahead of all the other models (in terms of F1-score) for L, PBC and PBCL feature sets.

Looking at the LR and SVM-L model columns of the results, we notice that the F1-score is 0 for the feature sets used in the Hacker and Academia classifiers. Upon investigating the confusion matrix in each iteration of the cross-validation training process, we noticed that both TPs and FPs were 0. Consequently, precision, recall and F1-score are all 0, which means all samples were predicted as negative cases.

To summarise the results, we noticed that Random Forest and Extra Tree models

Table 6.3: Overall experimental results for the four old classifiers

	Feature set	#F	#S	DT			RF			ET			LR			XGBoost			SVM-L			SVM-R		
				F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec	F1	Prc	Rec
Baseline	C	9	1882	0.88	0.90	0.89	0.91	0.90	0.95	0.91	0.90	0.95	0.91	0.93	0.90	0.91	0.89	0.94	0.91	0.92	0.92	0.92	0.91	0.95
	L	93	1882	0.81	0.81	0.82	0.88	0.89	0.89	0.88	0.87	0.90	0.86	0.86	0.88	0.89	0.89	0.90	0.88	0.87	0.90	0.88	0.87	0.90
	PBC	56	1974	0.88	0.89	0.89	0.91	0.90	0.95	0.91	0.90	0.94	0.90	0.91	0.89	0.91	0.90	0.94	0.90	0.91	0.91	0.90	0.91	0.91
	PBCL	149	1974	0.88	0.88	0.89	0.91	0.90	0.95	0.91	0.91	0.94	0.90	0.91	0.91	0.91	0.90	0.94	0.91	0.91	0.92	0.90	0.91	0.91
Individual	C	9	937	0.72	0.73	0.72	0.79	0.80	0.78	0.79	0.81	0.78	0.81	0.78	0.84	0.78	0.80	0.76	0.81	0.79	0.83	0.82	0.83	0.80
	L	93	937	0.84	0.83	0.85	0.88	0.92	0.85	0.86	0.90	0.82	0.86	0.91	0.82	0.89	0.90	0.89	0.85	0.92	0.79	0.85	0.93	0.78
	PBC	56	957	0.81	0.79	0.83	0.86	0.90	0.83	0.85	0.89	0.82	0.87	0.87	0.88	0.87	0.89	0.86	0.87	0.88	0.87	0.86	0.88	0.84
	PBCL	149	957	0.84	0.84	0.84	0.89	0.91	0.87	0.88	0.93	0.84	0.89	0.90	0.88	0.91	0.92	0.90	0.89	0.91	0.87	0.87	0.91	0.83
Hacker	K_UC-IDF	200	314	0.78	0.74	0.90	0.88	0.85	0.94	0.89	0.86	0.94	0.00	0.00	0.00	0.79	1.00	0.66	0.00	0.00	0.00	0.24	0.93	0.14
	K_UC-TFIDF	200	314	0.74	1.00	0.59	0.82	1.00	0.70	0.85	1.00	0.75	0.00	0.00	0.00	0.75	1.00	0.60	0.00	0.00	0.00	0.66	0.50	0.94
Acad.	K_UC-IDF	200	245	0.81	0.68	1.00	0.90	0.82	1.00	0.92	0.85	1.00	0.00	0.00	0.00	0.82	0.69	1.00	0.00	0.00	0.00	0.43	0.71	0.58
	K_UC-TFIDF	200	245	0.68	0.72	0.83	0.76	0.75	0.88	0.77	0.75	0.88	0.00	0.00	0.00	0.60	0.76	0.70	0.00	0.00	0.00	0.28	0.31	0.37

performed well across all the classification tasks. As for the feature sets, we found that for the Baseline and Individual classification tasks, the PBCL feature set seemed to be a good and stable choice, while for the Academia and Hacker classifications, the UC-IDF keyword-based feature set was still the best option for these classifiers.

6.6.1.5 Predicting Classes

6.6.1.5.1 Best Classifier

Before using the trained classifiers, we must first select the best-performing classifiers, which means choosing the combination of best-performing models and feature sets according to the results in Table 6.3. Consequently, for the Baseline and Individual classification tasks, we selected the trained classifiers that were built using the PBCL feature set and the Random Forest model, while for the Academia and Hacker classification tasks, we used the trained classifiers that were built using the UC-IDF keyword-based feature set and the Extra Tree model.

6.6.1.5.2 Prediction Results

Next, we extracted the required feature sets for the data collection as described in Section 5.5. After that, the selected trained classifiers were used to predict the class of each Twitter account in the data collection according to each classification task.

Each classifier predicted whether an account was a positive or negative case. The prediction statistics are listed in Table 6.4.

Table 6.4: The prediction results for the four old classifiers

Task	Features	Model	#(Samples)	Prediction Samples	Positive	Negative
Baseline	PBCL	RF	42,028	42,028	9,377	32,651
Individual	PBCL	RF	42,028	9,377	4,795	4,582
Academia	K:UC-IDF	ET	42,028	9,018	4,990	4,028
Hacker	K:UC-IDF	ET	42,028	9,018	7,409	1,609

It is worth mentioning that for the Individual, Academia and Hacker classifiers, we applied them in a cascaded way to the positive cases predicted by the Baseline classifier. In other words, if a Twitter account was predicted as a positive case using the Baseline classifier, then we have to pass it to the other sub-classifiers to predict the output class in each classification task. Thus, the number of samples used for prediction in each sub-classifier equals the number of positive cases minus the number of accounts that do not have the required features, e.g., if the used feature set is keyword-based and a Twitter account has no public tweets, then this account will not be in the prediction samples. That is why the number of samples in the feature extraction phase for the Academia and Hacker classifiers dropped from 42,028 to 33,790, as 8,238 accounts have either protected or empty timelines.

When combining or cascading different classifiers, the error will propagate from the former classifier to the later one, meaning the former classifier would be considered biased to any later classifier. To avoid that, we should not use the predicted positives, but instead, we should use the true positives if we have the whole dataset labelled, which was not feasible in our case. Thus, we acknowledge this limitation and know that among the predicted negatives from the former classifier, there will be false negatives, which should have been passed to the later classifier.

6.6.1.6 Manual Evaluation

To know whether the prediction results were good or not and to evaluate the performance of the trained classifiers on the prediction dataset, we had to manually verify the predictions by selecting a subset of accounts for each classification task and manually labelling them. Then, we compared the actual vs predicted labels to calculate the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which all together form the confusion matrix. Next, we calculated Accuracy, F1, Precision, and Recall scores. The manual verification results are reported in Table 6.5, where each row is for a classification task. The number of samples manually verified in each task is indicated in the #S column.

Table 6.5: Manual validation results for the classifier predictions

Task	#S	TP	TN	FP	FN	Acc	F1	Prec	Rec
Baseline	1,154	900	63	87	104	0.83	0.90	0.91	0.90
Individual	1,003	535	281	37	150	0.81	0.85	0.94	0.78
Academia	172	25	35	110	4	0.34	0.30	0.19	0.86
Hacker	172	13	52	103	3	0.38	0.20	0.11	0.81

For the Baseline classifier evaluation, we randomly selected 1,154 samples. The F1-score was 90%, which means a 2% drop in performance compared to the F1-score from the original training/testing results, reported in Chapter 5. For the Individual classifier, we selected 1,003 samples, and the F1-score was 85%, representing a 5% drop in performance. However, considering the significant difference in size between the original training dataset and our prediction dataset (2k vs. 42k accounts) and the relatively small performance drop, we can confidently assert that both the Baseline and Individual classifiers are good enough for our case study.

On the other hand, for the Academia and the Hacker classification tasks, the F1-score dropped significantly from 91% to 30% for the Academia classifier and from 88% to 20% for the Hacker classifier. We think there are several reasons for

this drop in performance. First, the original datasets used for training the classifiers were small; where for the Academia dataset, we had 250 samples, and for the Hacker dataset, we had around 320 samples, see Sections 5.6.4 and 5.6.3. The small datasets limited the generalisability of these two classifiers. Second, the Academia classification task was neither straightforward nor clearly defined compared to the classification tasks in the case of the Baseline and the Individual classifiers.

6.6.2 Research Related Classifier

The focus of the case study centres on exploring the cyber security research network within ACEs-CSR. Cyber security researchers constitute a subset of accounts within the broader category of cyber security related accounts. Consequently, the existing Baseline classifier and other sub-classifiers are not enough to capture this distinct group. To address this gap, it was imperative to develop a dedicated and well-defined classifier tailored specifically to identify Twitter accounts associated with research activities in the field of cyber security.

For the Research classifier, we considered a data sample as a “positive” case if it is involved with any research work or activity related to research. This is judged based on the account’s description and timeline. This makes any cyber security researcher a positive case, even if they do not work in academia or are not associated with any research organisation. This is the significant difference between the “Research” and “Academia” classifiers. We also noticed that the labelling process by human experts for the former classifier compared to the latter one was more straightforward as its definition is more precise and less confusing.

6.6.2.1 Feature Extraction

In addition to the features extracted for the Baseline and Individual classifiers, we incorporated a distinct set of features tailored specifically for the Research classifier, denoted as the Research (**R**) group. These features were meticulously selected to encapsulate the distinctive attributes and content trends exhibited by accounts

associated with cyber security research, thus enhancing the classifier’s ability to identify them with precision. The following section elaborates on the new features.

6.6.2.1.1 Connections with Seeds

This is a metric of two values. The first one is the number of seed accounts that follow the account, whereas the second one is the number of seed accounts that this account follows (i.e. following). The number of connections a Twitter account has with the seed accounts might indicate how close/related this account is to what these accounts represent.

6.6.2.1.2 Verified

This metric is a binary value corresponding to the **Verified** profile attribute in the Twitter account, which is indicated with the blue check mark.

6.6.2.1.3 Researcher Keywords

We compiled two lists of keywords that can be used in the Twitter Name or Description fields and may refer to an account that is related to research. The keywords list for the Name field contains 13 entries while the one for the Description contains 27 entries, making a total of 40 keywords, listed in Table 6.6. The metric representing these features contains 40 binary values. Each one is either “0”, which means the corresponding keyword does not exist or “1”, which means this keyword was found.

6.6.2.1.4 Website Category

This metric is derived from the “Website” field of the Twitter account’s profile. Sometimes a link for a page can tell a lot about the Twitter account owner, as many people use this field to share a personal or professional website. The latter can be, for example, their LinkedIn page or the website of the organisation they work for.

We processed the URL found in this field, identified the host of each URL, and then used some regular expressions with manually created lists of hosts, main domains, and top-level domains to assign the parsed URL to a category from below.

Table 6.6: The Researcher Keywords in the Twitter Name & Description fields

Name field		Description field			
Academic	Research	Academic	PGR	RA	Researching
Faculty	Researcher	Doctor of Science	PHD	Reader	Scientist
PhD	Scientist	EngD	PHD Candidate	Research	University
PROF	University	Faculty	PHD Researcher	Research Assistant	Doctoral
Professor	Doctoral	Lecturer	PHD Student	Research Associate	Postdoc
RA	Postdoc	MPHIL	PROF	Research Fellow	Postdoctoral
Reader		MRES	Professor	Researcher	

- “**Research**” category represents a website more likely related to research, such as a university or a research institute. Some entries used in this category’s domain list are “.edu”, “.ac.”, “.academy”, “orcid.org”, and “scholar.google%”. We noticed that in some countries, universities do not have a unified domain. Thus, we used also a list of university hosts (Hipo, 2022) to capture as many cases as possible.
- “**Mixed**”: here, the website is not specifically related to research, but it might be. Some examples of the hosts and domains in this category are “linkedin”, “medium”, “github”, “.info”, “.net” and “.com”.
- “**Other**”: any other websites that are less likely related to research and do not fall under the previous two categories.

The number and names of the above categories can be adjusted according to the studied use case and the required classification question.

6.6.2.2 Training Dataset

We needed a dataset to train the Research related classifier before using it for prediction. Thus, we created a training dataset from the primary data collection described in Section 6.5 as follows. After using the Baseline classifier to predict the labels of the 42k accounts, we kept only the accounts that were predicted as cyber

security related accounts, i.e. we selected the “predicted” positive cases. Then, we randomly selected around 1,200 samples to label them manually.

The selection and labelling processes were made in iterations until we got enough labelled samples. In each iteration, we randomly selected 100 accounts and manually labelled them. In the final iterations, we had to alter the selection step by adding some simple criteria to increase the likelihood of getting positive cases. Otherwise, we would have spent significantly longer time on labelling without guarantee of finding enough positive samples. This was necessary to create a balanced training dataset. At the end of the labelling process, the positive and negative labelled samples were 500 and 700, respectively. We chose around 500 samples from each class to get a balanced dataset of 1k data samples.

6.6.2.3 Machine Learning Models

We followed a similar approach for this classification task where we trained and tested various ML models, as in Sections 5.7 and 6.6.1.3. We utilised the same seven ML models, including Extra Trees and XGBoost to ensure consistency in our classifier development process. These models were employed to assess the performance of the classifiers and determine their effectiveness in this new context.

6.6.2.4 Experimental Results

Using the same Python library Scikit-Learn we used before and the aforementioned models, we experimented with the following feature sets: R, P, PR, B, BR, C, CR, PBC, PBCR, PBCL, and PBCLR. We tried different feature sets to compare their performance scores and report which ones were the best for this new classifier. All the chosen models were trained and tested with 5-fold stratified cross-validation as we did for the other classifiers. The same performance metrics were used as well to report the results and make comparisons. The experimental results for the Research classifier are listed in Table 6.7. The numbers of samples and features were added to the table as well. A colour scale from red to green was used for the F1 and Precision

metrics to make visual comparison and selection easier.

6.6.2.4.1 F1-Score

Judging by the F1-score, the highest score was 83% which was achieved several times using different combinations of feature sets and ML models. As for the feature sets, the best performance was achieved using the following ones: R, BR, CR, PBCR, and PBCLR. We can notice that the R feature set was part of all the combined feature sets that achieved the highest F1-score, which means R features are quite related to the classification question of the Research classifier. The opposite can be said for the feature sets that do not have the R group, and as a result, they were not considered for the prediction phase of the Research classifier.

Regarding the seven ML models used, the best performance was achieved using the SVM-R and ET models. We can also notice that the RF model performed well, achieving 81% for the F1-score. The DT model came last in terms of performance.

Table 6.7: Experimental results of the Research classifier

Feature set	#F	#S	DT			RF			ET			LR			XGBoost			SVM-L			SVM-R		
			F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec
R	32	1003	0.78	0.94	0.67	0.81	0.94	0.72	0.81	0.94	0.71	0.82	0.97	0.72	0.81	0.94	0.72	0.82	0.97	0.72	0.83	0.96	0.73
P	25	1003	0.55	0.55	0.55	0.63	0.61	0.65	0.62	0.61	0.64	0.61	0.61	0.62	0.62	0.60	0.64	0.61	0.61	0.61	0.59	0.57	0.61
PR	57	1003	0.76	0.77	0.76	0.81	0.89	0.74	0.82	0.92	0.74	0.82	0.96	0.71	0.78	0.82	0.74	0.82	0.97	0.71	0.82	0.97	0.71
B	22	1003	0.60	0.61	0.59	0.69	0.69	0.68	0.69	0.68	0.69	0.67	0.62	0.72	0.66	0.66	0.66	0.69	0.62	0.79	0.69	0.63	0.75
BR	54	1003	0.76	0.78	0.75	0.78	0.80	0.77	0.82	0.89	0.77	0.82	0.97	0.72	0.79	0.81	0.76	0.82	0.97	0.71	0.83	0.96	0.73
C	9	975	0.60	0.59	0.61	0.64	0.64	0.65	0.62	0.60	0.63	0.64	0.59	0.71	0.62	0.61	0.63	0.67	0.58	0.79	0.67	0.60	0.76
CR	41	975	0.75	0.76	0.74	0.79	0.88	0.72	0.81	0.91	0.73	0.82	0.96	0.71	0.76	0.80	0.72	0.82	0.97	0.71	0.83	0.96	0.73
PBC	56	1003	0.62	0.62	0.63	0.68	0.69	0.67	0.67	0.68	0.67	0.68	0.65	0.71	0.68	0.68	0.68	0.67	0.65	0.71	0.66	0.66	0.66
PBCR	88	1003	0.74	0.76	0.74	0.79	0.80	0.79	0.83	0.90	0.77	0.82	0.95	0.73	0.79	0.82	0.78	0.82	0.97	0.71	0.82	0.97	0.71
PBCL	149	1003	0.63	0.62	0.64	0.70	0.71	0.70	0.70	0.70	0.71	0.70	0.68	0.74	0.71	0.71	0.71	0.70	0.67	0.74	0.69	0.66	0.73
PBCLR	181	1003	0.74	0.76	0.73	0.77	0.79	0.76	0.83	0.88	0.79	0.82	0.94	0.73	0.81	0.82	0.79	0.82	0.97	0.71	0.81	0.97	0.70

6.6.2.4.2 Precision

Although we wanted to select the best classifier based on the F1-score, we had to consider the **Precision** as well since it corresponds to the accuracy of the positive class (i.e., the research related account). By choosing Precision over Recall, we decided to prioritise false positives (FPs) over false negatives (FNs) since our OSN

analysis required working with positive samples and inspecting their profiles, timelines and connections. Moreover, since we were studying the communities resulting from positive samples, we needed the predicted positive samples to be more accurate and the FPs to be as minimal as possible.

Looking again at the experimental results of the Research classifier in Table 6.7, and judging by Precision, the highest score was 97% and achieved 11 times using six feature sets and three models. For the feature sets, the best-performing ones has the R features. The best-performing models were SVM-R, SVM-L, and LR.

6.6.2.5 Predicting Classes

6.6.2.5.1 Best Classifier

After training and testing the Research classifier using the labelled dataset, we needed to predict the classes of all the accounts in the data collection. Thus, we had to select the best-performing classifier based on the experimental results in Table 6.7 by considering both the F1-score and Precision metrics.

We noticed that the SVM-R was the best model, where F1-score reached 83%, and Precision score was 96% for the following three feature sets: R, BR and CR. There were other cases that are worth considering where Precision was 97%, and the highest F1-score was 82%. However, for the prediction of the research related accounts in our data collection, we selected the best classifier based on first F1-score and then Precision. Consequently, we selected the trained Research classifier built using the R feature set and the SVM-R model (F1-score = 83%, Precision = 96%).

6.6.2.5.2 Prediction Results

Since the Research classifier is also a cascaded classifier following the Baseline classifier, we only considered the positive samples (9,377) predicted by the Baseline classifier as the input for this classifier. This was necessary as the Twitter accounts that we wanted to capture were the cyber security related accounts that are also Research related ones.

Table 6.8: The prediction results using the Research classifier

Task	Features	Model	#(Samples)	Prediction Samples	Positive	Negative
Baseline	PBCL	RF	42,028	42,028	9,377	32,651
Research	R	SVM-R	42,028	9,377	1,684	7,693

To achieve this, we took the positive samples from the Baseline classifier (9,377 samples) and extracted the required feature set (R) as described in Section 6.6.2.1. Then, the selected trained classifier, SVM-R, was used to predict the class of each account in the prediction samples. The prediction statistics are shown in Table 6.8. We got 1,684 positive cases and 7,693 negative cases. After that, we carried out some manual verification for the prediction results by inspecting the Twitter account profiles and timelines. Thus, some FNs were captured, and the final number of the cyber security related and research related accounts was 1,817.

6.7 Social Network Analysis

In this section, we explored the social structure of the ACEs-CSR research network on Twitter. To do this, we constructed a social graph representing the connections between accounts in this network. We then applied graph-based analysis techniques to identify and analyse the communities within the social graph. By examining the connections and relationships among the graph nodes, we gained insights into the social structure and community organisation of this network.

6.7.1 Social Graph Construction

The construction of the social graph for the ACEs-CSR network involved a systematic two-step process. Firstly, we focused on identifying the nodes within the graph, and secondly, we established the connecting edges between them. Node identification relied on the utilisation of the ML classifiers explained earlier in this chapter

(see Section 6.6). These classifiers were instrumental in pinpointing Twitter accounts that were relevant to both cyber security and research. Turning to the edges, all the connections detailed in Section 6.5.2 were subjected to a filtering process, in which only connections with both ends in the selected nodes were retained to represent the the follower/friend connections between the selected nodes.

Table 6.9: The different graphs within the ACEs-CSR Twitter network

Nodes	Edges	Graph
19	87	ACE-CSR seed accounts graph (Lv1 \rightleftharpoons Lv1)
42,028	3,301,730	ACE-CSR graph ($\{Lv1, Lv2\} \rightleftharpoons \{Lv1, Lv2\}$)
9,377	1,144,800	ACE-CSR cyber security graph
1,817	64,826	ACE-CSR cyber security research graph

Within the ACEs-CSR network, several distinct graphs can be distinguished, as exemplified in Table 6.9. First, the simplest graph which represents solely the seed accounts (Lv1), encompassing 19 nodes and 87 edges. The second graph is the comprehensive ACEs-CSR social graph, incorporating nodes from both Lv1 and Lv2, along with their connecting edges, totalling 42,028 nodes and 3.3m edges. The subsequent two graphs were derived using the developed ML classifiers. For the third graph, we applied the Baseline classifier to filter the full ACEs-CSR social graph, identifying nodes (i.e., accounts) related to cyber security and preserving their connecting edges. This yielded a graph with 9,377 nodes and 1,1m edges. Lastly, the fourth graph, which is the primary focus of this study. By using the Research-related classifier to identify nodes within the cyber security research field, a new graph resulted, which contains 1,817 nodes and 64,826 edges, representing the ACEs-CSR cyber security research network. For simplicity, we will refer to the ACEs-CSR cyber security research graph as the ACE-CSR graph.

6.7.2 Graph Visualisations

The constructed ACE-CSR OSN graph was visualised using Gephi (Bastian, Heymann, and Jacomy, 2009). In Figure 6.1, we can see a random view of the ACE-CSR graph, and it can be noticed that the graph is large. The node sizes are scaled using their in-degree centrality. We can clearly notice a few ACE-CSR nodes that are bigger than the rest of the ACE-CSR nodes. Using the predicted labels from the Individual classifier in Section 6.6.1, the node shape can be either a triangle (individual account) or a circle (non-individual account).

To have a better and simpler overview of how the ACEs-CSR nodes are connected to the other nodes in the graph, we created another visualisation of the graph where only two colours were used for the graph nodes, one for ACEs-CSR nodes and the other colour for the remaining nodes. Then, we used Force Atlas 2 (Jacomy et al., 2014) layout¹ on Gephi software, see Figure 6.2.

6.7.3 Network Statistics

Social Network Analysis techniques were employed to analyse the ACE-CSR network of friends and followers. In Table 6.10, we present several basic network statistics that provide insights into the structure and characteristics of the ACEs-CSR social graph. These statistics offer a quantitative overview of the network, allowing us to understand its size, density, and connectivity patterns.

Graph Density: The graph density measures the proportion of existing connections to the total possible connections. For more details, refer to Section 2.2.3.1. The graph density for the ACEs-CSR graph is calculated to be 0.020, indicating a relatively sparse network. This is expected as the researcher's graph is quite small compared to the original graph before applying the ML classifiers.

Network Diameter: The network diameter measures the longest shortest path between any two nodes in the graph. For more details, refer to Section 2.2.3.2.

¹Layout custom settings: Behavior Alternatives (Dissuade Hubs, LinLog mode, Prevent Overlap), Tuning (Scaling: 2.0, Stronger Gravity, Gravity: 1.0)

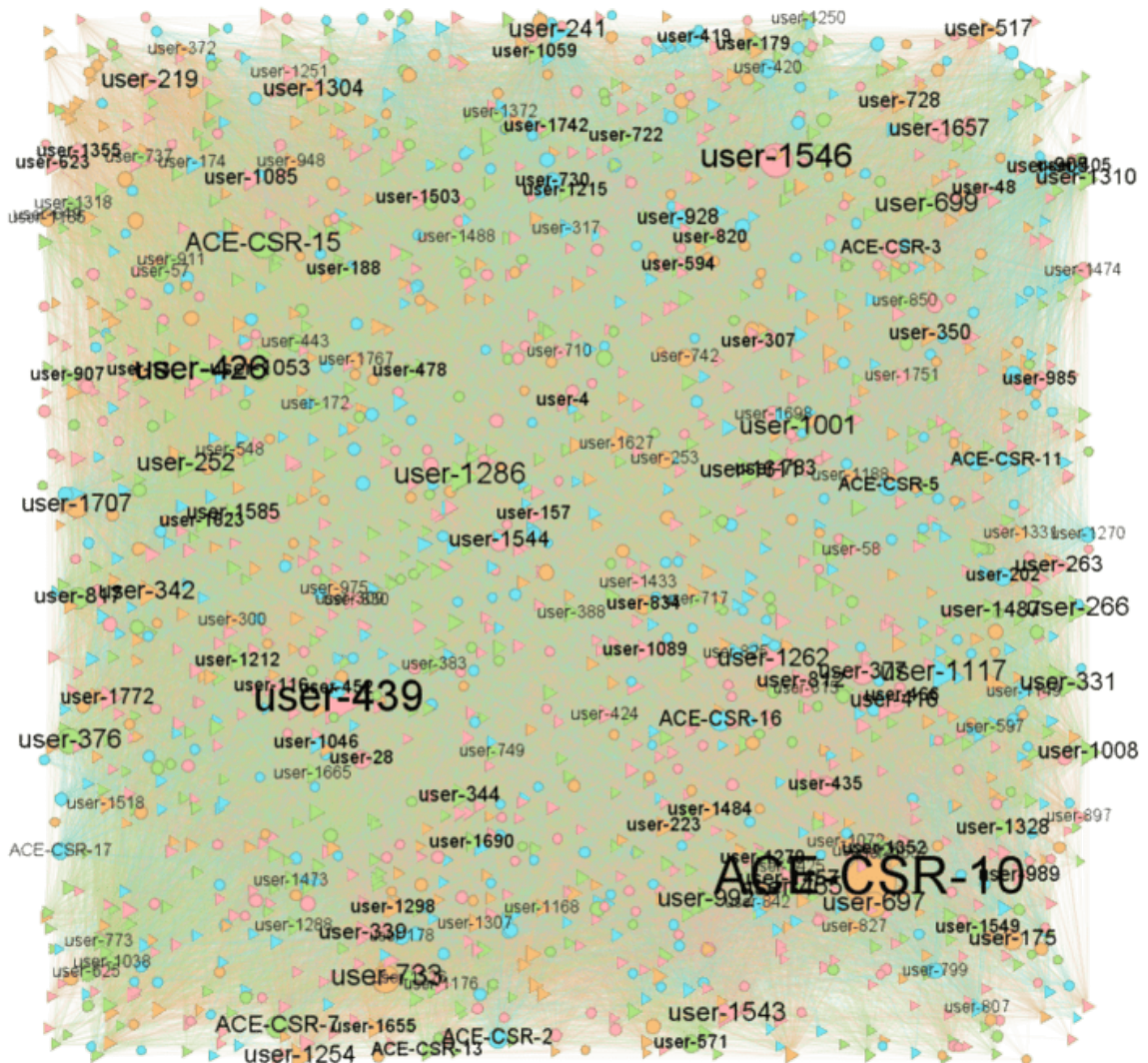


Figure 6.1: The Constructed Graph of ACEs-CSR Network (Random View)

The network diameter for the ACEs-CSR graph is 7, suggesting that information or influence can spread relatively efficiently within the network.

Average Path Length: The average path length signifies the mean count of edges connecting any two nodes within the graph. Refer to Section 2.2.3.3 for more details. Notably, in the ACEs-CSR graph, the computed average path length is 2.727. This implies that, on average, nodes are connected by less than 3 edges, reflecting swift information dissemination across the network.

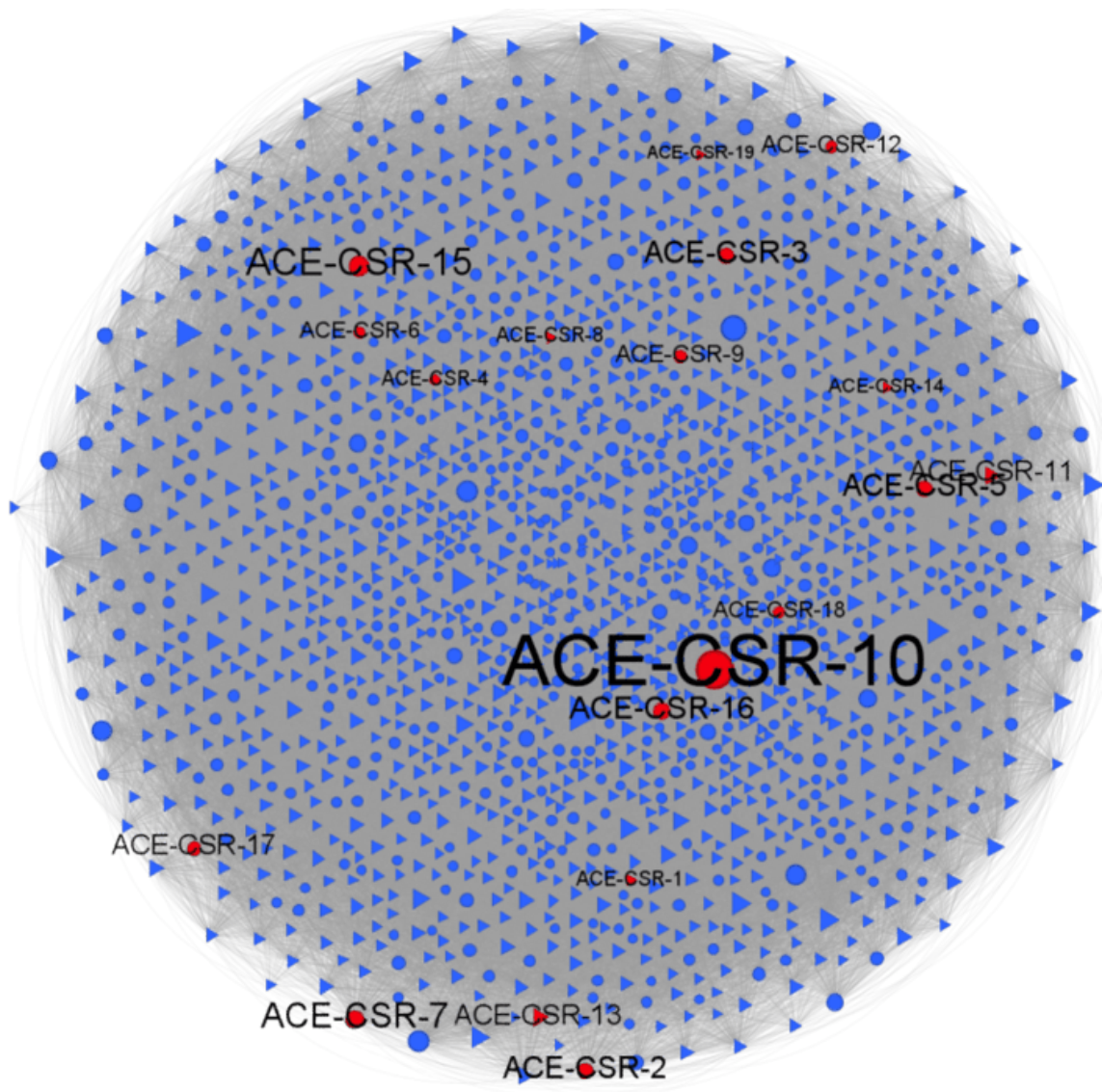


Figure 6.2: The Constructed Graph of ACEs-CSR Network (Focused View)

The average degree, a measure of the average number of connections per node, is 35.677, suggesting a relatively high level of connectivity in this network. Finally, the average clustering coefficient, which measures the degree to which nodes tend to cluster together, is 0.214, indicating the presence of clustering or community structures in the network. These metrics provide valuable insights into the network's structure, connectivity, and clustering patterns, shedding light on the dynamics and interactions within the ACEs-CSR research community on Twitter.

Table 6.10: Network Statistics of the ACE-CSR directed graph

Attribute	Value	Attribute	Value
Number of Nodes	1,817	Average Degree Centrality	35.677
Number of Edges	64,826	Average Path length	2.727
Graph Density	0.254	Average Clustering Coefficient	0.214
Network Diameter	7	Weakly Connected Components	1
		Strongly Connected Components	186

6.7.4 Community Detection & Analysis

Identifying the communities within the ACE-CSR network on Twitter is crucial for several reasons. Firstly, it provides a deeper understanding of the network's structure, revealing how cyber security experts interact and form clusters based on shared interests, affiliations, or geographical locations. By studying these communities, we can uncover patterns of collaboration, information exchange, and influence dynamics, which can inform strategies for fostering collaboration and knowledge sharing within the cyber security domain.

Studying these communities allows us to identify key individuals or groups that play central roles within the network. These central nodes may act as hubs for information dissemination, opinion leadership, or resource mobilisation, making them influential actors within the cyber security community. Understanding the position and influence of these nodes can help identify potential collaboration opportunities, leverage expertise, and amplify the impact of cyber security initiatives.

The message from studying the communities within the ACEs-CSR network on Twitter is that cyber security is a highly interconnected and collaborative field, with diverse communities of experts contributing to its advancement. By analysing these communities, we can gain valuable insights into the structure, dynamics, and potential influence of the cyber security landscape, enabling more effective collaboration, knowledge sharing, and strategic decision-making within the field.

In the subsequent sections, we outline our investigation into the communities within the ACEs-CSR social graph on Twitter. Initially, we introduce established and dependable algorithms for community detection. Subsequently, we detail our selection process for the appropriate algorithm and parameter configuration tailored to our study. Following this, we present the identified communities and discuss notable insights and findings obtained from our analysis. Furthermore, we delve into additional analyses of the detected communities, focusing on individual ratios and geographical distribution.

6.7.4.1 Community Detection Algorithms

The identification of communities holds significant value within the realms of sociology, biology, and computer science, where systems are frequently depicted as graphs. This challenge remains highly complex and has not been fully resolved, despite substantial endeavours from a broad, interdisciplinary community of researchers who have been dedicated to it in recent years (Fortunato, 2010). With the advance of OSNs' size and complexity, identifying and finding users who form communities within an extensive network became a challenge, which in turn, enables a wide range of SNA applications, such as maximising the diffusing of new ideas or technologies, improving recommendations suggestion and increasing the accuracy of expert finding tasks (Moscato and Sperli, 2021).

To study the big ACEs-CSR graph, we had to break it down into sub-graphs, where each graph represents a community (i.e., cluster) or a group of Twitter accounts that have something in common. A community in a network is defined by (Newman and Girvan, 2004) as a subset of nodes that are densely connected with each other but at the same time have a few connections to other network nodes. Since the graph nodes have no ground truth labels of any characteristic, using supervised classifiers to group or cluster these nodes was impossible. This is normal in such cases as we do not know the number of communities and whether they are roughly equal in size when we want to break a network into communities (Newman

and Girvan, 2004). As a result, we used unsupervised clustering techniques to divide the graph nodes into clusters (i.e., communities).

We tested several community detection algorithms that are widely adopted in the literature. First, we tried DBSCAN (Schubert et al., 2017), but it did not work with our dataset as the clustering results were not as good as the other methods. Then, we tried the Girvan-Newman algorithm (Girvan and Newman, 2002). Despite the long processing time, the results were also not good as sometimes it clustered all nodes in one cluster. After that, we examined modularity optimisation based algorithms as modularity is a well-known method for community detection (Newman and Girvan, 2004). We started to get good results using the Louvain algorithm (Blondel et al., 2008). However, due to some limitations in this algorithm (e.g., yielding arbitrarily poorly connected communities), we used the Leiden algorithm (Traag, Waltman, and van Eck, 2019) instead. These two algorithms use a resolution parameter (Lambiotte, Delvenne, and Barahona, 2014), which controls the size of the detected communities (Newman, 2016).

6.7.4.2 Community Detection in ACEs-CSR graph

Figure 6.3 shows the results of applying the Leiden algorithm on the ACEs-CSR graph with the resolution set to 1. The modularity was 0.406, and four communities were discovered in total. The statistics about these communities are listed in Table 6.12. The node size and label size are proportionate with the in-degree centrality score of the node. To emphasise the clusters, we used four colours, each dedicated to a cluster. Also, we placed the nodes that belong to the same cluster next to each other using the Circle Pack (Bostock, 2022) layout² in Gephi software.

Increasing the resolution parameter γ in the Leiden algorithm results in more communities while reducing it does the opposite (Traag, Waltman, and van Eck, 2019). To illustrate this, we presented six instances of applying the Leiden algorithm in Figure 6.4, using the following γ values: 0.5, 1, 1.5, 2, 2.5 and 3. The node size and

²Layout custom settings: Hierarchy was set to Cluster (Attribute).

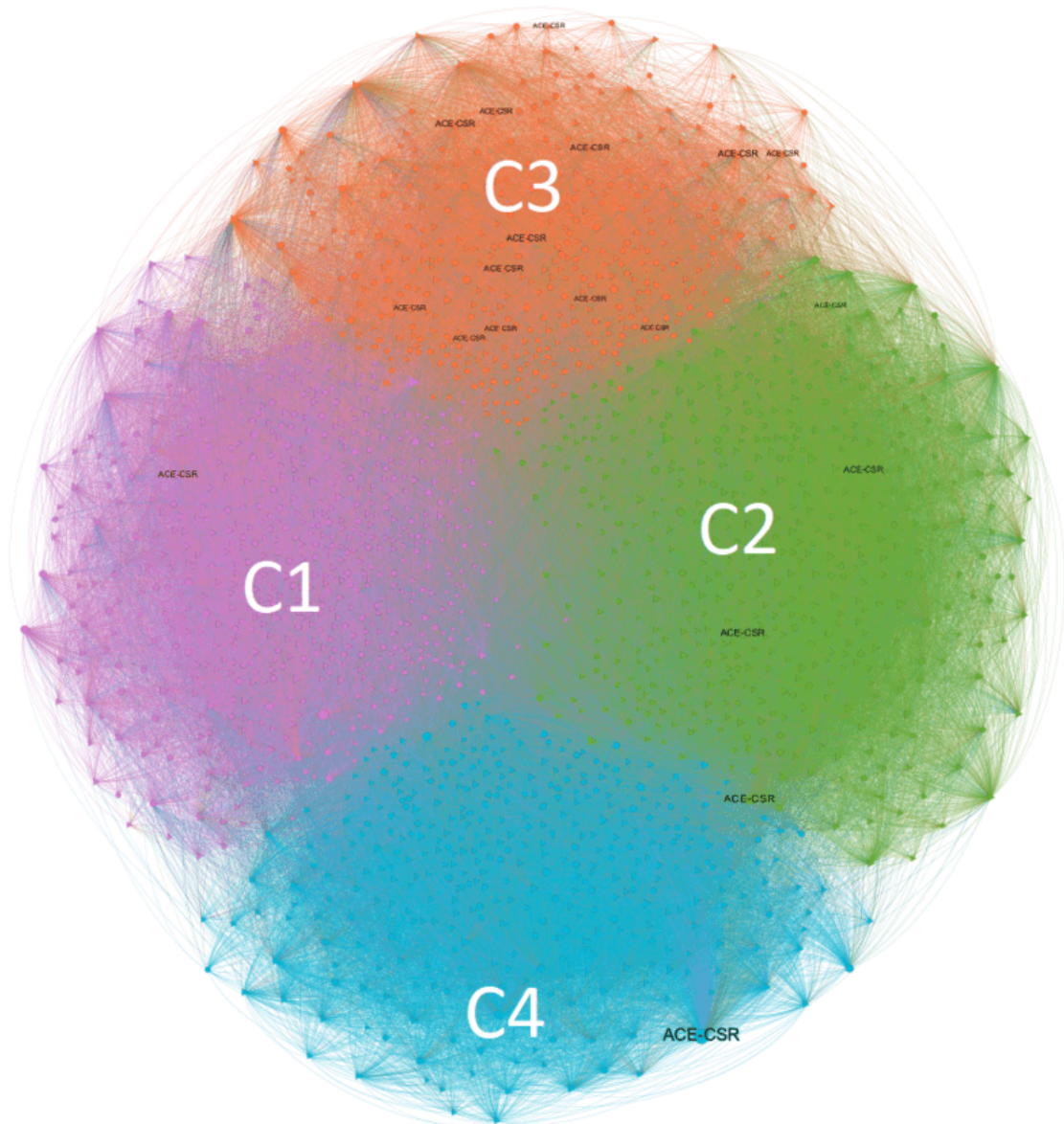
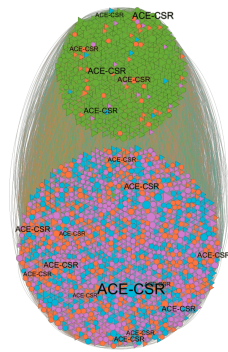
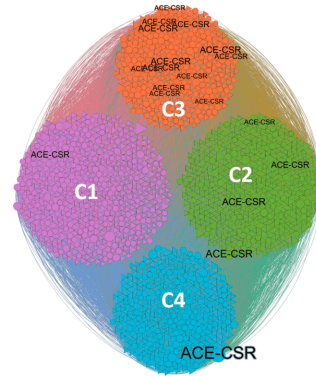


Figure 6.3: The communities within the ACEs-CSR social graph ($\gamma = 1$)

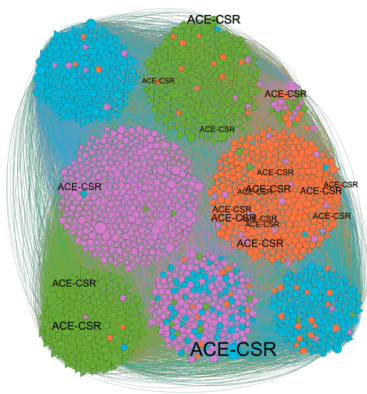
the label are proportionate with its in-degree centrality score. Also, we grouped the nodes that belong to the same cluster together using the Circle Pack (Bostock, 2022) layout with the “hierarchy” set to the “cluster” attribute in Gephi. To emphasise the size and members of the clusters, we used a distinctive colour for each cluster’s nodes when γ was 1. Then, we preserved these colours in the next applications of the Leiden algorithm to understand how these communities merge when the modularity



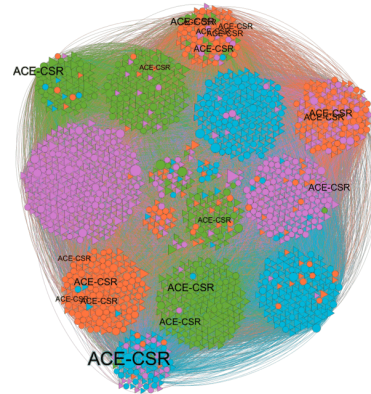
(a) $\gamma = 0.5$, $M = 0.566$, $C = 2$



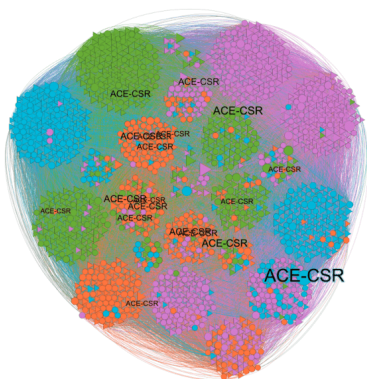
(b) $\gamma = 1$, $M = 0.406$, $C = 4$



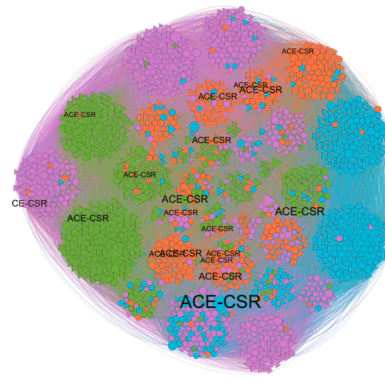
(c) $\gamma = 1.5$, $M = 0.305$, $C = 9$



(d) $\gamma = 2$, $M = 0.244$, $C = 18$



(e) $\gamma = 2.5$, $M = 0.206$, $C = 28$



(f) $\gamma = 3.0$, $M = 0.175$, $C = 40$

Figure 6.4: Six different visualisations of the ACEs-CSR network with different clustering parameters (C : the number of communities, M : modularity)

increases (i.e. case γ was 0.5), and how they split and create new sub-communities when the modularity decreases due to the increase in resolution.

For example, consider the two instances of applying the Leiden algorithm; the first one is resolution equals 1.0, represented in Figure 6.4b, and the second one is resolution equals 1.5, see Figure 6.4c. By comparing the two figures, we noticed that the four communities were split into sub-communities and small mixed sub-communities were formed as well. Thus when the resolution changed from 1.0 to 1.5, the number of communities increased from 4 to 9. On the other hand, we decreased the resolution from 1 to 0.5, three of the four communities were merged together to form one bigger community, and a few nodes were moved from these communities to the fourth original community.

Selecting the right resolution depends on how many communities we want to work with. Analysing hundreds of communities manually would be impossible, and analysing two or three communities would be less indicative. Thus, to select the appropriate resolution and the number of communities for our use case study, we experimentally tested how the modularity and number of discovered communities change with the γ , see the reported results in Table 6.11. We started with a resolution equal to 0.1 and increased it with a step of 0.1 until we reached 2, where we increased the step value. As expected, the number of communities increased with the resolution increase, but the modularity decreased.

As for the analysis of the detected communities, we could not list all the trials we had with each reasonable candidate resolution and its corresponding communities. Instead, we listed below a few examples of the insights we learned about the ACEs-CSR network and the sub-communities we discovered, see Figure 6.4.

A) Initially, we expected each ACE-CSR account to have a strong community around its node in the graph, but this was not the case for a few of them unless the modularity was significantly reduced. However, that would not reflect a strong and densely connected community. One of the explanations for this is that some ACE-CSR seed accounts are not well connected to other cyber security researchers.

Table 6.11: Resolution, Modularity and Communities using Leiden Algorithm

γ	Modularity	Communities	γ	Modularity	Communities
0.1	0.900	1	1.7	0.279	14
0.2	0.800	1	1.8	0.266	14
0.3	0.700	1	1.9	0.253	15
0.4	0.621	2	2.0	0.244	18
0.5	0.566	2	2.5	0.206	28
0.6	0.529	3	3.0	0.175	40
0.7	0.495	3	3.5	0.152	46
0.8	0.462	3	4.0	0.131	64
0.9	0.429	4	5.0	0.101	100
1.0	0.406	4	6.0	0.080	130
1.1	0.381	5	7.0	0.066	169
1.2	0.358	6	8.0	0.056	215
1.3	0.341	7	9.0	0.047	244
1.4	0.321	9	10.0	0.040	282
1.5	0.305	9	15.0	0.018	437
1.6	0.292	11	20.0	0.005	559

B) We noticed that some ACE-CSR nodes always appear in the same cluster regardless of the chosen granularity level (controlled by γ parameter). After manually inspecting several cases, one explanation for this might be that these ACEs-CSR are geographically close. We also had some personal observations about this, where we noticed that researchers across these ACEs-CSR have worked together. In two particular cases, some researchers moved from one ACE-CSR to another.

C) Using different values for the resolution parameter and checking the resulted communities each time, we observed some clusters that do not have any ACE-CSR nodes (see Figure 6.4c). We inspected these communities and checked their member Twitter profiles. We noticed they are also densely connected and represent a mix of

national, European, and international research institutions.

To maintain simplicity, and consistency, and to enhance explainability, our analysis will primarily focus on the communities identified using $\gamma = 1$. This resolution corresponds to a specific level of granularity in the community detection process. By focusing on the resolution being set to 1.0, we can effectively examine and analyse the main communities within the ACE-CSR network. Figure 6.3 provides a visual representation of these communities, allowing for a clearer understanding of their structure and composition. The properties of the corresponding four communities discovered in this case can be found in Table 6.12.

Table 6.12: Statistics of discovered communities ($\gamma = 1$)

Community	Colour	Members	Size	Individuals	Non-individuals
C1	Purple	595	32.75%	72.61%	27.39%
C2	Green	465	25.59%	79.14%	20.86%
C3	Orange	382	21.02%	51.83%	48.17%
C4	Blue	375	20.64%	70.13%	29.87%

6.7.4.3 Clusters Analysis – Individual Members

Knowing the percentage of individuals in the ACEs-CSR network is interesting as it gives insights into how many cyber security individual researchers these ACE-CSR accounts attracted on Twitter and how many non-individuals e.g., research centres, universities, and companies are connected to these ACEs-CSR. The “Individual” attribute was obtained using the Individual classifier (see Section 6.6.1). The percentages of individual and non-individual accounts in the graph were 69.40%, and 30.60%, respectively. Using the four communities in Figure 6.3 as an example, we calculated the individual/non-individual percentages of each community and showed the results in Table 6.12. Notably, the individual percentage reached 79.14% for Community C2, which is higher than the other communities. Upon inspecting C2,

we found that individuals in this community are often well-known researchers and figures in the cyber security research domain.

6.7.4.4 Clusters Analysis – Location

Analysing communities based on geographical location can provide valuable insights into the distribution, composition, and interactions of cyber security experts across different regions, cities, countries and even continents. This is quite important for several reasons, such as identifying regional expertise, understanding global collaboration, assessing regional vulnerabilities and threats, supporting policy and decision-making, and finally promoting diversity and inclusivity.

Geographical analysis allows us to identify clusters of experts in specific locations. This can help in understanding regional expertise, specialisation, and research focus areas within the cyber security domain. By examining the connections between experts from different regions, we can gain insights into global collaboration patterns. Understanding how experts from different parts of the world interact and collaborate can inform strategies for fostering international cooperation in cyber security research and practice.

Cyber security threats and vulnerabilities may vary based on geographical location due to factors such as regulatory frameworks, technological infrastructure, and geopolitical tensions. Analysing communities based on location can help identify regional cyber security challenges and inform targeted mitigation strategies. Furthermore, such analysis can provide policymakers and decision-makers with valuable information for shaping policies, allocating resources, and developing strategies to address cyber security challenges at regional and global levels.

The account’s “Location” field is optional on the Twitter user profile, so not all account holders provide such information. We calculated some statistics about the Location field in our data collection; see Table 6.13. The percentage of the accounts with the information provided in the whole data collection was 61.41%, while it was 77.55% for the ACEs-CSR research network. This higher percentage

indicates that cyber security research related accounts had a tendency to use this field more often compared to other users. We analysed the ACE-CSR communities based on their members' declared locations, hoping to gain more insights into how these communities were formed in the first place or what they represent.

Table 6.13: Twitter location field statistics in the studied datasets

Dataset	Total	Missing Location	Provided Location
Data Collection	42,030	16,220 (38.59%)	25,810 (61.41%)
Research Graph	1,817	408 (22.45%)	1,409 (77.55%)

6.7.4.4.1 Location Information Extraction

The “location” field is a free-formatted text where users can write anything they like. We observed names of places (e.g., towns, cities, countries, or even non-existing places), names of affiliations, GPS coordinates, postcodes, country codes (alphabetic such as “GB” and numeric such as “+44”), and Unicode symbols of national flags. Considering the different ways to indicate location information, we had to use a set of methods to extract such information.

For some “location” fields representing the location information as GPS coordinates, country codes and national flag symbols, we could extract such information using bespoke Python scripts. For other “location” fields that could not be processed using the previous method, we pre-processed them by removing any email address(es), URL(s), Twitter handle(s), special ASCII character(s), IP address(es)³ and isolated number(s), and then fed them to the Location Tagger Python library (Soni, 2022) to extract possible location information. Location Tagger is a process of tagging place names and filtering them out from the entities found using the Named Entity Recognition (NER) technique from text mining. The extracted

³IP addresses can sometimes carry location-related information. We considered such information less reliable and too complicated to process, so we decided to exclude it.

location information was automatically checked against cities’ names downloaded from the GeoNames website (GeoNames, 2022) to resolve the ambiguity that is usually raised when detecting location from free-formatted texts. Finally, in about 10% of “location” fields, the above-automated methods could not produce any location information, so we had to inspect them to recover such information manually.

It is worth mentioning that some accounts declared several locations, e.g. “London, Paris”, which we solved by choosing the first mentioned location. Also, while working with the location records, we had to deal with the ambiguity of a place name that can be resolved to several places, e.g. “Canterbury” is a UK city and a suburb in Australia. When we were resolving these cases, we prioritised the locations inside the UK, relying on the cities’ names list.

6.7.4.4.2 Location Statistics & Insights

Based on all the extracted location information of all accounts in each cluster, we calculated several geographical statistics about the nodes in the ACE-CSR network. Table 6.14 shows continent-specific statistics of the four communities shown in Figure 6.3. For comparison between communities, we visually represented the geographical statistics of these communities in Figure 6.5. We split Europe into two sub-groups, the UK, and Europe without the UK, to know which communities were more national (UK) or international (non-UK) from the perspective of ACEs-CSR.

Table 6.14: ACE-CSR community members distribution per continent

Community	UK	Europe-UK	N. America	S. America	Asia	Africa	Australia	Unknown
C1	95	53	132	9	55	16	13	222
C2	80	133	82	8	19	3	7	133
C3	241	21	11	0	6	1	10	92
C4	120	68	41	2	16	5	13	110

The location-based analysis revealed several interesting insights about the communities discovered in the ACEs-CSR network as follows.

A) First, for the four communities in Figure 6.3, Community C3 seems a more UK-centric one, while the remaining three communities exhibit a high level of international representation. Communities C1 and C2 are dominated by non-UK accounts, where most accounts were from North America for C1 and from the non-UK part of Europe for C2.

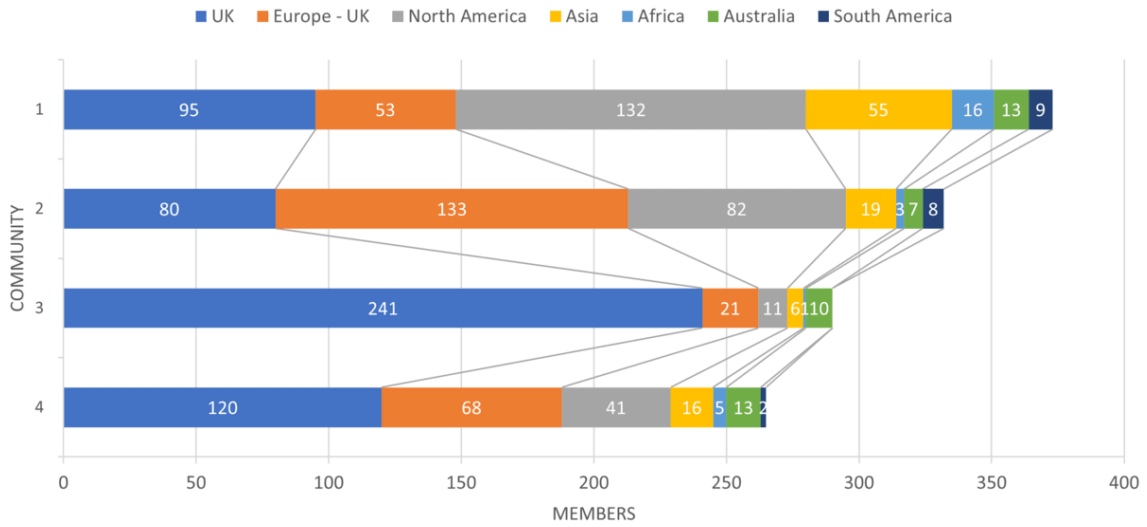


Figure 6.5: Location distribution per continent for ACE-CSR communities

B) Interestingly, a closer examination of all communities reveals a disparity in the representation of different geographic regions. Specifically, there is a notable underrepresentation of accounts from Africa, Australia, and South America, suggesting a bias towards connections with Europe, North America, and Asia.

C) Furthermore, among the communities, Community C1 stands out as the most internationally diverse cluster, with almost an equal number of accounts from Europe (excluding the UK) and Asia. The percentage of Asian accounts in C1 is substantially higher than the other three communities, indicating it may be the one representing the UK-Asia links.

D) Finally, when considering the ratio of UK accounts to non-UK accounts, Community C4 emerges as a more balanced cluster, with a roughly equal proportion of national and international accounts.

6.7.5 Influence Analysis

To investigate the influence within the ACEs-CSR research network, we employed SNA techniques, as described in Section 2.2. SNA allows us to study the relationships and interactions between cyber security research-related accounts in the network. As a proxy for quantifying influence, we utilised network centrality measures (explained in Section 2.2.4). Various centrality measures were applied in this analysis to capture different aspects of influence within the network.

The examination of influence in the ACEs-CSR research network was driven by several objectives. Firstly, our interest was in identifying the top influencers in this network. Discovering key influencers could provide insights into their contributions and impact on the cyber security research community on Twitter. Moreover, we sought to analyse the influence of the selected seed accounts (i.e., the ACE-CSR accounts) in the overall network, evaluating their roles in shaping discussions and interactions within the cyber security research community around them. Secondly, we aimed to explore the distribution of influence within the network, investigating whether it is evenly distributed among cyber security research-related accounts or if a small percentage of accounts holds a disproportionately high level of influence. This analysis provides a deeper understanding of the network's topology and the roles different nodes play within it.

6.7.5.1 Identifying Top Influential Nodes

As each centrality assesses the network position of a node differently (see Section 2.2.4), we used several centrality metrics to determine the influential nodes (users) according to each method. We used the following metrics: degree centrality (and we distinguished between the in-degree and out-degree scores), closeness centrality, betweenness centrality, eigenvector centrality, and PageRank centrality. We calculated the centrality scores for each node in the graph using the NetworkX Python library (Hagberg, Schult, and Swart, 2008). All the scores were normalised, and nodes were sorted according to their scores. The top 40 nodes according to each

used centrality method were listed in Table 6.15.

Each centrality measures the node's importance within the network from a distinct perspective. Looking at Table 6.15, we can notice that for most of the centralities we used, the top nodes were mixed between ACE-CSR accounts and other accounts. For example, for the Betweenness centrality, the first node was user-241, which is a non-ACE-CSR account, while For the Eigenvector centrality, the first and second nodes were user-439 and user-426, which are also non-ACE-CSR accounts. We analysed the scores of each centrality below, along with their implications and top influential nodes.

Degree Centrality indicates how many connections a node has. Nodes with high degree centrality scores act like hubs in the network, connecting to numerous other nodes. For instance, ACE-CSR-10, user-241 and user-439 have high degrees, suggesting that they are well-connected and can directly reach many other nodes.

In-degree Centrality focuses on incoming connections only. ACE-CSR-10, user-439 and user-426 have the highest in-degree centrality, which means that they are receiving interactions and attention from a significant number of other nodes. This could indicate their prominent positions in the network.

Out-degree Centrality looks at outgoing connections only. ACE-CSR-5, user-241 and user-1372 have the highest out-degree centrality, implying that they are actively reaching out to other nodes or sharing content.

Closeness Centrality emphasises how quickly a node can reach other nodes. ACE-CSR-10, user-439 and user-426 have the shortest average path lengths to others, meaning that they are the most connected and can spread information more efficiently through the network.

Betweenness Centrality identifies nodes that act as bridges between different parts of the network. The following nodes: user-241, ACE-CSR-10 and ACE-CSR-5 have the highest betweenness centrality, suggesting that they play a crucial role in maintaining connections between various groups of nodes. These nodes are essential for the overall network to be cohesive.

Table 6.15: Top 40 nodes ranked by different centrality measures

Rank	Degree	In-Degree	Out-Degree	Closeness	Betweenness	Eigenvector	PageRank
01	ACE-CSR-10	ACE-CSR-10	ACE-CSR-5	ACE-CSR-10	user-241	user-439	ACE-CSR-10
02	user-241	user-439	user-241	user-439	ACE-CSR-10	user-426	user-439
03	user-439	user-426	user-1372	user-426	ACE-CSR-5	ACE-CSR-10	user-426
04	ACE-CSR-5	user-1546	user-1046	user-1117	user-219	user-1117	user-1286
05	user-252	user-1286	user-252	user-1546	ACE-CSR-3	user-266	user-1546
06	user-1117	user-733	user-905	user-1254	user-697	user-1286	user-376
07	user-426	user-1117	user-1250	user-342	user-252	user-342	user-733
08	user-697	user-1001	user-1328	user-1707	user-905	user-699	user-266
09	user-219	user-1543	user-253	user-241	user-1117	user-1254	user-1310
10	user-1372	user-266	user-1298	user-1286	ACE-CSR-7	user-252	user-1117
11	user-1546	ACE-CSR-15	user-13	user-517	user-1372	user-992	user-465
12	user-905	user-699	user-219	ACE-CSR-7	user-1254	user-465	user-1001
13	user-1254	user-465	user-596	user-905	user-812	user-331	user-377
14	user-342	user-697	user-1331	user-733	user-1046	user-517	user-1657
15	user-1046	user-376	user-697	user-697	user-1544	user-241	user-342
16	user-1328	user-241	ACE-CSR-3	user-1304	user-342	user-1611	user-699
17	user-331	user-1262	user-1117	user-1457	user-1310	user-733	user-1543
18	user-1008	user-252	user-342	user-1543	user-1328	ACE-CSR-15	user-252
19	user-266	user-1254	user-1601	user-928	ACE-CSR-16	user-1304	user-625
20	user-13	user-331	user-1254	user-376	user-4	user-905	user-992
21	user-733	user-219	user-730	ACE-CSR-16	user-1008	user-1008	ACE-CSR-15
22	user-1286	user-1707	user-1008	user-1001	user-897	user-697	user-1585
23	user-1298	user-992	user-776	user-728	user-13	user-1585	ACE-CSR-7
24	user-253	user-342	user-331	ACE-CSR-2	user-1298	user-344	user-1707
25	ACE-CSR-3	ACE-CSR-7	user-239	user-266	user-1538	user-1352	user-1611
26	ACE-CSR-7	user-1611	user-174	ACE-CSR-15	user-1331	user-1487	user-1487
27	user-1250	user-1304	user-1335	user-992	user-253	user-1707	user-219
28	ACE-CSR-15	user-1008	user-949	user-405	user-730	ACE-CSR-7	user-517
29	user-699	user-517	user-29	user-219	user-174	user-1690	user-697
30	user-1304	user-416	user-471	user-1611	user-331	user-13	user-339
31	user-992	user-377	user-1488	user-817	ACE-CSR-17	user-1484	user-344
32	user-730	user-1657	user-1684	user-419	user-817	user-376	user-331
33	user-1544	user-1053	user-594	user-1262	user-989	user-1655	user-623
34	user-1331	user-339	user-630	user-699	user-266	user-728	user-175
35	ACE-CSR-16	user-817	user-1622	user-1310	user-1355	user-1457	user-905
36	user-1543	user-1310	user-168	user-1657	user-263	user-1742	user-1262
37	user-1611	user-1457	user-1717	user-1484	user-405	user-405	user-1254
38	user-1262	user-175	user-233	user-465	user-1089	user-1328	user-157
39	user-1001	user-1544	user-1077	user-1772	user-48	user-820	user-241
40	user-465	user-263	user-404	user-344	user-426	user-350	user-1008

Eigenvector Centrality considers not only a node's direct connections but also the connections of its neighbours. The following nodes: user-439, user-426, and ACE-CSR-10 have the highest eigenvector centrality because they are connected to other well-connected nodes. In essence, nodes with a higher eigenvector centrality are more connected to other important nodes.

PageRank Centrality is similar to eigenvector centrality, which evaluates a node's connections, particularly those from highly ranked nodes. ACE-CSR-10, user-439 and user-426 have the highest PageRank centrality, implying that they receive the most attention from nodes that are themselves influential.

This analysis reveals that distinct nodes wield influence through diverse metrics, underscoring the network's complexity and the different roles nodes play. For instance, nodes like ACE-CSR-10 and user-241 are not only well-connected but also serve as bridges, indicating their pivotal role in information flow. Similarly, ACE-CSR-5 stands out not only for its high degree and out-degree centrality but also for its quick information dissemination potential due to its closeness centrality score.

Interestingly, despite the focus on analysing the ACE-CSR network on Twitter and the selection of their accounts as seeds for constructing the social graph, the dominance of ACEs-CSR accounts in terms of influence and centrality was not universally evident. Examination of the top 15 nodes across all centrality measures revealed a maximum of three ACE-CSR accounts in the list, with the remaining nodes being non-ACE-CSR accounts, which means that non-ACE-CSR nodes were central and influential in this researchers' network.

In conclusion, each centrality provides a unique lens through which we can view the network's structure and influence dynamics, and the results suggest that certain nodes, particularly a small number of ACE-CSR accounts and a specific non-ACE-CSR account (e.g., user-241), hold the most significant influence in various ways.

6.7.5.2 Examining Influence Distribution

Having computed centrality scores using diverse measures and pinpointed the most influential nodes, we aimed to explore how influence is dispersed within the ACEs-CSR network. Our objective was to examine whether influence is uniformly distributed across the network or if a particular set of accounts, particularly those focused on cyber security research, dominate with the highest centrality scores and emerge as the top influencers. This investigation into influence distribution was driven by the need to uncover the underlying dynamics of the network.

Understanding the distribution of influence is crucial for characterising the network's structure. For instance, scale-free networks typically exhibit a small number of highly influential nodes, indicating a hierarchical structure where a few nodes hold significant sway over the network. Conversely, a random network tends to have a more uniform distribution of influential nodes, suggesting a decentralised structure where influence is more evenly spread.

To this end, we harnessed the insights gained from our centrality analysis and sought to visualise the patterns within the network. By doing so, we aimed to determine if the ACEs-CSR network aligns more closely with a scale-free model, characterised by a few dominant nodes, or with a random network, characterised by equal distribution of influence.

Using the eigenvector centrality score, we trimmed the ACEs-CSR graph, showing only the top 5% nodes. The resulting graph is depicted in Figure 6.6. Nodes were grouped together based on their original clusters. The node label's colour was also assigned the same cluster's colour. The node size was scaled corresponding to its in-degree centrality score. A colour scale was used for node colours to reflect the eigenvector score (the darker the colour, the higher the score is). Node's shape can be a triangle or circle corresponding to the Individual attribute.

To visually assess the distribution of influence, we created a separate chart between accounts (i.e., nodes) and the scores for each centrality method: degree centrality (Figure 6.7), in-degree centrality (Figure 6.8), out-degree centrality (Fig-

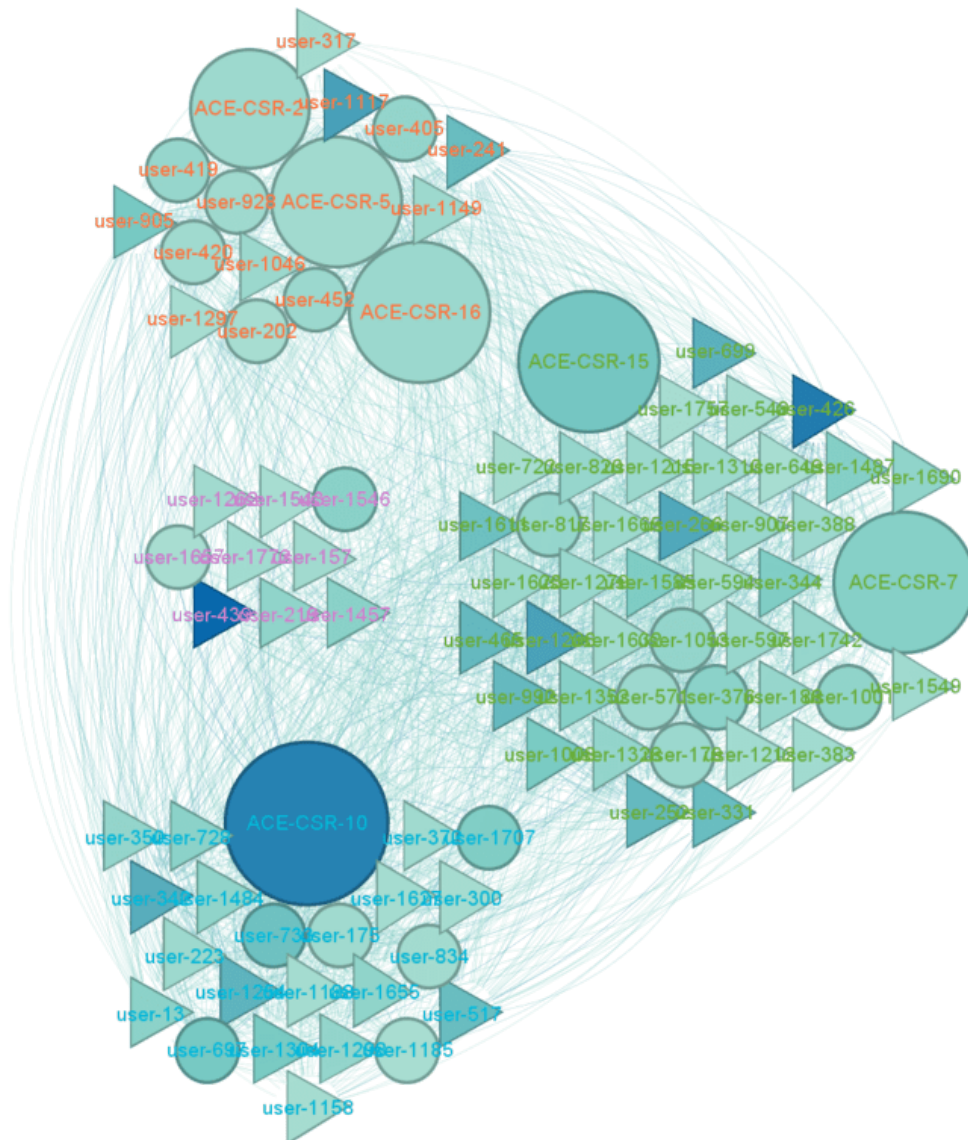


Figure 6.6: Top 5% nodes of the ACEs-CSR graph using eigenvector centrality

ure 6.9), betweenness centrality (Figure 6.10), eigenvector centrality (Figure 6.11), and PageRank centrality (Figure 6.12). We wanted to investigate whether each centrality distribution within the ACEs-CSR network follows a particular pattern.

Our analysis revealed that each centrality distribution exhibits characteristics of a long-tailed distribution, a hallmark feature of “scale-free” networks (Barabási and Albert, 1999). Several key observations emerge from these distributions as follows.

Connectivity: Within the ACEs-CSR network, a small subset of nodes exhibit

remarkably high centrality scores (i.e., “highly connected nodes”) and are often referred to as “hubs”, while the majority of nodes in the network possess relatively low centrality scores (i.e., low connectivity nodes), indicating only a few connections compared to the hubs. In scale-free networks, hubs serve as bridges between different sections of the network, facilitating efficient information dissemination. Thus, these hubs play a pivotal role in network connectivity.

Real-world relevance: Many real-world networks, such as social networks, the World Wide Web, and citation networks, manifest this type of centrality distribution (Barabási and Albert, 1999). For instance, in social networks, a few individuals (e.g., celebrities or influencers) may have an exceptionally high number of connections, while the majority of users maintain relatively few connections.

Resilience and vulnerability: Scale-free networks, typified by a long-tailed degree centrality distribution, exhibit resilience to random node failures but vulnerability to targeted attacks on hubs (Albert, Jeong, and Barabási, 2000). Removing a hub can significantly disrupt network functionality due to its high connectivity.

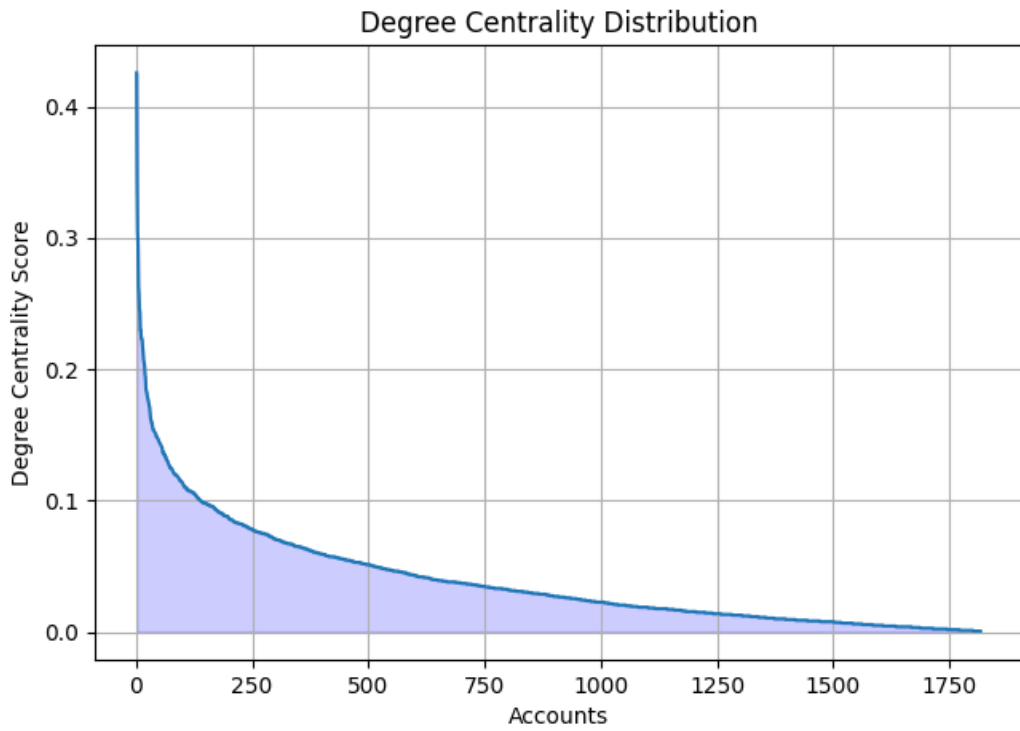


Figure 6.7: The degree centrality scores distribution, ACEs-CSR Network

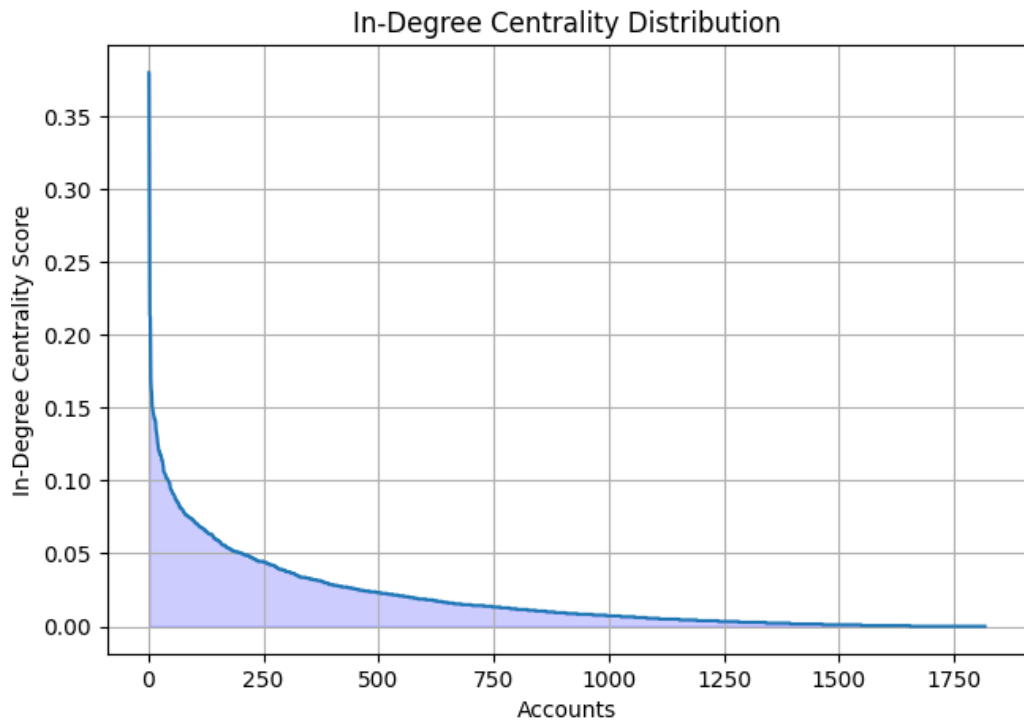


Figure 6.8: The in-degree centrality scores distribution, ACEs-CSR Network

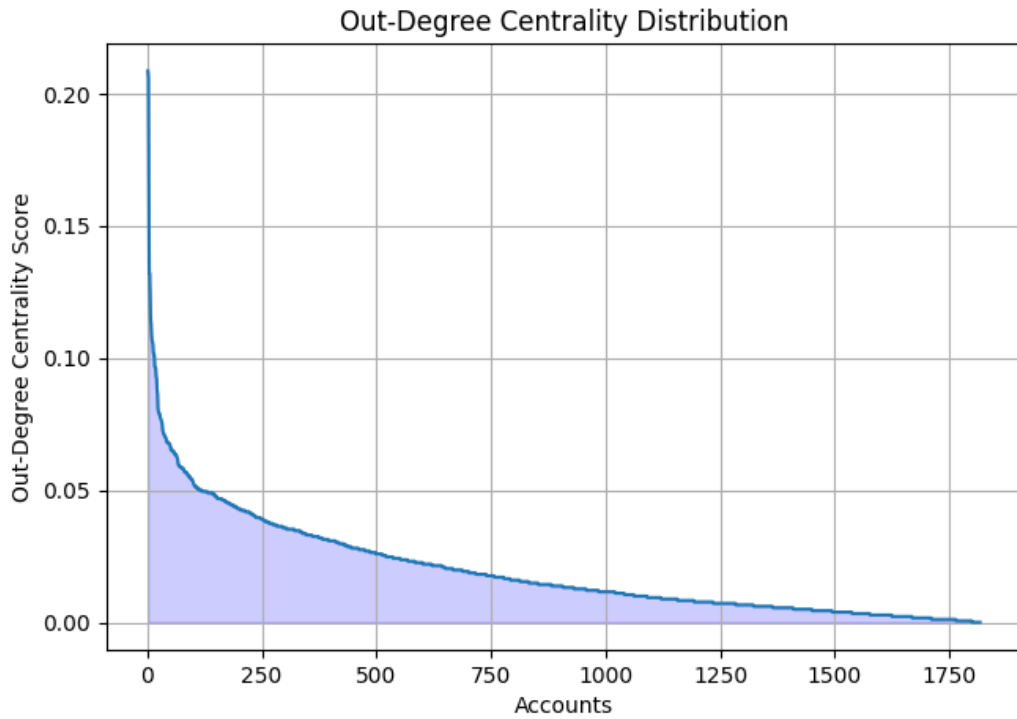


Figure 6.9: The out-degree centrality scores distribution, ACEs-CSR Network

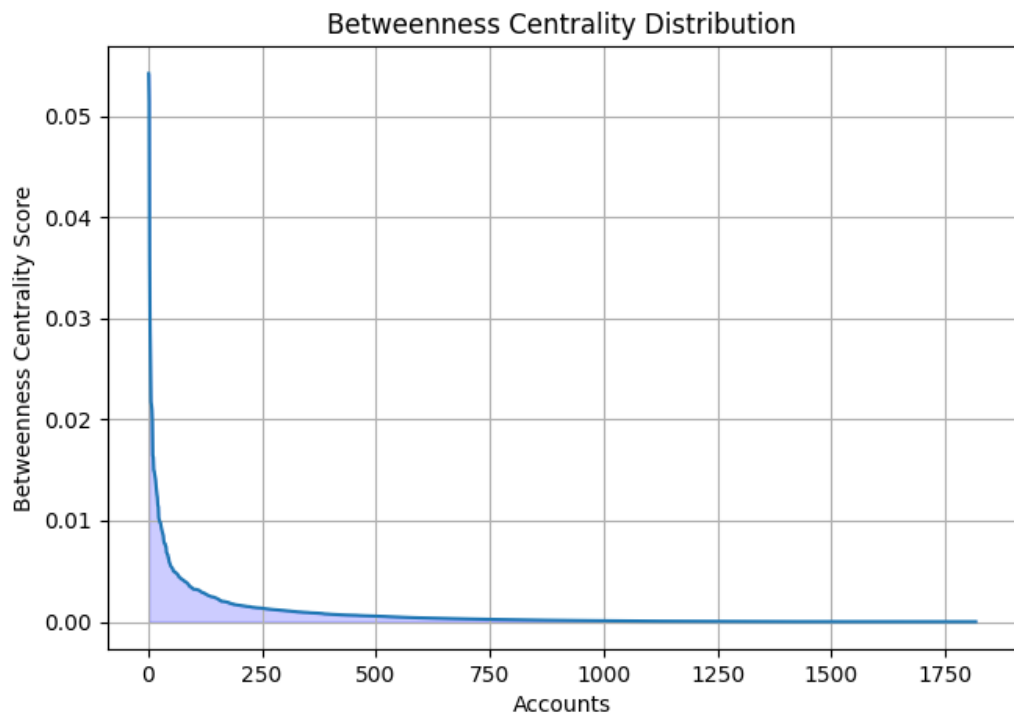


Figure 6.10: The betweenness centrality scores distribution, ACEs-CSR Network

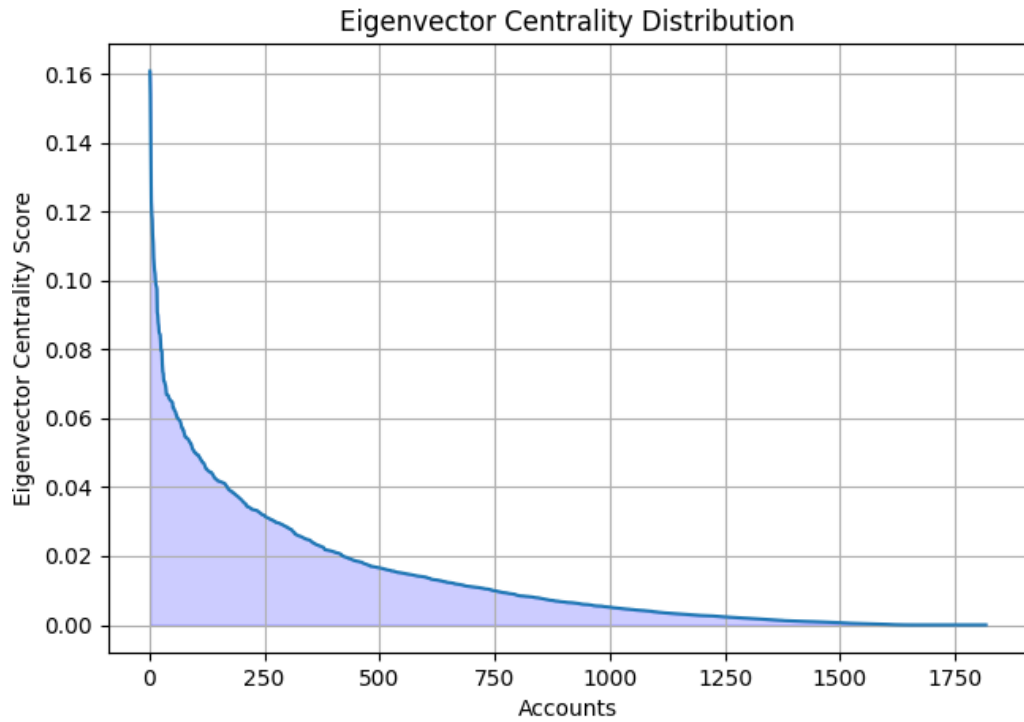


Figure 6.11: The eigenvector centrality scores distribution, ACEs-CSR Network

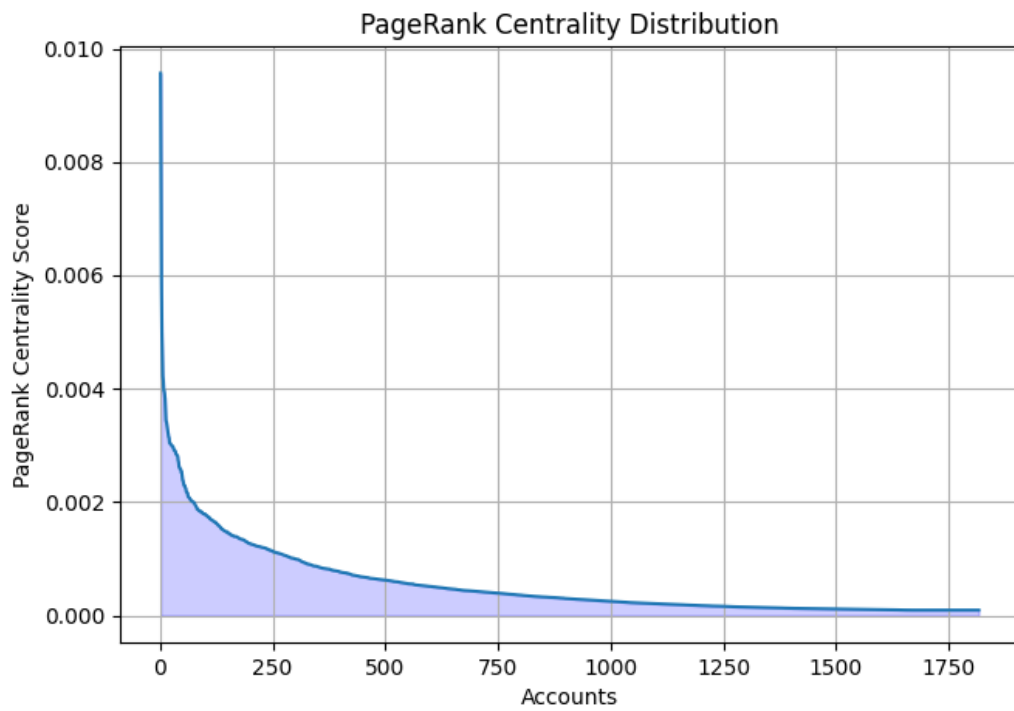


Figure 6.12: The PageRank centrality scores distribution, ACEs-CSR Network

6.8 Topic Modelling Analysis

Identifying and analysing the topics discussed by cyber security researchers on OSNs can reveal a lot of information such as their current focus and the trending technical and research subjects. A lot can be learned about cyber attacks and threats through the discussions of experts. Usually, researchers examine OSNs and forums of hackers and cyber criminals to learn more about their recent or previous nefarious activities. However, the technical details and how to prevent future attacks can be better found in the networks of cyber security experts.

Moreover, to learn more about a recent data breach, a new privacy law, or an industrial technology breakthrough in any of the cyber security sub-domains, you can start by analysing the discussions of experts and researchers, which are related to these topics. Also, insights from topic modelling analysis can inform decision-making processes within the cyber security research community, such as prioritising research areas, identifying knowledge gaps, or shaping research agendas.

RQ3.4 was explored through topical analysis, employing Topic Modelling Analysis to automatically identify the topics frequently discussed by cyber security and research-related accounts within the ACEs-CSR network on Twitter. Subsequent sections detail the timeline preparation process, the application of the LDA algorithm with appropriate parameters, and the presentation of extracted topics along with insights.

6.8.1 Documents Preparation

We used the LDA algorithm for our topic modelling analysis, refer to Section 2.5.4 for more details on LDA. We used the Scikit-Learn implementation for LDA (Sklearn LDA, 2022) to process the documents in our dataset and estimate the main topics. In our case, the documents were the Twitter timelines of the cyber security and research related accounts in the dataset, which were 1,817 accounts, but because some accounts' timelines were private and thus not publicly accessible, the actual number

of the timelines was 1,771. The document corresponding to each account was created by consolidating all the timeline's tweets into one text document. Then, the timelines went through a preprocessing phase, which included the following steps.

- Initial preprocessing was done by removing the Twitter handlers, URLs (including shortened links), email addresses and “RT” (retweet indicator).
- The text was then tokenised using the Gensim library (Řehůřek, 2022b).
- The tokens were further cleaned by removing punctuation, isolated numbers and short tokens (i.e., two characters or less).
- After that, stopwords were removed using a compiled list of NLTK stopwords, Gensim stopwords, and some manually added stopwords which contained some terms used on OSNs and are useless for topic modelling e.g., “yay”, “wow”, “oh”, “lol”, etc.
- Lemmatisation was then applied using the TextBlob library (Loria, 2022).

6.8.2 Applying LDA Algorithm

After the pre-processing, the tokens were passed to the LDA algorithm. In order to obtain good results in terms of the estimated main topics by the LDA algorithm, we needed to find the optimum values for the LDA parameters, i.e. k , the number of topics, and r , the maximum number of iterations (see Section 2.5.4 for more details). Therefore, we tried to find the optimum values for these parameters automatically by training the LDA model for a series of values for each parameter. In each training cycle, we used the coherence model for topic models from the Gensim Python library (Řehůřek, 2022a) to calculate the UCI coherence score of the created topics (Röder, Both, and Hinneburg, 2015) in the trained LDA model. Ultimately, we chose the best value of each parameter that corresponds to the highest coherence score across all training cycles.

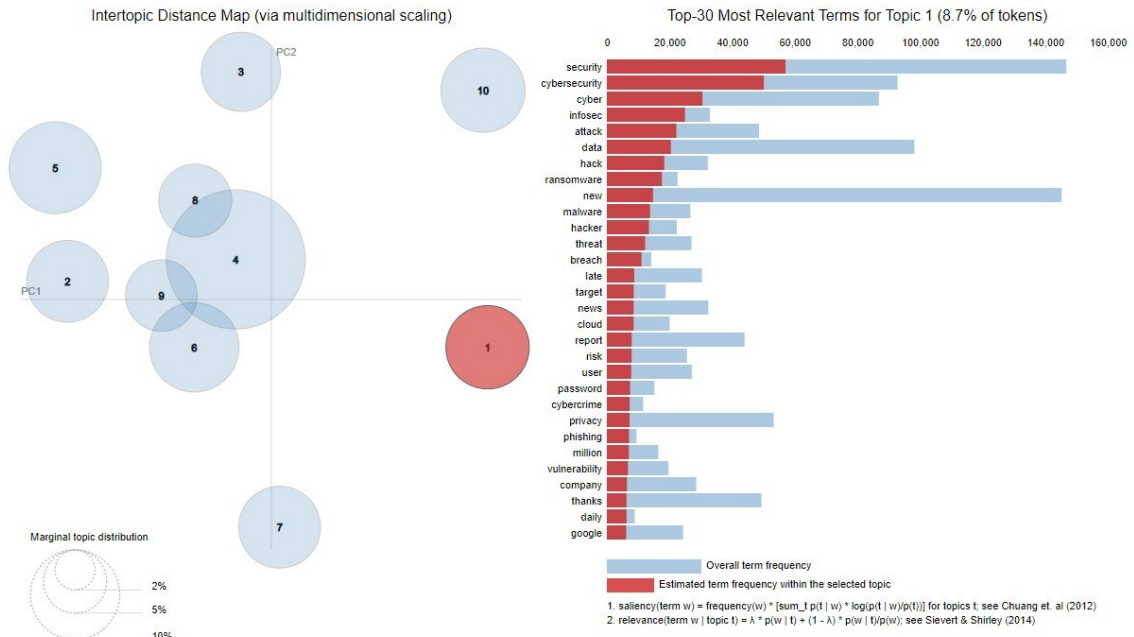


Figure 6.13: Visualisation of the estimated topics by the LDA algorithm

For k , the tested values were from 2 to 20 with a step size of 1 for each training cycle. The potential best values were 5 and 12. For r , the values were from 20 to 300 with a step size of 20. The potential best values were 200 and 220. While several past studies in the literature utilised coherence measures in similar experiments to find the best values for k (N. Pattnaik, Li, and Nurse, 2023; Jones, Nurse, and Li, 2020), several other studies agreed that a manual inspection approach for the estimated topics in each cycle is better for finding the best values of these parameters (Kigerl, 2018; Aslan, Li, et al., 2020), which was confirmed in our case as well.

Thus, we further inspected the topics generated in each training cycle aided by the topics visualisation generated by pyLDAvis Python library (Sievert and Shirley, 2014). The visualisation of pyLDAvis was quite beneficial for the topic interpretation as it provides a global view of the generated topics with the ability to focus on a certain topic at a time and see its top keywords. The most important aspect of pyLDAvis was the inter-topic distance map via multidimensional scaling, which shows topics overlapping. We found that the minimum overlapping between topics

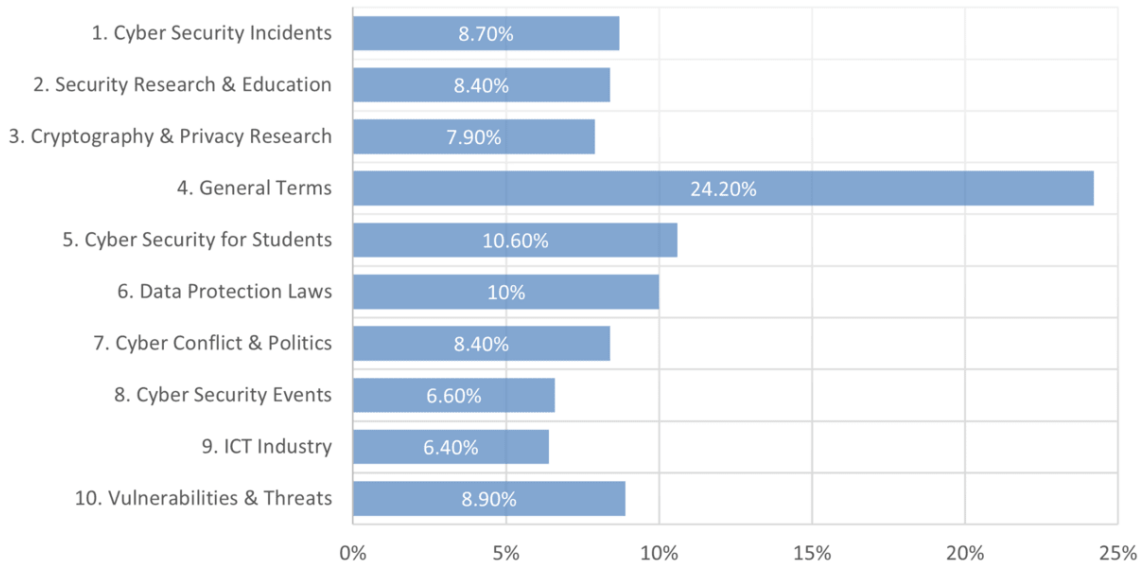


Figure 6.14: Size comparison of the estimated topics by the LDA algorithm

is better for topic interpretation. Considering the coherence model experiment, the manual inspection, and the visualisation-aid analysis, we chose $k = 10$ and $r = 200$.

6.8.3 Extracted Topics

The results in Figure 6.13 show the topics that were discussed by the cyber security research accounts in the ACEs-CSR network. In order to interpret the obtained topics, we had to label them manually by checking the top keywords in each one. Table 6.16 shows the labelled topics and the top 20 keywords in each one. Using the inter-topic distance map shown in Figure 6.13, we can notice that the correlation between topics is minimal. However, the main overlapping was caused mainly by topic T4, which is a generally mixed topic and non-related to cyber security domain. This kind of topic is expected to be found in similar textual sources like tweets.

The topic size comparison was presented in Figure 6.14. Apart from topic T4, all the other topics are relatively similar in size, ranging from 6.4% to 10.6% with an average of 8.4%. We can spot several topical themes by looking at the generated topics: research, privacy, education, technical, and politics. Ignoring T4, the top discussed topic was T5 (“Cyber Security for Students”, 10.6%), followed by T6 (“Data

Table 6.16: LDA topics with top 20 keywords, ranked in descending order by size

ID	Topic Name	Size (%)	Top Keywords
T4	General Terms	24.2	like, people, think, time, good, work, know, need, look, year, thing, day, great, want, way, thanks, come, love, right, try
T5	Cyber Security for Students	10.6	student, today, great, day, new, cyber, work, look, event, research, talk, join, team, uk, year, week, help, open, come, university
T6	Data Protection Laws	10	data, privacy, law, new, right, digital, eu, ai, internet, tech, work, protection, facebook, online, gdpr, public, surveillance, uk, need, report
T10	Vulnerabilities & Threats	8.9	new, security, malware, attack, tool, vulnerability, release, exploit, code, hack, blog, use, android, linux, update, file, bug, analysis, cve, learn
T1	Cyber Security Incidents	8.7	security, cybersecurity, cyber, infosec, attack, data, hack, ransomware, new, malware, hacker, threat, breach, late, target, news, cloud, report, risk, user
T2	Security Research & Education	8.4	research, new, work, security, social, read, join, look, digital, data, online, study, report, project, researcher, science, policy, paper, great, article
T7	Cyber Conflict & Politics	8.4	cyber, state, russia, new, russian, china, war, ukraine, government, attack, world, country, intelligence, military, report, trump, india, force, election, today
T3	Cryptography & Privacy Research	7.9	paper, security, work, research, new, privacy, talk, crypto, open, program, phd, bitcoin, student, computer, blockchain, present, attack, conference, year, post
T8	Cyber Security Events	6.6	cybersecurity, security, cyber, join, learn, new, register, ic, today, check, day, event, talk, team, course, help, conference, look, secure, great
T9	ICT Industry	6.4	ai, iot, technology, data, learn, new, business, tech, future, digital, market, innovation, report, industry, world, read, join, cloud, service, company

Protection Laws”, 10%), T10 (“Cyber Security Vulnerabilities & Threats”, 8.9%), T1 (“Cyber Security Incidents”, 8.7%), and T2 (“Security Research & Education”, 8.4%). Interestingly, politics-related and cyber conflict discussions in T7 also had a good share with 8.4%. Upon checking some tweets, we noticed sub-topics that many researchers discussed within politics, e.g., the Russia-Ukraine cyber conflict and the Trump elections. Finally, by checking the document-topic matrix, we found that the top two main topics across all documents are T5 and T3.

Overall, the topic modelling analysis revealed a diverse spectrum of discussions within this community, highlighting both conventional cyber security topics and emerging trends. These findings go beyond surface-level observations, offering deeper insights into the evolving discourse and interdisciplinary connections, such as with politics or industry, that shape discussions on cyber security issues. Additionally, the quantitative assessment of topic sizes and their order provides a nuanced view of the relative importance of different themes within the community over time.

6.9 Sentiment Analysis

For RQ3.4, sentiment analysis was utilised to know how the cyber security research community perceive the ACE-CSR programme and accounts on Twitter. The ACE-CSR programme started almost a decade ago, and such an analysis can provide useful insights about what to do in the future with the ACE-CSR programme.

6.9.1 Dataset Creation

We created a dataset of tweets by filtering the timelines of the 42,028 accounts in the dataset (i.e., Lv1 and Lv2 in the ACEs-CSR network), searching for tweets related to the ACE-CSR program or any of the ACEs-CSR using a set of selected keywords. Moreover, we added tweets that mentioned any of the 19 seed accounts we used during the data collection, as such mentions were considered direct or indirect interactions with an ACE-CSR. Finally, we excluded tweets created by the seed accounts as these accounts might be biased when they talk about the ACE-CSR program or themselves. In the end, a total of 21,374 tweets were obtained for the sentiment analysis. The tweets were pre-processed by removing Twitter handlers, URLs, email addresses, and the beginning word “RT” (for retweets).

6.9.2 Sentiment Analysis Algorithms

We examined the two most popular methods for sentiment analysis. The first one is the sentiment analyser in TextBlob (Loria, 2022), a popular Python library for text processing and NLP tasks standing on the giant shoulders of NLTK (NLTK Toolkit, 2023), which is a leading Python platform for working with human languages. TextBlob relies on a lexicon-based sentiment analyser with predefined rules to calculate a “polarity” score between -1 and 1. This score tells whether a text can be considered positive, neutral, or negative. The second method is VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based sentiment analyser with a simple rule-based model for general sentiment analysis (Hutto and Gilbert, 2014). The VADER sentiment analyser returns four scores for each piece of input text: “neg”, “neu”, “pos”, and “compound”. Each score corresponds to a sentiment type except the last, which is a normalised combined value of the first three scores. We used the VADER implementation in the NLTK library (NLTK Python, 2022).

Table 6.17: Examples of tweets wrongly classified by TextBlob sentiment analyser

Tweet	TextBlob	Vader
<i>Our Academic Centre of Excellence in Cyber Security Research becomes active this week.</i>	Negative	Positive
<i>Congratulations to @UniKent @KingsCollegeLon and @cardiffuni who join @UniofOxford and 13 other UK universities as Academic Centres of Excellence in Cyber Security Research, announced recently by the National Cyber Security Centre @NCSC and @EPSRC.</i>	Negative	Positive
<i>Academic Centre of Excellence in Cyber Security Research Open Day @ucl: @uclisec hosting an open day at the ACE center November 15th #infosec #CyberSecurity.</i>	Negative	Positive

We first applied the TextBlob, and upon checking the results, we noticed a lot of tweets in the dataset where TextBlob failed to classify them correctly and assigned the wrong sentiment types to them. Thus, we tried the Vader sentiment analyser and examined the same cases, and indeed Vader classified them correctly. For com-

parison, we listed in Table 6.17 a few examples of tweets that TextBlob wrongly classified. The results of the Vader sentiment analyser were improved significantly. After applying both sentiment analysers to our data and manually inspecting the results, we concluded that VADER was a better method.

6.9.3 Sentiment Analysis Results

The results of the VADER sentiment analyser were depicted in Figure 6.15. 65.8% of all tweets were classified as positive, 25.09% as neutral, and only 9.11% as negative. These results showed that the cyber security research community perceived the ACE-CSR program and accounts on Twitter largely positively.

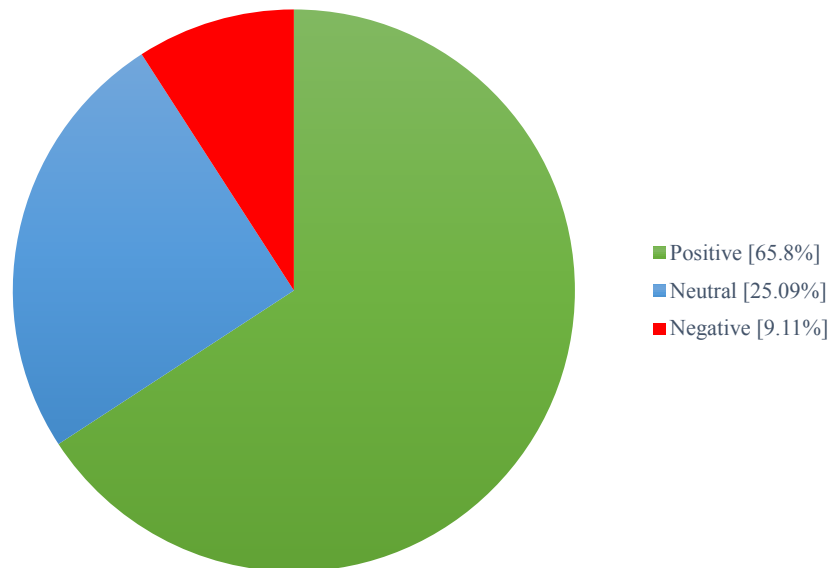


Figure 6.15: Sentiment Analysis, Statistics on Tweets Related to ACEs-CSR

Following the community analysis presented in Section 6.7.4, we wanted to know whether the sentiment analysis results would vary from one community to another, and between cyber security research related accounts and others in the ACEs-CSR network. Thus, we divided the tweets we selected into sub-datasets, each corresponding to an intended sub-group of accounts. Then, we analysed each sub-dataset using the VADER sentiment analyser. The results are listed in Table 6.18.

Table 6.18: Sentiment Analysis, Statistics on Tweets Related to ACEs-CSR

Accounts		Tweets	Positive		Neutral		Negative	
Main Group	Sub-group	Count	Count	(%)	Count	(%)	Count	(%)
Research related	C1	608	406	66.78	134	22.04	68	11.18
Research related	C2	1,613	988	61.25	459	28.46	166	10.29
Research related	C3	4,485	2,888	64.39	1,205	26.87	392	8.74
Research related	C4	753	476	63.21	188	24.97	89	11.82
Non Research related		13,915	9,306	66.88	3,377	24.27	1,232	8.85
All accounts		21,374	14,064	65.8	5,363	25.09	1,947	9.11

The sentiment analysis results of each sub-group are largely aligned with the main results for all. However, a few observations were noted, e.g., the percentage of the positive sentiment in Community C2 (the more “European” community) dropped to 61.25% while the negative percentage increased to 10.29%. On the other hand, the more UK-centric community C3 saw the lowest negative sentiment percentage (8.74%) across the four communities, while the positive sentiment percentage was 64.39%. Comparing the sentiment results of Communities C2 and C3, one may wonder if the accounts’ characteristics such as location can affect the results. One explanation for this observation is that UK-based accounts may be more interested in the ACE-CSR program than European-based accounts outside of the UK.

6.10 Conclusion

This chapter reported our study on the presence of cyber security experts on OSNs, focusing on the cyber security research accounts and taking the UK’s ACEs-CSR network on Twitter as a case study. In our analysis, we utilised the taxonomy created in Chapter 4 and the ML classifiers developed in Chapter 5 along with a new classifier developed in this chapter and other tools and techniques such as social structural analysis, influence analysis, topic modelling, and sentiment analysis.

The baseline and individual classifiers, developed in Chapter 5, were employed after re-validating their performance, and a third classifier was developed specifically to identify cyber security research accounts within the ACEs-CSR network. The results revealed the effectiveness of the reused classifiers, achieving 90% for F1-score using the RF model. Additionally, the Research classifier showcased its suitability in capturing the desired accounts of the case study, achieving an F1-score of 83% using the SVM-R model. The identified accounts, as cyber security related and research related, were used to construct the ACEs-CSR social graph on Twitter.

Then, we conducted a social structure analysis of the ACEs-CSR research network, influence analysis, topic modelling analysis, and sentiment analysis. The social structure analysis revealed some useful insights about the network's structure and sub-communities, e.g., a location-based analysis led to the discovery of a four-community structure: International, European, UK-centric, and balanced.

The conducted influence analysis within the ACEs-CSR network aimed to identify the nodes wielding the most substantial influence. The results illuminated an intriguing mix of influential nodes, comprising both the originally designated seed accounts and additional accounts within the network. Moreover, the examination of centrality distribution revealed interesting patterns reminiscent of a long-tailed distribution. This characteristic is highly significant, aligning with a key trait frequently observed in scale-free networks, implying a structure where a few nodes possess significantly higher centrality than others.

Topic modelling analysis revealed a wide range of topics which the research community of the ACEs-CSR network discussed on Twitter, e.g., cyber security incidents, system vulnerabilities, cyber threats, industry, data protection laws, and even politics and cyber conflicts. The sentiment analysis results showed that the accounts in the ACEs-CSR network talked about the ACE-CSR program and the ACEs-CSR mostly positively. Overall, our study has demonstrated the feasibility and usefulness of a largely automated data-driven approach for studying cyber security experts on OSNs and analysing their communities.

Chapter 7

Conclusion & Future Work

“The most important thing is to never stop questioning.”

- Albert Einstein

7.1 Conclusion

IN this thesis, we presented the findings of our comprehensive approach to investigate cyber security experts’ presence on OSNs, with a specific focus on cyber security research networks. As a case study, the UK ACEs-CSR network on Twitter was analysed. The main aim of this thesis was to design, develop and test the required tools to enable and support studying cyber security experts on OSNs. The thesis reported the work carried out to develop and test the required tools and methods to conduct the suggested analysis of cyber security experts.

We started developing the needed tools in Chapter 4 by creating a general cyber security taxonomy following a human-machine teaming based process, starting from a given set of textual sources and using mostly automated processing done by NLP and IR tools. The key features of the proposed taxonomy-building process are i) the ability to reduce human efforts needed due to the higher level of automation and ii) making the selection of the taxonomy concepts less subjective as it is data-driven and relies on the input sources.

Moving to Chapter 5, a new methodology was introduced to improve the automatic detection of cyber security related accounts and other sub-groups using supervised machine learning. The three-staged methodology consists of A) more systematic data-collection, B) crowdsourcing labelling experiment and C) the development of ML classifiers. The general cyber security taxonomy, created in Chapter 4, was crucial during the development of the ML classifiers in this chapter, as the terms within the taxonomy were used as features for the classifiers. The primary dataset was labelled using a crowdsourcing technique. A rich set of features was extracted and used to train the ML classifiers, experimenting with 63 feature sets and more than 1200 features. RF was the best model, achieving an F1-score ranging from 88% to 93%. Moreover, feature importance analysis was performed, discovering a minimal set of features enough to create a lightweight classifier with the same performance.

Ending with Chapter 6, where we utilised the tools and methods from Chapters 4 and 5 along with other techniques to analyse the case study of ACEs-CSR network on Twitter. Using 19 seed accounts representing the ACE-CSR accounts on Twitter, we extracted their network of friends and followers. Then, we employed the baseline and individual ML classifiers developed in Chapter 5 and developed a third classifier specifically designed to identify cyber security research accounts within the ACE-CSR network. The classifiers' results revealed the effectiveness of the reused classifiers, scoring 90% for F1-score using the RF model. The Research classifier also showcased its suitability in capturing the desired accounts, achieving an F1-score of 83% using the SVM-R model. The 1817 identified cyber security research accounts were used to construct the social graph corresponding to the research network of ACEs-CSR. Social structure analysis revealed valuable insights regarding the network's structure and the underlying sub-communities. By conducting a location-based analysis of community members, we identified four distinct types of communities within the ACE-CSR network: International, EU, UK, and a balanced mix cluster. Furthermore, the influence analysis was essential to iden-

tify the top influential nodes, which were a mix of ACE-CSR and non-ACE-CSR accounts. Centrality distribution showed an interesting characteristic of a scale-free network. Moreover, our topical analysis unveiled a diverse range of subjects discussed by cyber security researchers, including incidents, vulnerabilities, threats, industry trends, data protection laws, and even politics and cyber conflicts. Finally, to gain a deeper understanding of the sentiment within the ACEs-CSR network, we performed sentiment analysis on the communications taking place. The results demonstrated that accounts within the ACE-CSR network expressed positive sentiments towards the ACE-CSR program, indicating a favourable perception of the initiative. Furthermore, the analysis revealed that the overall communication between ACE-CSR and its network members was predominantly positive.

In summary, our study highlights the significant value of employing a data-driven approach alongside quantitative methods for analysing cyber security expert networks on OSNs, especially the research networks on Twitter. The findings provide meaningful insights into the composition, influence, discussions, and sentiment within the ACEs-CSR network, ultimately contributing to a more comprehensive understanding of cyber security experts' interaction and behaviour on OSNs.

7.2 Revisiting Research Questions

The human-machine teaming-based methodology for creating taxonomies and the created general cyber security taxonomy in Chapter 4 answered the first research question **RQ1** of this thesis. The enhanced methodology for automatic detection of cyber security related accounts, the labelling experiment, and the obtained results in Chapter 5 adequately answered **RQ2**. Finally, the utilisation of the tools developed in both Chapters 4 and 5, and the extensive analysis of the online communities of cyber security experts on OSNs using the presented case study addressed **RQ3** and the other sub-research questions set at the beginning of Chapter 6. Thus, the thesis addressed all the research questions and achieved its research aims.

7.3 Summary of Contributions

7.3.1 Data-Driven Human-Machine Teaming Approach for Constructing Taxonomies

We introduced a human-machine teaming approach for constructing taxonomies based on mixed textual documents. Our process involves a combination of manual and automated steps, utilising NLP and IR tools. By automating certain aspects, our approach was able to reduce human effort and introduced more data-driven decision-making, minimising subjectivity in the selection of the relevant terms. Moreover, this process facilitates easier maintenance of the constructed taxonomies.

7.3.2 Built General Cyber Security Taxonomy

To illustrate the effectiveness of our taxonomy-building approach, we provided an example of constructing a general taxonomy specifically tailored for the cyber security domain. This application of our approach demonstrates how the process efficiently combines human efforts with the automated collection of a large number of candidate terms from multiple data sources. This saves considerable time and mitigates the potential for errors that may arise when relying solely on human involvement. By employing the proposed process, taxonomies can be constructed and maintained more efficiently, ultimately enhancing the overall taxonomy-building experience.

7.3.3 Improved Methodology for Automatic Detection Of Cyber Security Related Accounts

In Chapter 5, we reported the improved methodology which we used to detect cyber security related accounts and other related sub-groups (individual accounts, hacker-related accounts, and accounts belonging to academia). Using a cyber security taxonomy, a systematically labelled dataset, and a crowdsourcing-based labelling experiment, a general three-staged methodology was proposed to ensure the dataset

is more representative and accurate. We trained and tested the four classifiers with five machine learning models using the labelled dataset and 63 different feature sets across five larger groups (with a total number of over 1,000 features). Furthermore, integrating the cyber security taxonomy into the feature sets of the machine learning classifiers contributed to the development of accurate and efficient automated systems for identifying and categorising cyber security related accounts on OSNs.

Our results showed that the Random Forest model is the best machine learning model for all four classifiers, with the best F1-score ranging from 88% to 93%. We also investigated the importance of different features and found that only a very small number of features (e.g., 6 for the baseline classifier) are already sufficient to produce a lightweight classifier with the same best performance.

7.3.4 Additional Classifiers to Detect Different Types of Cyber Security Accounts

In Chapter 5, we developed four classifiers, one for general accounts related to cyber security, and three others for typical sub-groups: individuals, those related to hacking, and those belonging to academia. The best machine learning model, Random Forest, performed well for all classifiers: 93% for the baseline classifier and 88-91% for the three sub-classifiers. Furthermore, in Chapter 6, we developed a new machine learning classifier to detect cyber security research related accounts with good performance.

7.3.5 Tested the ML Classifiers in a Real-World Setting

In the context of our research in Chapter 6, we performed a detailed evaluation of the machine learning classifiers that were developed in Chapter 5 and documented in (Mahaini and Li, 2021). The objective was to assess their effectiveness in identifying cyber security related accounts and sub-groups in a real-world setting. Through the results of the verification procedure, we gained confidence in the classifiers' per-

formance. As a result, we asserted that both the Baseline and Individual classifiers are sufficiently good and reliable for the ACEs-CSR analysis case study. This finding underscores the utility and applicability of these classifiers in real-world scenarios involving cyber security experts' analysis.

7.3.6 ACEs-CSR Research Network Analysis

Drawing upon the constructed ACEs-CSR network, our research encompassed a comprehensive analysis of various aspects of this network. This involved conducting social structure analysis, influence analysis, topic modelling analyses, and sentiment analyses to gain valuable insights. The social structure analysis shed light on the network's overall structure and revealed the existence of distinct sub-communities. Notably, a location-based analysis unravelled a four-community structure consisting of International, European, UK-centric, and balanced communities.

The influence analysis conducted within the network served the purpose of pinpointing the most influential nodes, revealing that these top influencers are a blend of both the seed accounts and other accounts within the network. Furthermore, the examination of centrality distribution unveiled patterns resembling a long-tailed distribution, a crucial hallmark commonly associated with scale-free networks.

Moving on to the topic modelling analysis, a diverse range of topics emerged from the discussions among cyber security researchers within the ACEs-CSR network on Twitter. These topics encompassed critical areas such as cyber security incidents, system vulnerabilities, cyber threats, industry-related matters, data protection laws, and even intersections with politics and cyber conflicts. Furthermore, sentiment analysis provided valuable insights into the overall sentiment expressed within the ACEs-CSR network. The results indicated that discussions surrounding the ACE-CSR programme and the ACEs-CSR were predominantly positive in nature. This finding highlights the favourable reception and positive sentiments towards this programme within the cyber security research community. This can help the decision-makers at NCSC to plan the next steps for the ACE-CSR programme.

7.3.7 Approach to Study Cyber Security Expert Networks

The presented study exemplifies the feasibility and usefulness of leveraging automated data-driven approaches to analyse and gain insights from cyber security expert networks on OSNs. By employing a combination of social structure analysis, topic modelling, influence analysis, and sentiment analysis, we have successfully uncovered valuable information about the network's structure, discussed topics, top influential, influence distribution, and prevailing sentiments. Such an approach shows great potential for researchers aiming to investigate further the dynamics of expert communities in the field of cyber security on OSNs. The methodologies and practical part of the work presented in this thesis were applied to Twitter as an example, but it can be applied to other OSNs that cyber security experts use.

7.4 Limitations & Future Work

Throughout this dissertation, we presented three main research questions, which correspond to Chapters 4, 5 and 6 respectively, and each main research question had its own related sub-research questions, challenges, solutions and limitations. We solved and presented what we have done during our research journey, but we highlighted the possible future work out of these identified limitations. The proposed approach for studying the networks of cyber security experts on OSNs should be expanded by further investigation. This section lists the potential research opportunities presented as future work for each of the main chapters.

7.4.1 Extend the Taxonomy-Building Approach

The work presented in Chapter 4 offers opportunities for various extensions. First, the tree-based taxonomy is not enough to capture complicated concepts and their relations. Also, the same issue is faced with taxonomies when a concept can fit into two classes at the same time. This is a known limitation of taxonomies. Thus, extending the taxonomy to a more complicated cyber security ontology will be needed

to support more advanced analysis, such as automatic reasoning based on input data (e.g., when a particular cyber security event happens, what consequences it will generate and what defences can be taken).

Second, we need to enhance the level of automation of the proposed process to further reduce the required human effort. For instance, more advanced NLP and IR tools can be used to reduce the number of irrelevant and wrongly extracted terms, and an automated recommendation system can map relevant terms to the defined taxonomy structure. It is also possible to automate the collection of input textual documents at a larger scale. Moreover, the process of building taxonomies can be significantly enhanced through the integration of AI and Large Language Models (LLMs). AI technologies excel in automating data processing tasks and efficiently handling large volumes of unstructured data to identify relevant terms and relationships for taxonomy construction. LLMs, powered by advanced NLP capabilities, play a pivotal role in understanding the context and semantics of terms. They can suggest accurate categorisations and relationships between terms, thereby enhancing the depth and accuracy of taxonomies.

Third, we want to consider using structural features to improve the n -grams ranking and terms selection. We will investigate the use of words' styles and positions to set weights for the extracted n -grams. For example, an n -gram that appears in a document title, abstract section, or section header should receive more importance than other n -grams in other parts of the document. Also, a word highlighted by a **bold** or *italic* style may be more important than other un-styled words.

Finally, with the aim of validating the generalisability and consistency of the proposed taxonomy-building method, the process can be conducted by independent experts and data sources, and then the results can be cross-validated and enhanced. We can also compare the taxonomy built using the proposed method with others built manually by experts without using a semi-automated approach. However, we would like to point out that a real ground truth can hardly be established for quality checking since the taxonomy is *qualitative* and *subjective* by nature.

7.4.2 Enhance the Automatic Detection of Cyber Security Related Accounts and other Sub-groups

There are some limitations in the reported work in Chapter 5 that can be addressed in future work. Using a list of cyber security related n -grams – derived from the cyber security taxonomy in Chapter 4 – was necessary to filter the huge data collection we harvested from Twitter. However, while being representative, this list is far from complete, especially for such fast-evolving domains as cyber security, where new concepts and terms keep appearing. We plan to help refine the cyber security taxonomy to make it more dynamically updated and then use it to update the developed classifiers.

The recruited participants in the labelling experiment were not as diverse as we wished, considering different demographic factors such as gender, age, employment status, and level of education, see Figure 5.5. For example, only about 25% of the participants were females. We plan to conduct future experiments to recruit more participants to enlarge our dataset with a more diverse participation pool. However, it is pretty challenging to address this easily, as it is common to have an unbalanced distribution for the participants across the mentioned demographics, especially when the targeted people are from a specific domain.

Furthermore, the performance of the Research machine learning classifier used in Chapter 6 was 83% for the F1-score. Although this is a good score, it is lower than the scores achieved in other related work (Aslan, Sağlam, and Li, 2018), including our previous results in Chapter 5, which were over 90%. However, this classification task was more challenging due to its cascaded nature, involving the detection of research accounts which is a subset within the broader realm of cyber security related accounts. Enhancing the Research classifier can be accomplished by considering additional features and employing a larger dataset, thereby enabling the use of other hybrid machine learning models such as deep learning.

We also plan to improve the classifiers so they can stand the test of time. The keyword-based features were extracted based on the content found in the Twitter

accounts' timelines that are in the training set, which means that the performance of any classifier that uses keyword-based features could decrease over time as the topics discussed by cyber security people change over time.

7.4.3 Further applications for the developed ML classifiers

As mentioned earlier in the introduction of Chapter 5.1, the developed classifiers can be used to analyse cyber security related accounts on OSNs for many different purposes. One example application for OSINT (open source intelligence) is to use the Hacker sub-classifier to help identify more hacker-related accounts so we can study more hacker-related phenomena beyond what has been reported (Jones, Nurse, and Li, 2020; Aslan, Li, et al., 2020). A second example is to automatically detect and monitor different types of cyber security related accounts to collect cyber threat intelligence. A third example is using the developed machine learning classifiers to study cyber security influencers on Twitter. Having multiple classifiers will allow us to detect influencers and influencees belonging to different sub-communities of the cyber security community, therefore allowing us to gain more insights about how such influence is formed and spread across an OSN platform and members of different (sub-)communities.

7.4.4 Open Opportunities for Studying Cyber Security Communities on OSNs

The study presented in Chapter 6 can be expanded upon in various ways to address the limitations we encountered and to develop and build on top of what we have learned through the introduced methodology and the case study of the ACEs-CSR network on Twitter. First and most importantly, study other types of cyber security experts and their networks, such as cyber security practitioners, innovators, vendors, end-users, etc. Second, it would be interesting to explore other networks of cyber security researchers. While the presented case study focused on the ACEs-CSR on

Twitter provided valuable insights, they may not be universally applicable to all other cyber security researchers' networks on Twitter or other OSNs.

Third, to gain a better understanding of cyber security research users on OSNs, it is necessary to expand the dataset used for such studies. Thus, a larger dataset is crucial for improved comprehension of this specific user group and their online communities. Fourth, it is of paramount importance to explore other social networks and websites such as LinkedIn and cyber security research groups on universities' websites. The cyber security researchers are usually listed on those websites and can be found on LinkedIn. By incorporating this additional data, we can further investigate research groups and gain deeper insights. Fifth, leveraging external data sources like Google Scholar can help us identify correlations between positive or negative sentiment and significant events in a researcher's career, such as publishing a paper or journal, delivering a talk, participating in international conferences or workshops, and more.

Finally, using an external data source focused on research and publications, we may be able to confirm whether a researcher's main topic aligns with their research interests as indicated by their list of publications. Additionally, we wish to explore whether researchers with specific research topics tend to connect with each other. Lastly, although our research primarily focused on cyber security researchers, we also intend to investigate research networks in other domains and draw comparisons for a more comprehensive understanding.

7.4.5 Extend the Work to Other Languages

The work presented in this thesis has primarily focused on the English language. The collected sources and extracted terms utilised in constructing the general cyber security taxonomy were written in English. However, it is worth mentioning that numerous cyber security accounts on OSNs use other languages. Consequently, to effectively study and analyse these accounts and their activities conducted in different languages, it becomes necessary to incorporate support for multiple languages

across all the different aspects of the presented work.

The methodology for building the taxonomy, as presented in Chapter 4, can be extended and generalised to encompass support for various languages. Thus, researchers can adapt the machine learning classifiers to accommodate the unique linguistic characteristics and nuances of different languages. This expansion would enable a more comprehensive understanding of cyber security networks and activities taking place in languages beyond English. Incorporating multiple languages in the classifiers will enhance the overall accuracy and applicability of the research, providing a broader perspective on cyber security communities on OSNs.

Moving forward, future studies should incorporate multilingual capabilities in analysing cyber security expert networks on social media. This would enable a more inclusive study of these networks across different languages, providing deeper insights into the global cyber security landscape and enhancing our understanding of cyber security on OSNs.

Appendix A

General Cyber Security Taxonomy Webpage

A dedicated webpage for the general cyber security taxonomy was created under the University of Kent domain, where we actively maintain the taxonomy, ensuring its accessibility and usability. Additionally, we offer both machine-readable files and interactive visualisations for the most stable iteration of the taxonomy, enabling users to readily download and utilise it. We attached the current version of the cyber security taxonomy to the thesis document, see Table A.1 to save the attached taxonomy files. However, for the latest version, please visit the taxonomy webpage¹.

Cyber Security Taxonomy

We have done a lot of work on since we published the paper in August 2019, and the current taxonomy contains more than 1900 nodes.

Visualization

For an interactive visualization of the taxonomy, please click [here](#).

We used the following libraries to create our visualizations: [JavaScript InfoVis Toolkit](#) and [D3](#)



Figure A.1: Taxonomy Visualisation Webpage

¹https://cyber.kent.ac.uk/research/cyber_taxonomy/

We extend an open invitation to fellow cyber security researchers and experts to collaborate in further advancing the taxonomy. Originally published in 2019, the taxonomy continues to be utilised by individuals and organisations, and we continue to receive valuable feedback and collaboration requests.

Table A.1: Cyber Security Taxonomy Files

Format	Size	Attached File
CSV	145 KB	Cybersecurity_Taxonomy.csv
SQL	307 KB	Cybersecurity_Taxonomy.sql
JSON	527 KB	Cybersecurity_Taxonomy.json
XML	316 KB	Cybersecurity_Taxonomy.xml
XLSX	97 KB	Cybersecurity_Taxonomy.xlsx

Appendix B

Labelling Experiment

Questionnaire

Below is the questionnaire that we used along with the labelling experiment. The questionnaire is divided into three parts as follows.

B.1 Part 1: Basic Demographics

The survey begins with a few questions about the basic demographics of the participants. In this section, we have five questions in total.

Q1. Gender

What is your gender?

- Male
- Female
- Other

Q2. Age Bracket

What is your age?

- 18-24
- 25-34
- 35-44

- 45-54
- 55-64
- 65-74
- 75 and above

Q3. Employment

What is your current employment status?

- Employed
- Self-employed
- Student
- Retired
- Unemployed

Q4. Education

What is your highest qualification? (even if you are still studying it)

- Doctorate degree
- Master's degree
- Bachelor's degree
- High school diploma or equivalent
- Other

Q5. Cyber Security Experience

Which of the following best describes you in the context of cyber security?

- I study or have a degree in cyber security
- I teach cyber security
- I research cyber security
- I work in the cyber security industry
- I work on cyber security for the government or a not-for-profit organisation
- I have hands-on skills that relate to cyber security
- I am not a cyber security professional. I am just interested in the latest updates on cyber security

- Other, please specify: [Text Answer]

B.2 Part 2: Cyber Security Experts on OSNs

There are a few questions about the cyber security related topics, communications and terminology used by cyber security professionals.

Q6. Definition of Cyber Security Expert

In your opinion, a cyber security expert is someone who:

- has a degree in cyber security
- teaches cyber security at a university
- researches cyber security at a research centre or university
- works in cyber security at a company or organisation
- has hands-on skills that relate to cyber security
- is a hacker (White/Grey/Black-Hat)
- Other, please specify: [Text Answer]

Q7. News Sources

What information sources do you use to stay updated about cyber security?

- Social media
- Conventional media (newspapers, radio, TV, ...)
- Official reports (issued by governments, companies, research centres, ...)
- Mailing lists
- Other, please specify: [Text Answer]

Q8. Keeping Up to Date

What forms of social media do you use to follow the latest updates and news about cyber security topics, incidents and attacks?

- General Online Social Networks (e.g. Facebook)
- Professional Social Networks (e.g. LinkedIn)
- Instant Messaging (e.g. WhatsApp)

- Microblogging (e.g. Twitter)
- Weblogs (e.g. Medium)
- Forums (e.g. Reddit)
- Video-sharing services (e.g. YouTube)
- Other, please specify: [Text Answer]

Q9. Activity on Online Social Networks (OSNs)

How often do you use OSNs to share or discuss cyber security related topics?

- Very often (at least once per day)
- Often (a few times per week)
- Sometimes (a few times per month)
- Rarely (at least once per month)
- Never

Q10. Your Audience

If you use OSNs, who is the intended audience of your posts about cyber security?

- Cyber security related professionals/colleagues
- Non-Cyber security related people
- Anyone
- Not sure
- Other, please specify: [Text Answer]
- N/A

Q11. Your Terminology

Do you use a different terminology when writing posts about cyber security related topics compared to your other non-cyber security writings?

- Yes, please describe that if possible: [Text Answer]
- Somehow
- No
- Don't know
- N/A

Q12. Use of Concepts vs Instances

When you read or write a post about a cyber security attack or incident, do you prefer the use of concepts, like the attack category (e.g. Ransomware), the actual attack name (e.g. WannaCry01) or both? Here are three statements about the same attack, please choose one.

- The ransomware attack on UK NHS systems in 2017 was massive
- The WannaCry01 attack on UK NHS systems in 2017 was massive
- The ransomware/WannaCry01 attack on UK NHS systems in 2017 was massive

Q13. Cyber Security Keywords

Please write down general keywords (not specific ones on sub-topics) that you often use when writing about cyber security on OSNs: [Text Answer]

B.3 Part 3: Labelling Process Feedback

We would like to ask you a few questions about your experience during the labelling process.

Q14. Twitter Profiles

On a scale from 1 to 5 (where 5 is the highest), how much were the Twitter profiles helpful for you during the labelling process?

Q15. Twitter Timelines

On a scale from 1 to 5 (where 5 is the highest), how much were the Twitter timelines helpful for you during the labelling process?

Q16. Google Search Results

On a scale from 1 to 5 (where 5 is the highest), how much were the Google search results helpful for you during the labelling process?

Q17. Other Factors

What other factors did you also consider while doing the labelling tasks?

- Account screen-name (username)
- Account name
- Tweets count
- Followers count
- Following count
- Favourites count
- Listed count
- Profile image
- Verified sign (the blue mark for a verified account)
- Other, please specify: [Text Answer]

Q18. Your Strategy

Did you use any strategy to help yourself finish the labelling tasks faster?

- If so, please describe the strategy you used: [Text Answer]

Bibliography

- Abulaish, Muhammad and Mohd Fazil (2020). “Socialbots: Impacts, Threat-Dimensions, and Defense Challenges”. In: *IEEE Technology and Society Magazine* 39.3, pp. 52–61. DOI: [10.1109/MTS.2020.3012327](https://doi.org/10.1109/MTS.2020.3012327).
- Adewole, Kayode Sakariyah, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak (2016). “Malicious accounts: Dark of the social networks”. In: *Journal of Network and Computer Applications* 79. September 2016, pp. 41–67. DOI: [10.1016/j.jnca.2016.11.030](https://doi.org/10.1016/j.jnca.2016.11.030).
- Agrafiotis, Ioannis, Jason R. C. Nurse, Michael Goldsmith, Sadie Creese, and David Upton (2018). “A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate”. In: *Journal of Cybersecurity* 4.1, ty006. DOI: [10.1093/cybsec/tyy006](https://doi.org/10.1093/cybsec/tyy006).
- Ahmad, Khurshid, Lee Gillam, Lena Tostevin, and Ai Group (2000). “University of Surrey Participation in TREC 8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)”. In: *The Text Retrieval Conference (TREC)*. NIST. URL: <https://trec.nist.gov/pubs/trec8/papers/surrey2.ps>.
- Albert, Réka, Hawoong Jeong, and Albert-László Barabási (2000). “Error and attack tolerance of complex networks”. In: *Nature* 406.6794, pp. 378–382. DOI: [10.1038/35019019](https://doi.org/10.1038/35019019).
- Andreotta, Matthew, Robertus Nugroho, Mark J. Hurlstone, Fabio Boschetti, Simon Farrell, Iain Walker, and Cecile Paris (2019). “Analyzing social media data: A mixed-methods framework combining computational and qualitative text analy-

- sis”. In: *Behavior Research Methods* 51, pp. 1766–1781. DOI: [10.3758/s13428-019-01202-8](https://doi.org/10.3758/s13428-019-01202-8).
- Antonakaki, Despoina, Paraskevi Fragopoulou, and Sotiris Ioannidis (2021). “A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks”. In: *Expert Systems with Applications* 164, p. 114006. DOI: [10.1016/j.eswa.2020.114006](https://doi.org/10.1016/j.eswa.2020.114006).
- Aslan, Çağrı B., Shujun Li, Fatih V. Celebi, and Hao Tian (2020). “The World of Defacers: Looking through the Lens of Their Activities on Twitter”. In: *IEEE Access* 8, pp. 204132–204143. DOI: [10.1109/ACCESS.2020.3037015](https://doi.org/10.1109/ACCESS.2020.3037015).
- Aslan, Çağrı B., Rahime Belen Sağlam, and Shujun Li (2018). “Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter as an Example”. In: *Proceedings of the 9th International Conference on Social Media and Society*. ACM, pp. 236–240. DOI: [10.1145/3217804.3217919](https://doi.org/10.1145/3217804.3217919).
- Aswani, Reema, Arpan Kumar Kar, and P. Vigneswara Ilavarasan (2018). “Detection of Spammers in Twitter marketing: A Hybrid Approach Using Social Media Analytics and Bio Inspired Computing”. In: *Information Systems Frontiers* 20.3, pp. 515–530. DOI: [10.1007/s10796-017-9805-8](https://doi.org/10.1007/s10796-017-9805-8).
- Bagozzi, Richard P and Utpal M Dholakia (2006). “Open Source Software User Communities: A Study of Participation in Linux User Groups”. In: *Management Science* 52.7, pp. 1099–1115. DOI: [10.1287/mnsc.1060.0545](https://doi.org/10.1287/mnsc.1060.0545).
- Bailey, Kenneth D. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques*. Quantitative Applications in the Social Sciences 102. SAGE Publications. ISBN: 9780803952591. URL: <https://books.google.co.uk/books?id=1TaYulGjhLYC>.
- Bakshy, Eytan, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts (2011). “Everyone’s an Influencer: Quantifying Influence on Twitter”. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM ’11. Association for Computing Machinery, pp. 65–74. DOI: [10.1145/1935826.1935845](https://doi.org/10.1145/1935826.1935845).

- Barabási, Albert-László and Réka Albert (1999). “Emergence of Scaling in Random Networks”. In: *Science* 286.5439, pp. 509–512. DOI: [10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509).
- Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). “Gephi: An Open Source Software for Exploring and Manipulating Networks”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 3.1. DOI: [10.1609/icwsm.v3i1.13937](https://doi.org/10.1609/icwsm.v3i1.13937).
- Basu, Aparna (2014). “Social Network Analysis: A Methodology for Studying Terrorism”. In: *Social Networking: Mining, Visualization, and Security*. Ed. by Mrutyunjaya Panda, Satchidananda Dehuri, and Gi-Nam Wang. Vol. 65. Springer, pp. 215–242. DOI: [10.1007/978-3-319-05164-2_9](https://doi.org/10.1007/978-3-319-05164-2_9).
- Bedell, Doug (1998). “Meeting your new best friends Six Degrees widens your contacts in exchange for sampling Web sites”. In: *The Dallas Morning News* 4. URL: <https://web.archive.org/web/20010104125400/http://www.dougbedell.com/sixdegrees1.html>.
- Beekun, Rafik I. (1989). “Assessing the Effectiveness of Sociotechnical Interventions: Antidote or Fad?” In: *Human Relations* 42.10, pp. 877–897. DOI: [10.1177/001872678904201002](https://doi.org/10.1177/001872678904201002).
- Benevenuto, Fabricio, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida (2010). “Detecting Spammers on Twitter”. In: *Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*. CEAS 2010. URL: <https://homepages.dcc.ufmg.br/~fabricio/download/ceas10.pdf>.
- Bhol, Seema Gupta, JR Mohanty, and Prasant Kumar Pattnaik (2021). “Taxonomy of cyber security metrics to measure strength of cyber security”. In: *Materials Today: Proceedings* 80, pp. 2274–2279. DOI: [10.1016/j.matpr.2021.06.228](https://doi.org/10.1016/j.matpr.2021.06.228).
- Bilge, Leyla, Thorsten Strufe, Davide Balzarotti, and Engin Kirda (2009). “All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks”. In: *Proceedings of the 18th International Conference on World*

- Wide Web*. WWW '09. Association for Computing Machinery, pp. 551–560. DOI: [10.1145/1526709.1526784](https://doi.org/10.1145/1526709.1526784).
- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer Science+Business Media. ISBN: 0387310738.
- Blei, David M. (2012). “Probabilistic Topic Models”. In: *Communications of the ACM* 55.4, pp. 77–84. DOI: [10.1145/2133806.2133826](https://doi.org/10.1145/2133806.2133826).
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: <https://www.jmlr.org/papers/v3/blei03a.html>.
- Bloch, Francis, Matthew O. Jackson, and Pietro Tebaldi (2023). “Centrality measures in networks”. In: *Social Choice and Welfare* 61.2, pp. 413–453. DOI: [10.1007/s00355-023-01456-4](https://doi.org/10.1007/s00355-023-01456-4).
- Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008). “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008:1–P10008:12. DOI: [10.1088/1742-5468/2008/10/p10008](https://doi.org/10.1088/1742-5468/2008/10/p10008).
- Borgatti, Stephen P., Martin G. Everett, and Jeffrey C. Johnson (2018). *Analyzing Social Networks*. Core Textbook. Sage. ISBN: 9781526418487. URL: <https://books.google.co.uk/books?id=-gpEDwAAQBAJ>.
- Bostock, Mike (2022). *d3-hierarchy: 2D layout algorithms for visualizing hierarchical data*. URL: <https://github.com/d3/d3-hierarchy>.
- Boyd, Danah M. and Nicole B. Ellison (Oct. 2007). “Social Network Sites: Definition, History, and Scholarship”. In: *Journal of Computer-Mediated Communication* 13.1, pp. 210–230. DOI: [10.1111/j.1083-6101.2007.00393.x](https://doi.org/10.1111/j.1083-6101.2007.00393.x).
- Boyd, Danah M., Scott Golder, and Gilad Lotan (2010). “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter”. In: *2010 43rd Hawaii International Conference on System Sciences*. IEEE Computer Society, pp. 1–10. DOI: [10.1109/HICSS.2010.412](https://doi.org/10.1109/HICSS.2010.412).

- Bradley, David (2008). “Six Degrees of Separation”. In: *Science*. URL: <https://www.sciencebase.com/science-blog/six-degees-of-separation.html>.
- Bradshaw, Samantha and Philip N. Howard (2019). *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*. Tech. rep. Oxford Internet Institute. URL: <https://digitalcommons.unl.edu/scholcom/207>.
- Breiman, Leo (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- Brin, Sergey and Lawrence Page (1998). “The anatomy of a large-scale hypertextual Web search engine”. In: *Computer Networks and ISDN Systems* 30.1. Proceedings of the Seventh International World Wide Web Conference, pp. 107–117. DOI: [10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X).
- Brownlee, Jason (2019). “14 Different Types of Learning in Machine Learning”. In: *MachineLearningMastery.com*. URL: <https://machinelearningmastery.com/types-of-learning-in-machine-learning>.
- Burger, Eric W., Michael D. Goodman, Panos Kampanakis, and Kevin A. Zhu (2014). “Taxonomy Model for Cyber Threat Intelligence Information Exchange Technologies”. In: *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security*. ACM, pp. 51–60. DOI: [10.1145/2663876.2663883](https://doi.org/10.1145/2663876.2663883).
- Canbek, Gürol, Seref Sagiroglu, and Nazife Baykal (2016). “New Comprehensive Taxonomies on Mobile Security and Malware Analysis”. In: *International Journal of Information Security* 5.4, pp. 106–138.
- Cao, Qiang, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro (2012). “Aiding the Detection of Fake Accounts in Large Scale Social Online Services”. In: *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. NSDI 2012. USENIX, pp. 197–210. URL: <https://dl.acm.org/doi/10.5555/2228298.2228319>.

- Carley, Kathleen M. (2020). “Social cybersecurity: an emerging science”. In: *Computational and Mathematical Organization Theory* 26.4, pp. 365–381. DOI: [10.1007/s10588-020-09322-9](https://doi.org/10.1007/s10588-020-09322-9).
- Chinchor, Nancy (1992). “MUC-4 Evaluation Metrics”. In: *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pp. 22–29. URL: <https://aclanthology.org/M92-1002>.
- Cimiano, Philipp, Andreas Hotho, and Steffen Staab (2004). “Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text”. In: *Proceedings of the 16th European Conference on Artificial Intelligence. ECAI’04*. IOS Press, pp. 435–439. URL: <https://dl.acm.org/doi/abs/10.5555/3000001.3000093>.
- CogNub (2018). *Cognitive Computing And Machine Learning*. URL: <https://web.archive.org/web/20181023121517/http://www.cognub.com/index.php/cognitive-platform>.
- Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson (2014). “Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data”. In: *Journal of Communication* 64.2, pp. 317–332. DOI: [10.1111/jcom.12084](https://doi.org/10.1111/jcom.12084).
- Cortes, Corinna and Vladimir Vapnik (1995). “Support-Vector Networks”. In: *Machine Learning* 20.3, pp. 273–297. DOI: [10.1007/BF00994018](https://doi.org/10.1007/BF00994018).
- Costa, Daniel, Michael Albrethsen, Matthew Collins, Samuel Perl, George Silowash, and Derrick Spooner (2016). *An Insider Threat Indicator Ontology*. Tech. rep. Cmu/sei-2016-tr-007. Pittsburgh, PA, USA: Software Engineering Institute, Carnegie Mellon University. URL: <http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=454613>.
- Danyliw, R. (2016). *The Incident Object Description Exchange Format Version 2*. IETF RFC 7970. URL: <https://tools.ietf.org/html/rfc7970>.

- Disney, Andrew (2020). “PageRank centrality & EigenCentrality”. In: *Cambridge-Intelligence.com*. URL: <https://cambridge-intelligence.com/eigencentality-pagerank>.
- Dodds, Peter Sheridan, Roby Muhamad, and Duncan J. Watts (2003). “An Experimental Study of Search in Global Social Networks”. In: *Science* 301.5634, pp. 827–829. DOI: [10.1126/science.1081058](https://doi.org/10.1126/science.1081058).
- Elahi, Golnaz, Eric Yu, and Nicola Zannone (2009). “A Modeling Ontology for Integrating Vulnerabilities into Security Requirements Conceptual Foundations”. In: *Proceedings of the 28th International Conference on Conceptual Modeling, Gramado, Brazil*. Springer, pp. 99–114. DOI: [10.1007/978-3-642-04840-1_10](https://doi.org/10.1007/978-3-642-04840-1_10).
- Ellison, Nicole B. and Danah M. Boyd (2013). “Sociality Through Social Network Sites”. In: *The Oxford Handbook of Internet Studies*. Oxford University Press. DOI: [10.1093/oxfordhb/9780199589074.013.0008](https://doi.org/10.1093/oxfordhb/9780199589074.013.0008).
- Elnagdy, Sam Adam, Meikang Qiu, and Keke Gai (2016). “Understanding Taxonomy of Cyber Risks for Cybersecurity Insurance of Financial Industry in Cloud Computing”. In: *Proceedings of 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing*. IEEE, pp. 295–300. DOI: [10.1109/CSCloud.2016.46](https://doi.org/10.1109/CSCloud.2016.46).
- European Commission, Joint Research Centre (JRC) (2021). *JRC Cybersecurity Taxonomy*. [Dataset]. URL: <http://data.europa.eu/89h/d2f56334-a0df-485b-8dc8-2c0039d31122>.
- Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini (2016). “The Rise of Social Bots”. In: *Communications of the ACM* 59.7, pp. 96–104. DOI: [10.1145/2818717](https://doi.org/10.1145/2818717).
- Fortunato, Santo (2010). “Community detection in graphs”. In: *Physics Reports* 486.3, pp. 75–174. DOI: [10.1016/j.physrep.2009.11.002](https://doi.org/10.1016/j.physrep.2009.11.002).
- Fowler, Bella (2019). “The Exact History of ‘Six Degrees of Kevin Bacon’”. In: *NZ Herald*. URL: <https://www.nzherald.co.nz/entertainment/the-exact-history-of-six-degrees-of-kevin-bacon/NCVVAU73UZ4TNCZAK726ENOBQ>.

- Freeman, Linton C. (1977). “A Set of Measures of Centrality Based on Betweenness”. In: *Sociometry* 40.1, pp. 35–41. DOI: [10.2307/3033543](https://doi.org/10.2307/3033543).
- (1978). “Centrality in Social Networks Conceptual Clarification”. In: *Social Networks* 1.3, pp. 215–239. DOI: [10.1016/0378-8733\(78\)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7).
- (2004). *The Development of Social Network Analysis – with an Emphasis on Recent Events*. SAGE. DOI: [10.4135/9781446294413](https://doi.org/10.4135/9781446294413).
- Friedman, Jerome H. (2001). “Greedy Function Approximation: A Gradient Boosting Machine”. In: *The Annals of Statistics* 29.5, pp. 1189–1232. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451).
- GeoNames (2022). *Cities*. URL: <http://www.geonames.org>.
- Georgescu, Tiberiu and Ion Smeureanu (2017). “Using Ontologies in Cybersecurity Field”. In: *Informatica Economică* 21.3, pp. 5–15. DOI: [10.12948/issn14531305/21.3.2017.01](https://doi.org/10.12948/issn14531305/21.3.2017.01).
- Gibb, Will (2013). *Back to Basics Series: OpenIOC*. URL: <https://www.fireeye.com/blog/threat-research/2013/09/basics-series-openioc.html>.
- Girvan, M. and Mark E. J. Newman (2002). “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences (PNAS)* 99.12, pp. 7821–7826. DOI: [10.1073/pnas.122653799](https://doi.org/10.1073/pnas.122653799).
- González-Bailón, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno (2014). “Assessing the bias in samples of large online networks”. In: *Social Networks* 38, pp. 16–27. DOI: <https://doi.org/10.1016/j.socnet.2014.01.004>.
- Grieve, Guy (2020). “Explained SIEMply: Machine Learning”. In: *Log Point*. URL: <https://www.logpoint.com/en/blog/explained-siemply-machine-learning>.
- Groshek, Jacob, Vincent Mees, and Rob Eschmann (2020). “Modeling influence and community in social media data using the digital methods initiative-Twitter capture and analysis toolkit (DMI-TCAT) and Gephi”. In: *MethodsX* 7, 101164:1–101164:4. DOI: [10.1016/j.mex.2020.101164](https://doi.org/10.1016/j.mex.2020.101164).

- Grove, Adam J. and Dale Schuurmans (1998). “Boosting in the Limit: Maximizing the Margin of Learned Ensembles”. In: *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*. AAAI '98/IAAI '98. American Association for Artificial Intelligence, pp. 692–699. URL: <https://dl.acm.org/doi/10.5555/295240.295766>.
- Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart (2008). “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference (SciPy 2008)*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, pp. 11–15.
- Hajeer, Mustafa H., Alka Singh, Dipankar Dasgupta, and Sugata Sanyal (2013). *Clustering online social network communities using genetic algorithms*. DOI: [10.48550/arXiv.1312.2237](https://doi.org/10.48550/arXiv.1312.2237).
- Hanneman, Robert A. and Mark Riddle (2005). *Introduction to Social Network Methods*. URL: <http://faculty.ucr.edu/~hanneman/nettext>.
- Heartfield, Ryan, George Loukas, Sanja Budimir, Anatolij Bezemskij, Johnny R.J. Fontaine, Avgoustinos Filippoupolitis, and Etienne Roesch (2018). “A taxonomy of cyber-physical threats and impact in the smart home”. In: *Computers & Security* 78, pp. 398–428. DOI: [10.1016/j.cose.2018.07.011](https://doi.org/10.1016/j.cose.2018.07.011).
- Hedden, Steve (2019). “Who is the most important person in the film industry?” In: *Towards Data Science*. URL: <https://towardsdatascience.com/who-is-the-most-important-person-in-the-film-industry-61d4fd6980be>.
- Heidemann, Julia, Mathias Klier, and Florian Probst (2012). “Online Social Networks: A Survey of a Global Phenomenon”. In: *Computer Networks* 56.18, pp. 3866–3878. DOI: [10.1016/j.comnet.2012.08.009](https://doi.org/10.1016/j.comnet.2012.08.009).
- Hines, Kristi (2022). “The History Of Social Media”. In: *Search Engine Journal*. URL: <https://www.searchenginejournal.com/social-media-history/462643>.
- Hipo (2022). *University Domains*. URL: <https://github.com/Hipo/university-domains-list>.

- Huberman, Bernardo, Daniel Romero, and Fang Wu (2008). “Social Networks that Matter: Twitter Under the Microscope”. In: *First Monday* 14. DOI: [10.2139/ssrn.1313405](https://doi.org/10.2139/ssrn.1313405).
- Hutto, C. J. and Eric Gilbert (2014). “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1, pp. 216–225. DOI: [10.1609/icwsm.v8i1.14550](https://doi.org/10.1609/icwsm.v8i1.14550).
- ISO, International Organization for Standardization (2018). *Information technology – Security techniques – Information security management systems – Overview and vocabulary*. ISO/IEC 27000:2018. URL: <https://www.iso.org/standard/73906.html>.
- Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian (2014). “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software”. In: *PLoS ONE* 9.6, e98679:1–e98679:12. DOI: [10.1371/journal.pone.0098679](https://doi.org/10.1371/journal.pone.0098679).
- James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer International Publishing. DOI: [10.1007/978-3-031-38747-0](https://doi.org/10.1007/978-3-031-38747-0).
- Japkowicz, Nathalie and Mohak Shah (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press. DOI: [10.1017/CB09780511921803](https://doi.org/10.1017/CB09780511921803).
- Jiang, Tammy, Jaimie L. Gradus, and Anthony J. Rosellini (2020). “Supervised Machine Learning: A Brief Primer”. In: *Behavior Therapy* 51.5, pp. 675–687. DOI: [10.1016/j.beth.2020.05.002](https://doi.org/10.1016/j.beth.2020.05.002).
- Jiang, Yuning, Manfred Jeusfeld, Yacine Atif, Jianguo Ding, Christoffer Brax, and Eva Nero (2018). “A Language and Repository for Cyber Security of Smart Grids”. In: *Proceedings of the IEEE 22nd International Enterprise Distributed*

-
- Object Computing Conference (EDOC)*. IEEE, pp. 164–170. DOI: [10.1109/edoc.2018.00029](https://doi.org/10.1109/edoc.2018.00029).
- Jones, Keenan, Jason R. C. Nurse, and Shujun Li (2020). “Behind the Mask: A Computational Study of Anonymous’ Presence on Twitter”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 1. AAAI, pp. 327–338. DOI: [10.1609/icwsm.v14i1.7303](https://doi.org/10.1609/icwsm.v14i1.7303).
- (2022). “Out of the Shadows: Analyzing Anonymous’ Twitter Resurgence during the 2020 Black Lives Matter Protests”. In: *Proceedings of the International Conference on Web and Social Media* 16.1, pp. 417–428. DOI: [10.1609/icwsm.v16i1.19303](https://doi.org/10.1609/icwsm.v16i1.19303).
- June, Abbas (2010). *Structures for Organizing Knowledge: Exploring Taxonomies, Ontologies, and Other Schema*. Neal-Schuman Publishers, Inc.
- Kanade, Vijay (2022). “What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices”. In: *SpiceWorks.com*. URL: <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression>.
- Kaplan, Andreas M. and Michael Haenlein (2010). “Users of the World, Unite! The Challenges and Opportunities of Social Media”. In: *Business Horizons* 53.1, pp. 59–68. DOI: [10.1016/j.bushor.2009.09.003](https://doi.org/10.1016/j.bushor.2009.09.003).
- Karinthy, Frigyes (1929). “Chain-links”. In: *Everything is different*, pp. 21–26. URL: https://djjr-courses.wdfiles.com/local--files/soc180:karinthy-chain-links/Karinthy-Chain-Links_1929.pdf.
- Kemp, Simon (2021). *Digital 2021: Global Overview Report*. Data Reportal. URL: <https://datareportal.com/reports/digital-2021-global-overview-report>.
- Kigerl, Alex (2018). “Profiling Cybercriminals: Topic Model Clustering of Carding Forum Member Comment Histories”. In: *Social Science Computer Review* 36.5, pp. 591–609. DOI: [10.1177/0894439317730296](https://doi.org/10.1177/0894439317730296).
- Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom (1975). “Derivation of new readability formulas (Automated Readability Index,

- Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel”. In:
URL: <https://stars.library.ucf.edu/istlibrary/56>.
- Kiss, Christine, Andreas Scholz, and Martin Bichler (2006). “Evaluating Centrality Measures in Large Call Graph”. In: *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services*. CEC/EEE’06. DOI: [10.1109/cec-eee.2006.44](https://doi.org/10.1109/cec-eee.2006.44).
- Krithiga, R. and E. Ilavarasan (2019). “A Comprehensive Survey of Spam Profile Detection Methods in Online Social Networks”. In: vol. 1362. 1. IOP Publishing. DOI: [10.1088/1742-6596/1362/1/012111](https://doi.org/10.1088/1742-6596/1362/1/012111).
- Krzyk, Kamil (2018). “Types of Machine Learning”. In: *TowardsDataScience.com*. URL: <https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d>.
- Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon (2010). “What is Twitter, a Social Network or a News Media?” In: *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. Association for Computing Machinery, pp. 591–600. DOI: [10.1145/1772690.1772751](https://doi.org/10.1145/1772690.1772751).
- Kwasnik, Barbara H (1999). “The Role of Classification in Knowledge Representation and Discovery”. In: *Library Trends* 48.1, pp. 22–47. URL: <https://api.semanticscholar.org/CorpusID:9769901>.
- Lambe, Patrick (2014). *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Chandos Knowledge Management. Elsevier Science. ISBN: 9781780632001. URL: <https://www.sciencedirect.com/book/9781843342274/organising-knowledge>.
- Lambiotte, Renaud, Jean-Charles Delvenne, and Mauricio Barahona (2014). “Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks”. In: *IEEE Transactions on Network Science and Engineering* 1.2, pp. 76–90. DOI: [10.1109/tnse.2015.2391998](https://doi.org/10.1109/tnse.2015.2391998).

- Laubheimer, Page (2022). *Taxonomy 101: Definition, Best Practices, and How It Complements Other IA Work*. Accessed: 2024-05-19. URL: <https://www.nngroup.com/articles/taxonomy-101/>.
- Lee, Kuo-Chan, Chih-Hung Hsieh, Li-Jia Wei, Ching-Hao Mao, Jyun-Han Dai, and Yu-Ting Kuang (2017). “Sec-Buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation”. In: *Soft Computing* 21.11, pp. 2883–2896. DOI: [10.1007/s00500-016-2265-0](https://doi.org/10.1007/s00500-016-2265-0).
- Lippmann, Richard P., William M. Campbell, David J. Weller-Fahy, Alyssa C. Mensch, Giselle M. Zeno, and Joseph P. Campbell (2017). “Toward Finding Malicious Cyber Discussions in Social Media”. In: *Proceedings of AAAI 2017 Workshops*. AAAI, pp. 203–209.
- Liu, Leqi, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen Ebrahimi Moghadam, and Lyle Ungar (2016). “Analyzing Personality through Social Media Profile Picture Choice”. In: *Proceedings of the International Conference on Web and Social Media 2016*. AAAI, pp. 211–220. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14738>.
- Loria, Steven (2022). *TextBlob: Simplified Text Processing*. URL: <https://textblob.readthedocs.io/en/dev>.
- Loukas, George, Eirini Karapistoli, Emmanouil Panaousis, Panagiotis Sarigiannidis, Anatolij Bezemskij, and Tuan Vuong (2019). “A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles”. In: *Ad Hoc Networks* 84, pp. 124–147. DOI: [10.1016/j.adhoc.2018.10.002](https://doi.org/10.1016/j.adhoc.2018.10.002).
- Luijff, H.A.M. and A.H. Nieuwenhuijs (2008). “Extensible threat taxonomy for critical infrastructures”. In: *International Journal of Critical Infrastructures* 4.4, pp. 409–417. DOI: [10.1504/ijcis.2008.020159](https://doi.org/10.1504/ijcis.2008.020159).
- Ma, Feicheng (2015). ““Six Degrees of Separation” and “Small World””. In: *Information Communication*. Springer International Publishing, pp. 77–78. DOI: [10.1007/978-3-031-02293-7_8](https://doi.org/10.1007/978-3-031-02293-7_8).

- Mahaini, Mohamad Imad and Shujun Li (2021). “Detecting Cyber Security Related Twitter Accounts and Different Sub-Groups: A Multi-Classifer Approach”. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, pp. 599–606. DOI: [10.1145/3487351.3492716](https://doi.org/10.1145/3487351.3492716).
- (2023). “Cyber Security Researchers on Online Social Networks: From the Lens of the UK’s ACEs-CSR on Twitter”. In: *Security and Privacy in Social Networks and Big Data: 9th International Symposium, SocialSec 2023, Canterbury, UK, August 14–16, 2023, Proceedings*. Vol. 14097. Lecture Notes in Computer Science. Springer, pp. 129–148. DOI: [10.1007/978-981-99-5177-2_8](https://doi.org/10.1007/978-981-99-5177-2_8).
- Mahaini, Mohamad Imad, Shujun Li, and Rahime Belen-Sağlam (2019). “Building Taxonomies based on Human-Machine Teaming: Cyber Security as an Example”. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. ACM, 30:1–30:9. DOI: [10.1145/3339252.3339282](https://doi.org/10.1145/3339252.3339282).
- Malvern, D., B. Richards, N. Chipere, and P. Durán (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan UK. ISBN: 9780230511804. URL: <https://books.google.co.uk/books?id=R78WDAAAQBAJ>.
- Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60. URL: <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- Mao, Yuchen, Lijun Zhou, and Naixue Xiong (2020). “Identify Influential Nodes in Online Social Network for Brand Communication”. In: *ArXiv*. DOI: [10.48550/arXiv.2006.14104](https://doi.org/10.48550/arXiv.2006.14104).
- Martins, António Rocha (2019). “The zoon politikon: Medieval Aristotelian Interpretations”. In: *Revista Portuguesa de Filosofia* 75.3, pp. 1539–1574. ISSN: 08705283, 2183461X. DOI: [10.17990/RPF/2019_75_3_1539](https://doi.org/10.17990/RPF/2019_75_3_1539).

- Mavroeidis, Vasileios and Siri Bromander (2017). “Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence”. In: *Proceedings of 2017 European Intelligence and Security Informatics Conference*. IEEE, pp. 91–98. DOI: [10.1109/eisic.2017.20](https://doi.org/10.1109/eisic.2017.20).
- Maxwell, John C. (2019). *Your Influence Inventory*. URL: <https://johnmaxwellleadershippodcast.com/episodes/john-maxwell-your-influence-inventory>.
- McAlaney, John, Sarah Hambidge, Emily Kimpton, and Helen Thackray (2020). “Knowledge is power: An analysis of discussions on hacking forums”. In: *Proceedings of IEEE European Symposium on Security and Privacy Workshops*. EuroS&PW 2020. IEEE, pp. 477–483. DOI: [10.1109/EuroSPW51379.2020.00070](https://doi.org/10.1109/EuroSPW51379.2020.00070).
- McLaughlin, G. Harry (1969). “SMOG Grading-a New Readability Formula”. In: *Journal of Reading* 12.8, pp. 639–646. ISSN: 00224103. URL: <http://www.jstor.org/stable/40011226>.
- Milgram, Stanley (1967). “The Small-World Problem”. In: *Psychology Today* 2.1, pp. 60–67. URL: <http://snap.stanford.edu/class/cs224w-readings/milgram67smallworld.pdf>.
- Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill.
- Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of Machine Learning*. MIT Press. ISBN: 9780262351362.
- Mohsen, Fadi, Cornelis Zwart, Dimka Karastoyanova, and Georgi Gaydadjiev (2022). “A Taxonomy for Large-Scale Cyber Security Attacks”. In: *EAI Endorsed Transactions on Cloud Systems* 7.21. DOI: [10.4108/eai.2-3-2022.173548](https://doi.org/10.4108/eai.2-3-2022.173548).
- Moscato, Vincenzo and Giancarlo Sperli (2021). “A survey about community detection over On-line Social and Heterogeneous Information Networks”. In: *Knowledge-Based Systems* 224, 107112:1–107112:13. DOI: [10.1016/j.knosys.2021.107112](https://doi.org/10.1016/j.knosys.2021.107112).
- Nai Fovino, Igor, Ricardo Neisse, Jose Luis Hernandez Ramos, Nineta Polemi, Gian Luigi Ruzzante, Malgorzata Figwer, and Alessandro Lazari (2019). *A Proposal for*

- a *European Cybersecurity Taxonomy*. EUR 29868 EN. JRC118089. Publications Office of the European Union. DOI: [10.2760/106002](https://doi.org/10.2760/106002).
- Navlani, Avinash (2023). “Decision Tree Classification in Python Tutorial”. In: *DataCamp.com*. URL: <https://www.datacamp.com/tutorial/decision-tree-classification-python>.
- NCSC, National Cyber Security Centre (2019). *Academic Centres of Excellence in Cyber Security Research*. URL: <https://www.ncsc.gov.uk/information/academic-centres-excellence-cyber-security-research>.
- NeCS (2015). *European Network for Cyber Security*. URL: <http://necs-project.eu/about>.
- Nettleton, David (2014). “Chapter 11 - Text Analysis”. In: *Commercial Data Mining*. Morgan Kaufmann, pp. 171–179. DOI: [10.1016/B978-0-12-416602-8.00011-X](https://doi.org/10.1016/B978-0-12-416602-8.00011-X).
- New Idea Engineering Inc. (2018). *What’s the difference between Taxonomies and Ontologies? - Ask Dr. Search*. URL: <http://www.ideaeng.com/taxonomies-ontologies-0602>.
- Newman, Mark E. J. (2008). “The Mathematics of Networks”. In: *The New Palgrave Encyclopedia of Economics 2*, pp. 1–12. DOI: [10.1057/9780230226203.1064](https://doi.org/10.1057/9780230226203.1064).
- (2016). “Equivalence between modularity optimization and maximum likelihood methods for community detection”. In: *Physical Review E* 94 (5), 052315:1–052315:8. DOI: [10.1103/PhysRevE.94.052315](https://doi.org/10.1103/PhysRevE.94.052315).
- (2018). *Networks*. Second. Oxford University Press. DOI: [10.1093/oso/9780198805090.001.0001](https://doi.org/10.1093/oso/9780198805090.001.0001).
- Newman, Mark E. J. and M. Girvan (2004). “Finding and evaluating community structure in networks”. In: *Physical Review E* 69 (2), 026113:1–026113:15. DOI: [10.1103/PhysRevE.69.026113](https://doi.org/10.1103/PhysRevE.69.026113).
- Nickerson, Robert C., Upkar Varshney, and Jan Muntermann (2013). “A method for taxonomy development and its application in information systems”. In: *European Journal of Information Systems* 22.3, pp. 336–359. DOI: [10.1057/ejis.2012.26](https://doi.org/10.1057/ejis.2012.26).

- NLTK Python (2022). *NLTK VADER: Sentiment Analysis Tools*. URL: <https://www.nltk.org/api/nltk.sentiment.vader.html>.
- NLTK Toolkit (2023). *NLTK: Natural Language Toolkit*. URL: <https://www.nltk.org>.
- Nouh, Mariam and Jason R. C. Nurse (2015). “Identifying Key-Players in Online Activist Groups on the Facebook Social Network”. In: *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop*. IEEE, pp. 969–978. DOI: [10.1109/icdmw.2015.88](https://doi.org/10.1109/icdmw.2015.88).
- Open, OASIS (2019). *Introduction to STIX*. URL: <https://oasis-open.github.io/cti-documentation/stix/intro>.
- Otte, Evelien and Ronald Rousseau (2002). “Social network analysis: a powerful strategy, also for the information sciences”. In: *Journal of Information Science* 28.6, pp. 441–453. DOI: [10.1177/016555150202800601](https://doi.org/10.1177/016555150202800601).
- Pasca, Marius and Benjamin Van Durme (2007). “What You Seek is What You Get: Extraction of Class Attributes from Query Logs”. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 7. Morgan Kaufmann Publishers Inc., pp. 2832–2837. DOI: [10.5555/1625275.1625731](https://doi.org/10.5555/1625275.1625731).
- Pastrana, Sergio, Daniel R. Thomas, Alice Hutchings, and Richard Clayton (2018). “CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale”. In: *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, pp. 1845–1854. DOI: [10.1145/3178876.3186178](https://doi.org/10.1145/3178876.3186178).
- Pattnaik, Nandita, Shujun Li, and Jason R. C. Nurse (2023). “Perspectives of non-expert users on cyber security and privacy: An analysis of online discussions on Twitter”. In: *Computers & Security* 125, 103008:1–103008:15. DOI: [10.1016/j.cose.2022.103008](https://doi.org/10.1016/j.cose.2022.103008).
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu

- Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: <https://jmlr.org/papers/v12/pedregosa11a.html>.
- Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh (2011). “Predicting age and gender in online social networks”. In: *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*. SMUC ’11. ACM, pp. 37–44. DOI: [10.1145/2065023.2065035](https://doi.org/10.1145/2065023.2065035).
- Pennacchiotti, Marco and Ana-Maria Popescu (2011a). “A Machine Learning Approach to Twitter User Classification”. In: *Proceedings of the International Conference on Web and Social Media*. AAAI, pp. 281–288. DOI: [10.1609/icwsm.v5i1.14139](https://doi.org/10.1609/icwsm.v5i1.14139).
- (2011b). “Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter”. In: *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*. KDD ’11. ACM, pp. 430–438. DOI: [10.1145/2020408.2020477](https://doi.org/10.1145/2020408.2020477).
- Pinheiro, Carlos Andre Reis (2011). *Social Network Analysis in Telecommunications*. John Wiley & Sons. ISBN: 9781118010952. URL: <https://www.wiley.com/engb/Social+Network+Analysis+in+Telecommunications-p-9781118010952>.
- Prat, Nathalie, Isabelle Comyn-Wattiau, and Jacky Akoka (2015). “A Taxonomy of Evaluation Methods for Information Systems Artifacts”. In: *Journal of Management Information Systems* 32.3, pp. 229–267. DOI: [10.1080/07421222.2015.1099390](https://doi.org/10.1080/07421222.2015.1099390).
- Preoțiuc-Pietro, Daniel, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras (2015). “Studying User Income through Language, Behaviour and Affect in Social Media”. In: *PLOS ONE* 10.9. DOI: [10.1371/journal.pone.0138717](https://doi.org/10.1371/journal.pone.0138717).
- Prieto-Díaz, Rubén (1991). “Implementing Faceted Classification for Software Reuse”. In: *ACM* 34.5, pp. 88–97. DOI: [10.1145/103167.103176](https://doi.org/10.1145/103167.103176).

- Priss, Uta (2006). “Formal Concept Analysis in Information Science”. In: *Annual Review of Information Science and Technology* 40.1, pp. 521–543. DOI: [10.1002/aris.1440400120](https://doi.org/10.1002/aris.1440400120).
- Quinlan, J. Ross (1987). “Generating Production Rules from Decision Trees”. In: *Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 1*. Vol. 87. IJCAI’87. Citeseer. Morgan Kaufmann Publishers Inc., pp. 304–307. DOI: [10.5555/1625015.1625078](https://doi.org/10.5555/1625015.1625078).
- Radmand, Pedram, Alex Talevski, Stig Petersen, and Simon Carlsen (2010). “Taxonomy of Wireless Sensor Network Cyber Security Attacks in the Oil and Gas Industries”. In: *Proceedings of 2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE, pp. 949–957. DOI: [10.1109/aina.2010.175](https://doi.org/10.1109/aina.2010.175).
- Ralph, Paul (2019). “Toward Methodological Guidelines for Process Theories and Taxonomies in Software Engineering”. In: *IEEE Transactions on Software Engineering* 45.7, pp. 712–735. DOI: [10.1109/TSE.2018.2796554](https://doi.org/10.1109/TSE.2018.2796554).
- Razzaq, Abdul, Khalid Latif, H. Farooq Ahmad, Ali Hur, Zahid Anwar, and Peter Charles Bloodsworth (2014). “Semantic security against web application attacks”. In: *Information Sciences* 254, pp. 19–38. DOI: [10.1016/j.ins.2013.08.007](https://doi.org/10.1016/j.ins.2013.08.007).
- Řehůřek, Radim (2022a). *Gensim: Topic Coherence for Topic Models*. URL: <https://radimrehurek.com/gensim/models/coherencemodel.html>.
- (2022b). *Gensim: Topic Modelling for Humans*. URL: <https://radimrehurek.com/gensim/index.html>.
- Reynolds, Patrick (1999). *The Oracle of Bacon*. URL: <https://oracleofbacon.org/graph.php>.
- Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). “Exploring the Space of Topic Coherence Measures”. In: *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM, pp. 399–408. DOI: [10.1145/2684822.2685324](https://doi.org/10.1145/2684822.2685324).

- Sabidussi, Gert (1966). “The Centrality Index of a Graph”. In: *Psychometrika* 31.4, pp. 581–603. DOI: [10.1007/BF02289527](https://doi.org/10.1007/BF02289527).
- Saini, Anshul (2023). “Guide on Support Vector Machine (SVM) Algorithm”. In: *AnalyticsVidhya.com*. URL: <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners>.
- Saura, Jose Ramon, Daniel Palacios-Marqués, and Domingo Ribeiro-Soriano (2021). “Using data mining techniques to explore security issues in smart living environments in Twitter”. In: *Computer Communications* 179, pp. 285–295. DOI: [10.1016/j.comcom.2021.08.021](https://doi.org/10.1016/j.comcom.2021.08.021).
- Schneider, Fabian, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger (2009). “Understanding Online Social Network Usage from a Network Perspective”. In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. IMC '09. Association for Computing Machinery, pp. 35–48. DOI: [10.1145/1644893.1644899](https://doi.org/10.1145/1644893.1644899).
- Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu (2017). “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN”. In: *ACM Transactions on Database Systems* 42.3, 19:1–19:21. DOI: [10.1145/3068335](https://doi.org/10.1145/3068335).
- Sedjelmaci, Hichem and Sidi Mohamed Senouci (2018). “Cyber Security Methods for Aerial Vehicle Networks: Taxonomy, Challenges and Solution”. In: *Journal of Supercomputing* 74.10, pp. 4928–4944. DOI: [10.1007/s11227-018-2287-8](https://doi.org/10.1007/s11227-018-2287-8).
- Sharma, Abhishek (2023). “Random Forest vs Decision Tree — Which Is Right for You?” In: *AnalyticsVidhya.com*. URL: <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm>.
- Sievert, Carson and Kenneth Shirley (2014). “LDAvis: A method for visualizing and interpreting topics”. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. ACL, pp. 63–70. DOI: [10.3115/v1/W14-3110](https://doi.org/10.3115/v1/W14-3110).

- Simha, Anirudha (2021). “Understanding TF-IDF for Machine Learning”. In: *CapitalOone*. URL: <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf>.
- Simmons, Chris, Charles Ellis, Sajjan Shiva, Dipankar Dasgupta, and Qishi Wu (2009). “AVOIDIT: A Cyber Attack Taxonomy”. In: *CTIT Technical Reports Series*. URL: <https://api.semanticscholar.org/CorpusID:349528>.
- Sklearn LDA (2022). *Scikit-learn: Latent Dirichlet Allocation*. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>.
- Smith, M., A. Ceni, N. Milic-Frayling, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec, and C. Dunne (2010). *NodeXL: Network Overview, Discovery and Exploration for Excel*. URL: <https://www.smrfoundation.org/nodexl>.
- Sokolova, Marina and Guy Lapalme (2009). “A systematic analysis of performance measures for classification tasks”. In: *Information Processing & Management* 45.4, pp. 427–437. DOI: [10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002).
- Soni, Kaushik (2022). *locationtagger*. URL: <https://pypi.org/project/locationtagger>.
- Stanford NLP Group (2019). *CoreNLP for Natural Language Processing in Java*. URL: <https://stanfordnlp.github.io/CoreNLP/>.
- Stecanella, Bruno (2019). “Understanding TF-IDF for Machine Learning”. In: *MonkeyLearn*. URL: <https://monkeylearn.com/blog/what-is-tf-idf>.
- Subasi, Abdulhamit (2020). “Chapter 3 - Machine learning techniques”. In: *Practical Machine Learning for Data Analysis Using Python*. Academic Press, pp. 91–202. DOI: [10.1016/B978-0-12-821379-7.00003-5](https://doi.org/10.1016/B978-0-12-821379-7.00003-5).
- Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. 2nd. MIT Press. ISBN: 0262039249.
- Syafrizal, Melwin, Siti Rahayu Selamat, and Nurul Azma Zakaria (2021). “AVOID-ITALS: Enhanced Cyber-attack Taxonomy in Securing Information Technology

- Infrastructure”. In: *International Journal of Computer Science & Network Security* 21.8, pp. 1–12. DOI: [10.22937/IJCSNS.2021.21.8.1](https://doi.org/10.22937/IJCSNS.2021.21.8.1).
- Szopinski, Daniel, Thorsten Schoormann, and Dennis Kundisch (2020). “Criteria as a Prelude for Guiding Taxonomy Evaluation”. In: *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*. ScholarSpace, pp. 1–10. DOI: [10.24251/HICSS.2020.622](https://doi.org/10.24251/HICSS.2020.622).
- Tausczik, Yla R. and James W. Pennebaker (2010). “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods”. In: *Journal of Language and Social Psychology* 29. DOI: [10.1177/0261927X09351676](https://doi.org/10.1177/0261927X09351676).
- Tavabi, Nazgol, Nathan Bartley, Andres Abeliuk, Sandeep Soni, Emilio Ferrara, and Kristina Lerman (2019). “Characterizing Activity on the Deep and Dark Web”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, pp. 206–213. DOI: [10.1145/3308560.3316502](https://doi.org/10.1145/3308560.3316502).
- Thakur, Kutub, Thaier Hayajneh, and Jason Tseng (2019). “Cyber Security in Social Media: Challenges and the Way Forward”. In: *IT Professional* 21.2, pp. 41–49. DOI: [10.1109/MITP.2018.2881373](https://doi.org/10.1109/MITP.2018.2881373).
- Thelwall, Mike (2009). “Chapter 2 Social Network Sites: Users and Uses”. In: *Social Networking and The Web*. Vol. 76. Advances in Computers. Elsevier, pp. 19–73. DOI: [10.1016/S0065-2458\(09\)01002-X](https://doi.org/10.1016/S0065-2458(09)01002-X).
- Traag, Vincent A., Ludo Waltman, and Nees Jan van Eck (2019). “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific Reports* 9.1, 5233:1–5233:12. DOI: [10.1038/s41598-019-41695-z](https://doi.org/10.1038/s41598-019-41695-z).
- Treadstone71 (2014). *Treadstone71 Cyber Attack Taxonomy*. Website: The Cyber Shafarat – Treadstone 71. URL: <https://treadstone71llc.files.wordpress.com/2014/11/cyber-attack-taxonomy-treadstone-71.jpg>.
- Trevor, Hastie, Tibshirani Robert, and Friedman Jerome (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).

- Tsvetovat, Maksim and Alexander Kouznetsov (2011). *Social Network Analysis for Startups*. O'Reilly Media Inc. ISBN: 9781449306465.
- UK Ministry of Defence (2018). *Human-Machine Teaming*. Joint Concept Note 1/18. URL: <https://www.gov.uk/government/publications/human-machine-teaming-jcn-118>.
- Unterkalmsteiner, Michael and Waleed Adbeen (2023). “A compendium and evaluation of taxonomy quality attributes”. In: *Expert Systems* 40.1. DOI: [10.1111/exsy.13098](https://doi.org/10.1111/exsy.13098). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13098>.
- Usman, Muhammad, Ricardo Britto, Jürgen Börstler, and Emilia Mendes (2017). “Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method”. In: *Information and Software Technology* 85, pp. 43–59. DOI: [10.1016/j.infsof.2017.01.006](https://doi.org/10.1016/j.infsof.2017.01.006). URL: <https://www.sciencedirect.com/science/article/pii/S0950584917300472>.
- Vapnik, Vladimir Naumovich (1999). *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer. ISBN: 9780387987804. URL: <https://books.google.co.uk/books?id=sna9BaxVbj8C>.
- Wang, Ju An and Minzhe Guo (2009). “OVM: An Ontology for Vulnerability Management”. In: *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research*. CSIIRW '09. ACM, pp. 1–4. DOI: [10.1145/1558607.1558646](https://doi.org/10.1145/1558607.1558646).
- Warner, Andrew (2014). “The father of social networking”. In: *Mixergy*. URL: <https://mixergy.com/interviews/andrew-weinreich-sixdegrees>.
- Wasserman, Stanley and Katherine Faust (1994). *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge University Press. DOI: [10.1017/CB09780511815478](https://doi.org/10.1017/CB09780511815478).
- Watts, Duncan J (2004). *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, Inc.

- We Are Social Inc. (2023). *Digital 2023: What We Learned*. Special Report. URL: <https://wearesocial.com/uk/blog/2023/01/digital-2023>.
- Wellman, Barry and S.D. Berkowitz (1988). *Social Structures: A Network Approach*. Cambridge: Cambridge University Press.
- West, Douglas Brent (2001). *Introduction to Graph Theory*. Vol. 2. Prentice Hall Upper Saddle River. ISBN: 0130144002.
- Wood, Thomas (2019). “What is the F-score?” In: *DeepAI.org*. URL: <https://deepai.org/machine-learning-glossary-and-terms/f-score>.
- Yang, Hui and Jamie Callan (2009). “A Metric-based Framework for Automatic Taxonomy Induction”. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. ACL, pp. 271–279. URL: <https://aclanthology.org/P09-1031>.
- Yang, Yiming and Jan O. Pedersen (1997). “A Comparative Study on Feature Selection in Text Categorization”. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. ICML '97. Morgan Kaufmann Publishers Inc., pp. 412–420. URL: <https://dl.acm.org/doi/10.5555/645526.657137>.
- Zamfira, Andrei C. and Horia Ciocarlie (2018). “Developing An Ontology Of Cyber-Operations In Networks Of Computers”. In: *Proceedings of 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing*. IEEE, pp. 395–400. DOI: [10.1109/iccp.2018.8516644](https://doi.org/10.1109/iccp.2018.8516644).
- Zeng, Yutong, Honghao Yu, Tiejun Wu, Yong Chen, Xing Lan, and Cheng Huang (2023). “CSCD: A Cyber Security Community Detection Scheme on Online Social Networks”. In: *Digital Forensics and Cyber Crime*. Springer Nature Switzerland, pp. 135–150. DOI: [10.1007/978-3-031-36574-4_8](https://doi.org/10.1007/978-3-031-36574-4_8).