FACT-CHECKING ECOSYSTEM: FROM MODELLING TO SEMI-AUTOMATED STAKEHOLDER-BASED DETECTION OF FALSE INFORMATION

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT

IN THE SUBJECT OF COMPUTER SCIENCE

FOR THE DEGREE

OF PHD.

By

Enes Altuncu

January 2024

Acknowledgements

First and foremost, I would like to thank Allah the Almighty for giving me strength throughout my PhD studies and granting me to complete this thesis.

Then, I would like to thank my amazing PhD supervisors, Shujun Li and Jason Nurse, for their continuous support, guidance, and encouragement. I learned a lot from them during my PhD journey, which made me feel better prepared to survive the jungle of academia as an independent researcher.

I would also like to thank my beloved mother, father, and brother for their endless support and prayers, which kept me motivated even during the hardest times of my PhD. I also extend thanks to my big family for their prayers for my success.

I also owe thanks to my supervisory panel members, Marek Grzes, Sally Fincher, and Rogerio de Lemos, for their valuable feedback in the annual reviews. And, special thanks to Haiyue Yuan for his feedback on the thesis. In addition, I would like to thank all the members of the School of Computing and the Institute of Cyber Security for Society at the University of Kent for their collaboration. Finally, huge thanks to my PhD Viva examiners, Anna Jordanous, Yulan He, and Harith Alani, for the colourful discussion and their invaluable feedback on my thesis.

Last but not least, I would like to thank the Ministry of Education, Republic of Türkiye, for funding my PhD studies through the YLSY scholarship program, which made this entire journey possible.

Abstract

Humanity has suffered from false, misleading, and malicious information, in short, problematic information, throughout history. Nevertheless, the advances in digital technologies, including the Internet, social media, and smart devices, accelerated and facilitated the dissemination of problematic information.

As a complex phenomenon with several aspects to be studied, problematic information remains within many disciplines' areas of interest. Nevertheless, the lack of a unified understanding of the phenomenon complicates the development of more effective solutions that benefit from multidisciplinary literature to detect problematic information. Focusing on a common type of problematic information, false information, this thesis aims to contribute to filling this gap by presenting a more comprehensive approach to understanding and combating false information. To this end, a topical analysis of the relevant literature was conducted through topic modelling over research publications. This was complemented by the generation of an entity-relationship model of the ecosystem to develop a better conceptual understanding of the landscape and to produce a tool that can be used to guide conceptual analysis of false information-related work and scenarios. Taking a closer look at the proposed model, then, user attitudes and behaviours towards fact-checking tools were examined through an online user survey designed with the help of a new taxonomy of fact-checking tools to collect some empirical evidence for the design of more trustworthy solutions. Most importantly, the

conducted survey revealed that semi-automated fact-checking systems, combining human intelligence with automation speed and scalability, can build a higher level of trust among users, compared to manual and fully automated solutions. Therefore, based on the derived conceptual understanding and supported by the obtained empirical evidence, a semi-automated fact-checking tool that is entirely based on expert opinion discovery called *aedFaCT* was designed, implemented and tested.

The presented research in this thesis indicates the potential of semi-automated approaches for the development of more trustworthy and usable fact-checking solutions. The impact of the presented work can be further enhanced by addressing its limitations and incorporating new technological advancements, including large language models (LLMs) and explainable artificial intelligence (xAI).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

The issue of problematic information, i.e., an umbrella term referring to inaccurate, misleading, inappropriately attributed, or fully fabricated information (Jack 2017), dates back to the beginning of humanity. History is rife with examples of the propagation of false, misleading, and malicious information, that had negative implications on society. This varies from organised disinformation campaigns to relatively innocent urban legends. One of the earliest known examples of problematic information having a significant impact was the successful propaganda campaign led by Octavian against Antony during Roman times, which paved the way for him to become the first Roman Emperor (Posetti and Matthews 2018). After the invention of the printing press in 1493, the dissemination of problematic information was dramatically amplified with the help of printed materials. This helped the generation of the first widespread hoax, known as *The Great Moon Hoax of 1835*, which consisted of a series of articles published by New York's *The Sun* newspaper on the discovery of life on the moon, featuring drawings of fictitious creatures.

Coming to more recent times, the proliferation of digital technologies, including the Internet, smart devices, and social media, has facilitated and accelerated the transmission of problematic information to a wider audience. This also complicated the solution to the problem and boosted its negative impacts on society. For instance, fake news articles during the 2016 US presidential elections had serious political implications in the US (Allcott and Gentzkow 2017). More recently, many people lost their lives as they refused to get vaccinated due to conspiracy theories circulated during the COVID-19 pandemic (Islam et al. 2020). According to the Global Risks Report published by the World Economic Forum in 2024, the issue of problematic information is expected to become the most severe threat in the next two years, and the fifth most severe threat in the next decade (World Economic Forum, in partnership with Marsh & McLennan Companies and Zurich Insurance Group 2024).

Other than problematic information, the research community has coined a plethora of terms to describe the phenomenon, including information pollution (Orman 1984), information disorder (Wardle and Derakhshan 2017), and infodemic (Rothkopf 2003). However, the diversity of disciplines studying the phenomenon complicated the research community to agree on a unified terminology (Altay et al. 2023), even an umbrella term defining the problem. This also applies to the specific types of problematic information – many concepts have been proposed and used in the literature, including false information, misinformation, disinformation, misleading information, and fake news. Although each of these is different from the others, they are sometimes used interchangeably in the literature (Saunders 2022), revealing the inconsistency and confusion in the terminology. This thesis considers *problematic information* as an umbrella term to indicate any information that is likely to cause harm to society, extending the original definition to cover other kinds of information, such as rumours, hate speech and confidential

information. Nevertheless, it particularly focuses on a common type of problematic information, false information, which refers to inaccurate, untrue, or incorrect information. That is, other types of problematic information, such as misleading information and malinformation, are beyond the scope of this thesis, although the presented findings can be useful to counter them as well.

The efforts to fight against false information can be categorised into two different strategies – detection and prevention. The detection approach attempts to capture false information after being generated to avoid its further propagation. This can be achieved in multiple ways, e.g., by performing content moderation or using fact-checking tools. Nevertheless, the prevention approach aims to take proactive measures in the first place to avoid the generation, or at least, further propagation of false information. Such measures include establishing deterrence via law enforcement, educating people to improve their critical thinking abilities and digital media literacy, forewarning people with general warning messages, and exposing them to false information through *inoculation games* to immunise them against problematic content online with *prebunking* (Lewandowsky and van der Linden 2021). It is worth noting that detection and prevention have a close relationship, e.g., prevention approaches might follow the detection of false claims to avoid their further dissemination, or more accurate detection approaches can help better prevention. Besides, detection and prevention approaches can also be supported by studies that attempt to understand false information and its surrounding ecosystem through conceptual models, theoretical frameworks, and empirical analyses.

False information detection often requires an investigative process, called *fact-checking*, which includes accomplishing different tasks, such as identifying claims, searching for evidence, assessing the accuracy of claims, and providing perspective to claims (Cazalens et al. 2018). Depending on to what extent humans are involved, fact-checking can be designed in three ways – *manual*, *semi-automatic*, and

automatic fact-checking. *Manual fact-checking* refers to determining the veracity of information based on conventional investigation practices, e.g., contacting the claimer, consulting experts, etc. (Graves 2017). It has commonly been followed by journalists, especially before the introduction of digital tools and services. Although *manual fact-checking* can provide promising accuracy, it is unable to provide a solution at scale, considering the amount of false information circulated online (Hassan et al. 2015).

Fact-checking processes can be facilitated and accelerated using digital tools and/or services, accomplishing some or all of the fact-checking tasks. With this respect, *automatic fact-checking* aims to automatically identify false information employing digital technologies, including artificial intelligence (AI) and natural language processing (NLP). Even though the scalability issue in manual approaches, and the advancements in AI and NLP bring automatic fact-checking forward, the current progress of such methods is far from sufficient to solve the problem due to several challenges. To begin with, automatic fact-checking solutions are limited by the accuracy of the AI models employed and the amount of data used for training the models (Hassan et al. 2015; Graves 2018). Additionally, existing datasets are generally imbalanced and insufficient in covering different domains (Zeng, Abumansour and Zubiaga 2021). This also indicates the difficulties around the automatic construction of training datasets, meaning that even fully automated solutions need human annotators for dataset construction. Other technical challenges involve handling multilingual and multimodal settings, ambiguity in claims, system bias, and limited use of contextual information (Nakov et al. 2021). From the psychological perspective, however, fully automated solutions with a black-box design could be ineffective in persuading users, or even produce a "backfire effect" increasing misbelief in them, especially when the developed solution outputs only the veracity of claims without any justification (Ecker et al. 2022). Besides, both fact-checkers and common readers express scepticism and

distrust towards the use of automation and AI in fact-checking processes (Juneja and Mitra 2022; Philipp Schmidt and Teubner 2020).

Due to the limitations of human-only and fully automated systems, hybrid solutions leveraging the strengths of humans and automation emerged. Examples of such solutions include human-in-the-loop systems in which automation benefits from human input and human-machine teaming approaches where humans and automation collaborate to reduce their weaknesses. With this respect, *semi-automatic fact-checking* seeks to effectively join forces of humans and digital automated systems against false information by leveraging the strengths and reducing the weaknesses of both sides. More precisely, it combines human intelligence and expertise with the fast processing ability of digital technologies to provide a more effective solution in the short term. In such fact-checking solutions, humans and digital systems can be teamed up in several ways. For instance, different human actors, such as fact-checkers, experts, and the crowd (e.g., online users or recruited crowdsourcing workers), can be involved in different stages of the fact-checking process. In addition, digital tools can support the human component to accomplish different tasks during fact-checking (Nakov et al. 2021). However, the effectiveness of such different hybrid designs has not been fully explored in the literature (Nakov et al. 2021). Furthermore, existing research has paid far less attention to human aspects (e.g., usability, intelligibility, and trust) when designing such solutions (Das et al. 2023) despite recent findings revealing that hybrid solutions are perceived as more trustworthy (Juneja and Mitra 2022; Hrckova et al. 2022).

1.2 Research Objectives

As a complex phenomenon, false information and fact-checking require comprehensive approaches to increase the effectiveness of the proposed solutions. To this

end, the main aim of this thesis is to explore how better fact-checking solutions can be developed, informed by enhanced conceptual and empirical understanding of the false information and fact-checking ecosystem. To achieve this aim, the following research questions are studied:

- **RQ1:** What are the common concepts that have been studied by the research community in the false information literature?
- **RQ2:** What stakeholders exist in the false information and fact-checking ecosystem, and how are they related?
- **RQ3:** What are the user perspectives on fact-checking, the stakeholders involved, and the fact-checking tools?
- **RQ4:** How the findings obtained from the answers to the previous research questions can be used for the development of a new more trustworthy and effective fact-checking tool?

Answering these research questions is believed to contribute to the efforts to combat false information by taking a step forward to develop a more complete and systematic understanding of the phenomenon and depicting how this can be leveraged to implement more usable (e.g., more effective and more efficient) solutions. The contributions towards RQ1 and RQ2 are expected to provide a better understanding of existing concepts in the literature and the actors involved in the ecosystem, respectively. Then, answering RQ3 is expected to shed light on user perspectives based on empirical evidence. Finally, the answer to RQ4 is hoped to guide researchers studying false information and fact-checking in how to benefit from theoretical and empirical research findings in a technical solution.

1.3 Contributions

This thesis aims to make several contributions to the literature through research conducted towards answering the research questions presented. Specifically, the key contributions of this thesis are as follows:

- Addressing RQ1, a topical analysis of the interdisciplinary literature on false information is presented through topic modelling. The analysis explores commonly studied concepts by the research community and reveals different groups among the identified topics. The findings can guide researchers on potential research gaps and the generated topic models can be used as an initial step to construct an ontology of false information.

- Addressing RQ2, to the best of our knowledge, the most comprehensive conceptual model of the false information and fact-checking ecosystem is introduced for a better understanding of the current landscape. The model is an EER (enhanced entity-relationship) model, which contains actors involved in fact-checking as entities and the relationships between them. Potential applications of the model are exemplified with two use cases – modelling real-world scenarios and reviewing the literature. Furthermore, it can be used to accomplish many other tasks, including developing a computational ontology of the ecosystem to facilitate automated reasoning.

- Addressing RQ3, a user survey was conducted to discover users' attitudes and behaviours towards fact-checking tools. The survey questionnaire is prepared based on a built taxonomy of existing fact-checking tools and approaches. The presented taxonomy is unique in terms of the aspects considered for categorisation, e.g., their level of automation, their contribution to fact-checking, and supported platforms. The survey results offer some design cues to develop more trustworthy and usable fact-checking tools.

- Addressing RQ4, to the best of our knowledge, the first fact-checking tool that is entirely based on expert opinion discovery, namely aedFaCT, is presented. Prioritising trustworthiness, aedFaCT seeks evidence from credible sources, including news articles from mainstream and scientific news outlets with high credibility ratings, and peer-reviewed scientific publications. Moreover, it is designed as a semi-automatic tool for evidence retrieval, leaving the veracity assessment to users. With these features, aedFaCT demonstrates an example of fact-checking tools that can build trust among its users more conveniently.

1.4 Thesis Structure

This thesis contains six more chapters. As shown in Figure 1, Chapters 2 to 6 are closely related and supported by each other. More precisely, the rest of this thesis is organised as follows.

Chapter 2 provides a detailed review of the relevant literature in multiple disciplines. It first provides a conceptual background with proposed conceptual models and the definitions of a number of key relevant concepts. Then, it explains the different stages and types of fact-checking. The chapter concludes with some discussions on leveraging different stakeholders and considering human aspects in fact-checking.

Chapter 3 aims to address RQ1 by analysing the recent literature in terms of the studied topics through two generated topic models. The generated models use the references of relevant survey papers as the training set. Thus, the dataset construction partially benefits from the literature review provided in Chapter 2. Yet, the conducted topical analysis extends Chapter 2 with more insights from the literature.

Inspired by the closely tied topics revealed in Chapter 3, Chapter 4 attempts

to answer RQ2 with an EER model of the false information and fact-checking ecosystem. The entities of the proposed model are based on an initial list of entities derived from existing conceptual models mentioned in Chapter 2. The chapter concludes with a demonstration of how the proposed model can be utilised via two example applications.

In Chapter 5 which aims to address RQ3, user perspectives on fact-checking tools, corresponding to a part of the EER model introduced in Chapter 4, are examined with a survey based on a taxonomy of fact-checking tools. The presented study seeks answers to several research questions regarding trust in different stakeholders, common fact-checking practices, and trust in different fact-checking approaches.

Then, Chapter 6 attempts to address RQ4 with the implementation of aedFaCT, a semi-automated fact-checking tool based on expert opinion discovery. As a tool aiming to accelerate fact-checking practices and prioritising trustworthiness, the design of aedFaCT is informed by the EER model presented in Chapter 4 and supported by the findings of Chapter 5.

Finally, Chapter 7 concludes the thesis with a brief discussion of the summary of contributions, limitations, and future work.

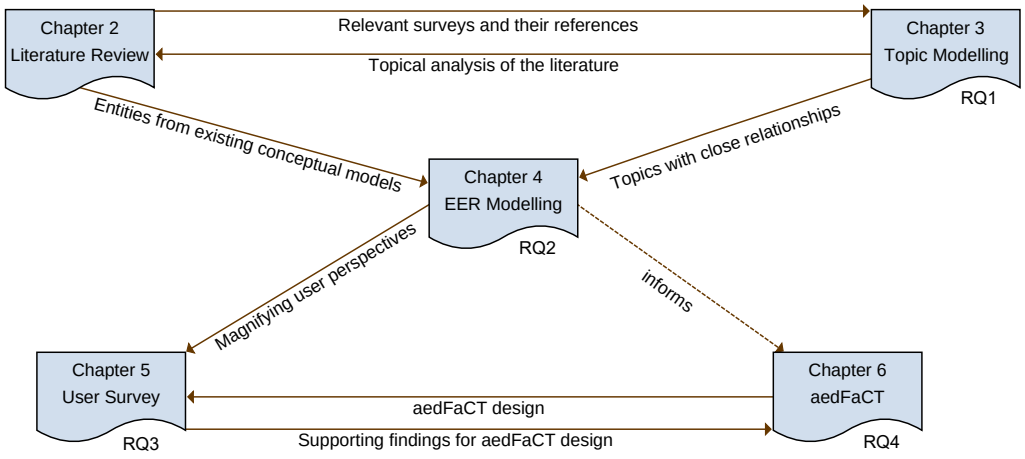


Figure 1: The relationships between different chapters of the thesis

Chapter 2

Literature Review

This chapter overviews the literature from several aspects, including current conceptual understanding of false information and fact-checking, different stages and types of false information detection, and discussion on utilising stakeholders and human aspects in fact-checking. Nonetheless, a more systematic analysis of the literature using a computational approach based on the analysis of many survey papers and their references is presented in the next chapter, i.e., Chapter 3.

2.1 Conceptual Models on Problematic Information

Entity-based conceptual models aim to categorise false information, its different types, and relevant concepts to provide a common understanding of the topic. With this respect, several attempts have been made to define and classify false information and related concepts through conceptual models, such as typologies and taxonomies (Fard and Cunningham 2019). One of the most commonly used conceptual models is Wardle's typology of *information disorder* (Wardle and Derakhshan 2017). It subdivides information disorder into misinformation (i.e., false information without intent to harm), disinformation (i.e., with intent to harm),

and malinformation (i.e., harmful information based on reality). Moreover, the typology classifies false information (i.e., the combination of misinformation and disinformation) into seven categories, which are satire or parody, misleading content, imposter content, fabricated content, false connection, false context, and manipulated context (Wardle 2017). Kumar and Shah (2018) introduced a simple high-level taxonomy of false information, in which they categorised false information into misinformation and disinformation according to intent, and opinion-based and fact-based in terms of knowledge. Zannettou et al. (2019) proposed another categorisation of false information, focusing on the Web ecosystem, which suggested eight types – fabricated, propaganda, conspiracy theories, hoaxes, biased or one-sided, rumours, click-bait, and satire news. They also covered the types of actors that constitute false information propagation and their motives.

Apart from the conceptual models of false information, the literature includes conceptual models of its different types, such as fake news, misinformation, and disinformation. For instance, Rubin, Chen and Conroy (2015) simply categorised fake news into serious fabrications, large-scale hoaxes, and humorous fakes. On the other hand, Tandoc Jr., Lim and Ling (2018) presented a more systematic typology of fake news based on 34 academic articles published between 2003 and 2017. The typology involves six types of fake news – news satire, news parody, fabrication, manipulation, advertising, and propaganda, aligning with the false information typology proposed by Wardle and Derakhshan. As another example, Lemieux and Smith (2018) proposed a disinformation taxonomy leveraging archival theory. While the taxonomy does not contain categorisations of different terms, it describes each concept in terms of the *accuracy*, *reliability*, and *authenticity* attributes. More recently, Molina et al. (2021) identified seven types of fake news, which are false news, polarized content, satire, misreporting, commentary, persuasive information, and citizen journalism.

2.2 Key Concepts

Given the conceptual models on problematic information, this section is intended to discuss some key concepts that are widely used in the literature.

2.2.1 Dimensions of Problematic Information

While problematic information (also known as bad information) originally refers to inaccurate, misleading, inappropriately attributed, or fully fabricated information (Jack 2017), its definition can be extended to cover all kinds of information within the scope of information disorder. Precisely, it can be considered as an umbrella term for any information that is undesired and causes harm to society. By combining different perspectives in the literature, problematic information can be described in three dimensions, which are *veracity*, *misleadingness*, and *harmfulness*:

Veracity corresponds to what extent a piece of information reflects the truth, and problematic information without veracity is generally named as *false information*. Although there is no widely adopted definition of veracity (García Lozano et al. 2020), many researchers have attempted to define it based on different attributes of information. For example, Rubin and Lukoianova (2013) explored veracity across its three main theoretical dimensions – *objectivity*, *truthfulness*, and *credibility*. Similarly, Lovelace et al. (2016) defined veracity as "the degree of truthfulness associated with a data set", with a focus on big data. Besides, Samuel and Zaïane (2018) used *credibility* and veracity interchangeably as referring to the factual accuracy of information. Finally, Wang, Luo and Liu (2015) associated veracity with *trustworthiness*.

The veracity status (i.e., false, true, partially true) of a piece of information could be unknown due to the lack of evidence. Such information with unverified veracity is called **rumour**. Although this definition is widely accepted in the

literature, other opinions exist claiming that rumours also cover false information (Zubiaga et al. 2018; Cao et al. 2018).

As a term more discussed in the philosophy literature, **misleadingness** addresses whether a piece of information is likely to cause false beliefs (Fallis 2015) and information having misleadingness is called *misleading information*. Misleading information differs from false information in the sense that it does not have to be false itself. Fallis (2015) conceptualised this difference by introducing the term "true disinformation" which covers true, but intentionally misleading information. Then, Søe (2017) coined "true misinformation" to extend this conceptualisation to involve true, but unintentionally misleading information. In addition, false information may not always be misleading, e.g., irony (Søe 2021).

Lastly, **harmfulness** means the degree of causing harm, for a piece of information. Although false and misleading information causes harm to society, *harmful information* is not limited to those. With this respect, Wardle and Derakhshan (2017) proposed a term called **malinformation**, covering factual, but harmful information, such as leaked private data and hate speech.

2.2.2 False Information

Aligned with the scope of this thesis, false information deserves more attention. It corresponds to inaccurate, wrong, or untrue information. A more common concept used as a synonym of false information is *misinformation*, which has been used for hundreds of years (Harjule et al. 2023).

Based upon the intelligence operations of the Soviets during the Cold War, a new word translated from Russian came into use, *disinformation*, referring to the strategic dissemination of false reports to mislead public opinion (Froehlich 2020). Following the introduction of disinformation, a widely accepted categorisation of false information has become based on the intent to share it – false information that is unintentionally shared is called *misinformation* while *disinformation* refers

to false information that is deliberately shared to mislead and deceive (Kumar and Shah 2018). Nonetheless, it has been also argued that misinformation and disinformation are types of misleading information, meaning that they are not necessarily false (Søe 2021; Fallis 2015).

More recently, another concept used for false information is *fake news*. It has become widespread during the 2016 US presidential election. Jaster and Lanius defines fake news as "news that lacks truth and truthfulness". Molina et al. (2021), however, used fake news as an umbrella term covering not only false information but also different types of problematic content, such as polarised and persuasive information. Despite its wide usage in the literature, many scholars agree that the term fake news is an inadequate and misleading term to describe the complexity of false information (Wardle and Derakhshan 2017; Kalsnes 2018). Besides, the High-Level Expert Group (HLEG) on fake news and online disinformation set up by the European Commission suggests abandoning the term fake news altogether and using the term *disinformation* instead (High-Level Expert Group (HLEG) on Fake News and Online Disinformation 2018).

2.2.3 Fact-checking

One of the essential steps to combat false information is its detection as soon as possible and debunking to minimise the number of people impacted. Once being an internal quality assurance process in news media, fact-checking has become a growing demand with the increased amount of false information online (Hanselowski 2020). American Press Institute defines fact-checking as "researching the purported facts in published/recorded statements made by politicians and anyone whose words impact others' lives and livelihoods" (Elizabeth 2014). Another definition for fact-checking proposed by Vlachos and Riedel (2014) is "the task of assessing the truthfulness of claims made by public figures". Cazalens et al. (2018) provided a more detailed task-based definition of fact-checking

as "an investigative process consisting of extracting claims from some discourse, searching for the facts the claims are based on, assessing the accuracy of the claim with regards to those backing facts, and providing perspective to claims for which there is no straightforward settlement". As a key concept and the main focus of this thesis, the next section elaborates more details on fact-checking.

2.3 Fact-checking Pipeline and Relevant Tasks

Since fact-checking is a complex phenomenon, a typical fact-checking process requires several tasks. With the advancements in digital technologies, there has been a growing interest in automating these tasks for more effective and efficient fact-checking. While the ultimate goal of this interest is developing highly automated end-to-end fact-checking systems with sufficiently high accuracy and minimum human intervention, a more feasible short-term alternative is to design fact-checking tools that can assist users in different fact-checking tasks. Based on the existing literature on different steps of fact-checking (Graves 2017; Kotonya and Toni 2020; Augenstein 2021; Nakov et al. 2021; Zeng, Abumansour and Zubiaga 2021; Guo, Schlichtkrull and Vlachos 2022; Das et al. 2023), these tasks can be grouped into four stages: 1) claim determination; 2) evidence retrieval; 3) veracity assessment; and 4) explanation generation, as shown in Figure 2.

2.3.1 Claim Determination

Given how frequently new suspicious claims can emerge and how fast they can be circulated online, time is limited for fact-checking such claims before they go viral. Therefore, determining what to fact-check in a given document is crucial as is the initial stage of a standard fact-checking pipeline. This can be achieved in several ways. One proposed approach is distinguishing potentially falsifiable claims,

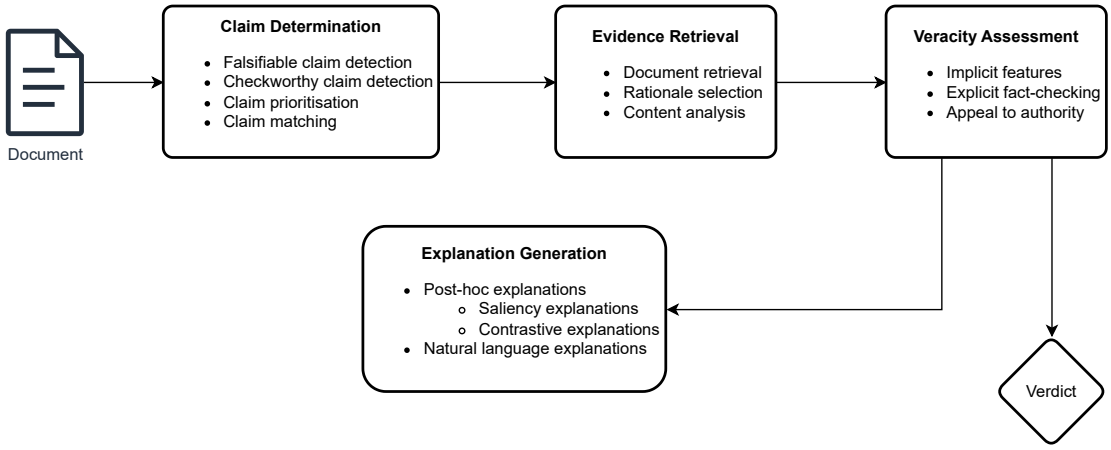


Figure 2: The summary of a typical fact-checking pipeline with relevant tasks

such as quantifiably verifiable statements (Jandaghi and Pujara 2023), from subjective opinions (Konstantinovskiy et al. 2021). Another approach is identifying *check-worthy* claims which might arouse more interest from the general public as to their veracity (Hassan, Li and Tremayne 2015). This approach also enables the prioritisation of more important claims to reduce the time consumed for fact-checking (Kartal and Kutlu 2023). Finally, some identified claims might have been fact-checked previously. To avoid repeated work for such claims, *claim matching* task seeks to identify whether a particular claim exists in databases of previously fact-checked claims and can be addressed by a previous fact-check (Zeng, Abumansour and Zubiaga 2021).

Other than identifying claims explicitly, automatic keyword (or keyphrase) extraction (AKE) methods can be utilised as a less-structured way of claim determination to obtain several representative keywords that can potentially provide insights about the claims mentioned in a given document. The most common way of incorporating AKE methods into fact-checking is to generate search queries with the extracted keywords to seek relevant evidence (Deka, Jurek-Loughrey and Deepak 2022; Sundriyal et al. 2022; Martín et al. 2022).

2.3.2 Evidence Retrieval

Once what to fact-check is determined, the next step is to seek relevant and trustworthy evidence that can help users or fact-checkers reach a verdict about the veracity of the investigated claim. A typical evidence retrieval process in fact-checking consists of two consecutive tasks – *document retrieval* and *rationale selection* (Zeng, Abumansour and Zubiaga 2021). Document retrieval aims to identify relevant documents that are more likely to contain relevant evidence for fact-checking. This task is commonly achieved by utilising search engines such as Google (or any database that provides searching) through their API services or using similarity measures (e.g., cosine similarity, BM25) in information retrieval to identify the relevance of documents to the claim. After identifying the relevant documents to the claim, specific parts (e.g., words, sentences, or paragraphs) from the document that might serve as evidence are selected in the rationale selection task. This task can be succeeded through different approaches, including keyword matching, sentence similarity scoring, and supervised ranking.

Apart from text-based claims, evidence retrieval may need to cover other data modalities, such as images and videos. For example, when the claim is about an image, reverse image search engines (e.g., TinEye[1], Google Image Search) are leveraged to identify other contexts in which the multimedia content was used. This can help detect whether the image was used out of context or was manipulated (Nakov et al. 2021). In addition, multimedia content can be analysed with automated tools such as InVID (Teyssou 2019) to obtain useful evidence, including contextual information, metadata, and copyright information.

[1] `https://tineye.com/`

2.3.3 Veracity Assessment

Based on the collected evidence, veracity assessment aims to reach a verdict regarding the veracity of the claim. While it is common for fully automated fact-checking systems to have an automated veracity assessment based on classifiers, the assessment might be left to users in semi-automated fact-checking systems. When it comes to automated veracity assessment, the simplest approach is binary classification with *true* and *false* class labels (Guo, Schlichtkrull and Vlachos 2022). Alternatively, the classification can be based on the alignment of the retrieved evidence (e.g., *supporting* or *contradicting*) with respect to the claim, which is considered to be a preferable approach due to the lack of assessment of the retrieved evidence (Zeng, Abumansour and Zubiaga 2021; Guo, Schlichtkrull and Vlachos 2022). Nevertheless, binary classification is insufficient to cover different cases, e.g., when information is partially true or there is a lack of evidence to reach a verdict. Therefore, several studies employed multi-class classification with labels corresponding to finer-grained truth values and levels of agreement between the claim and evidence (Guo, Schlichtkrull and Vlachos 2022).

In terms of the indicators adopted for classification in veracity assessment, three major approaches exist – leveraging *implicit features*, e.g., stylometric text features, URL features, or user account features, to distinguish veridical claims from non-veridical ones, performing *explicit fact-checking* by comparing the claim with existing knowledge to check if it matches, and *appealing to authority* with an assumption that the claim is veridical if claimed by an authoritative source. While the first approach is the most commonly used one in the literature, the mentioned approaches are often combined to obtain better results (García Lozano et al. 2020).

2.3.4 Explanation Generation

In real-world applications, providing basic labels such as true or false as the fact-checking outcome is likely to be insufficient to persuade users, especially for black-box automated fact-checking systems considering users' scepticism towards automation (Juneja and Mitra 2022). Therefore, generating human-understandable justification for the verdict reached by a fact-checking system is a promising research direction to increase transparency of fact-checking systems. Besides, such explanations help debug as well as measure bias and fairness (Atanasova 2024). However, many existing fact-checking systems lack of the functionality of providing explanations (Kotonya and Toni 2020).

In the literature, the two most common types of explanations generated in fact-checking solutions are *post-hoc explanations* and *natural language explanations* (Augenstein 2021; Atanasova 2024). The former refers to explanations that can be generated after a fact-checking model has already been trained. The most prominent type of such explanations is *saliency explanations* – highlighted parts of the input based on their importance for the model's verdict. Another post-hoc explanation type is *contrastive explanations*, seeking small changes to the input that can lead to a change in the model's verdict. Nonetheless, natural language explanations are human-understandable explanations provided by the fact-checking model in free text. Such explanations are not limited to the input – the model can generate further explanations on how it reaches a verdict for the input. Unlike post-hoc explanations, natural language explanations require the model to be trained to both perform fact-checking and generate free text explanations.

2.4 Types of Fact-checking

Fact-checking can be performed at different levels of automation. Broadly speaking, fact-checking types include *manual*, *automatic*, and *semi-automatic* fact-checking. The next subsections provide more details on each type of fact-checking with a comparison of them in Table 1.

2.4.1 Manual Fact-checking

The idea of fact-checking originated as an internal quality assurance process within news media (Hanselowski 2020). Before the advancements in digital systems, fact-checking had no other option than being performed manually with conventional methods based on the practices of investigative journalism. However, the inability of conventional journalism to combat false information led to the emergence of several fact-checking initiatives across the world, dedicated to investigating suspicious claims that are likely to go viral (Nieminen and Rapeli 2019). A typical manual fact-checking process includes several steps, including gaining familiarity with the topic, claim identification, evidence aggregation, source credibility check, claim validation, reasoning chain validation, and fallacy checking (Hanselowski 2020). These steps contain many investigative journalism tasks, such as contacting the author of claims, trusted official sources, and relevant experts, cross-checking from multiple sources, and transparently reporting the final verdict about claims (Graves 2017).

Although it can provide high accuracy, the most obvious limitation of manual fact-checking is scalability due to its laborious nature (Hanselowski 2020; Nakov et al. 2021). In addition, inconsistencies in accuracy ratings (Marietta, Barker and Bowser 2015; Lim 2018) and different cognitive biases (Soprano et al. 2024) negatively impact trust in manual fact-checking processes. Finally, fact-checkers following a manual approach need to have a certain level of expertise, especially

to accomplish specific tasks such as providing explanations (Kotonya and Toni 2020).

2.4.2 Automatic Fact-checking

Despite various fact-checking initiatives, it is insufficient to resolve the scalability issue in the presence of a growing number of false claims circulating online. This led to a growing interest in fully automated fact-checking solutions leveraging the advances in technologies, such as artificial intelligence (AI), natural language processing (NLP), information retrieval (IR), and data mining. Other than being fully automated, such solutions are also ideally expected to be instant, accurate, and accountable (Hassan et al. 2015). Automatic approaches have been designed in different ways, including utilising knowledge bases, training machine learning models on large datasets, and constructing a pipeline with several automated components corresponding to different stages of fact-checking (Hanselowski 2020). Apart from such content-based approaches, user networks in social media have been utilised in network-based approaches, based on research findings indicating that users spreading false information are likely to be involved in the same networks (Augenstein 2021).

Unlike manual approaches, automatic fact-checking offers high scalability and requires no expertise of the user. Furthermore, it can achieve high accuracy in certain tasks. Nevertheless, the current status of fully automated solutions indicates many challenges for the research community to overcome for the practical use of such solutions. From the computer science perspective, fully automated solutions are limited by the accuracy of underlying AI models and the amount of data used for training the models (Hassan et al. 2015; Graves 2018; Nakov et al. 2021). In addition, existing datasets are commonly imbalanced and unable to cover different domains sufficiently (Zeng, Abumansour and Zubiaga 2021). This might lead AI-based systems to have system bias reducing their performance,

similar to fact-checkers' perceived bias in manual fact-checking. Other than technical challenges, both professionals, e.g., fact-checkers, and the general public are sceptical of using fully automated fact-checking solutions due to their black-box nature (Philipp Schmidt and Teubner 2020; Juneja and Mitra 2022). Psychologically speaking, such solutions may not build trust among users, or may even have a "backfire effect" making users more sceptical of the system, particularly when no explanations are provided (Ecker et al. 2022).

2.4.3 Semi-automatic Fact-checking

The limitations of manual and fully automated fact-checking paved the way for hybrid solutions benefiting from the capabilities of human intelligence and the scalability provided by automated approaches. Such hybrid approaches can be designed in several ways, such as using crowdsourcing in certain tasks, active learning, interactive machine learning, and decision support systems where the final verdict is reached by users (Das et al. 2023). Besides, semi-automatic approaches can involve automated tools supporting fact-checkers in certain tasks during fact-checking (Nakov et al. 2021).

Similar to automatic fact-checking, semi-automatic fact-checking can offer higher scalability and require no (or limited) expertise. Moreover, its human component can be leveraged to improve the accuracy of the automated part, e.g., correcting the automation output. However, semi-automatic approaches also have some limitations in the current landscape. Despite a wide range of designs of semi-automatic approaches, they are not fully explored by the research community (Nakov et al. 2021). Thus, more research is needed to validate their effectiveness. Moreover, the inclusion of human component(s) might still introduce cognitive biases although their effects can be limited by system design (Soprano et al. 2024). Nevertheless, recent research findings suggest that such hybrid solutions are perceived as more trustworthy by fact-checkers (Juneja and Mitra 2022;

Table 1: The comparison of the three types of fact-checking

Type	**Pros**	**Cons**
Manual	High overall accuracy	Low scalability
		Inconsistency in accuracy ratings
		Cognitive biases
		Expertise required
Automatic	High scalability	Limited overall accuracy
	No expertise required	Low domain adaptability
	High accuracy in specific tasks	Scepticism towards automation
		System bias
Semi-automatic	Medium scalability	Designs not fully explored
	Improved automation accuracy	Validation of effectiveness needed
	Higher trustworthiness	Cognitive biases
	Less or no expertise required	

Hrckova et al. 2022).

2.5 Stakeholder-based Fact-checking

As a complex phenomenon, a typical fact-checking process involves many stakeholders that play different roles in the process. Neumann, De-Arteaga and Fazelpour (2022) categorised these stakeholders into four major groups – *seekers of information*, *sources of information*, *subjects of information*, and *sources of evidence*. Based on this categorisation, **stakeholder-based fact-checking** refers to fact-checking approaches that rely on sources of information, subjects of information, and sources of evidence in order to help protect seekers of information from false information.

Sources of information correspond to those making claims, such as the author

of an article or a politician asserting a claim. As an investigative journalism technique and for the sake of fairness, engaging with sources of information can provide useful evidence for fact-checking, e.g., the primary source of the data mentioned in the claim or the context to which the claim belongs (Graves 2017). Another group of stakeholders that can potentially help improve fact-checking is subjects of information, i.e., individuals or organisations mentioned in claims. They are among the frequently considered inputs to automated fact-checking approaches which require subject-predicate-object triplets (Thorne and Vlachos 2018). Furthermore, subjects can be automatically extracted from claims by utilising NLP techniques, although this task is not always straightforward due to ambiguities (Hidey et al. 2020) and ethical concerns (Neumann, De-Arteaga and Fazelpour 2022).

In the literature, the most common usage of stakeholders in fact-checking is engaging with sources of evidence, especially for the evidence retrieval stage. This covers a wide range of stakeholders, including domain experts, crowdsourcing workers, and professional fact-checkers. For example, expert-based approaches leverage expert comments, research papers, expert data annotations, and feedback from domain experts for different stages of fact-checking (Bandhakavi, Hoffmann and Lear 2022; He, Hu and Pei 2023). As another example, crowdsourcing-based approaches benefit from "the wisdom of crowds" (Surowiecki 2005) phenomenon by aggregating the judgements of crowds of laypeople to effectively identify false information online at scale (Allen et al. 2021; Martel et al. 2023). Crowdsourcing has also been employed by social media platforms, such as X[2] (formerly Twitter) and Facebook[3], for content moderation. Finally, there exist hybrid approaches combining the efforts of multiple stakeholders (Botambu Collins and Hwang 2021; La Barbera, Roitero and Mizzaro 2022; Lampou and Antonopoulos 2023).

[2] `https://communitynotes.twitter.com/guide/en/about/introduction`
[3] `https://www.facebook.com/formedia/tools/crowdtangle`

2.6 Key Human Aspects in Fact-checking

Fact-checking is a complex socio-technical phenomenon involving multiple stakeholders, rather than a purely technical task which can be achieved by developing fully automated solutions without considering their applicability in the real world (Juneja and Mitra 2022). Therefore, fact-checking solutions need to take into account different human aspects although existing solutions insufficiently concern such aspects (Das et al. 2023). While they overlap and influence each other in many cases, some of the key aspects commonly discussed in the literature are presented in this section.

2.6.1 Trust

As well as how accurate it is, a fact-checking system is required to be trustworthy to be effective in fighting against false information. With this respect, unlike most existing approaches, it is crucial that fact-checking systems only retrieve evidence from credible sources instead of collecting any information from the largest available sources, e.g., Wikipedia or web search engines (Guo, Schlichtkrull and Vlachos 2022).

Apart from source credibility, another bottleneck for trust in fact-checking is users' trust in different stakeholders. Existing research provides evidence that users are sceptical of automation (Juneja and Mitra 2022; Das et al. 2023) and fact-checkers (Petter Bae Brandtzaeg and Ángeles Chaparro Domínguez 2018) when employed individually in the fact-checking process. Therefore, human-machine teaming approaches in which automation and humans balance and check each other can be a promising research direction in terms of developing more trustworthy solutions.

2.6.2 Usability

Usability is an important human factor to ensure users are provided with maximum benefit during their fact-checking experience. In the context of fact-checking, developing solutions that require minimal technical knowledge and minimising cognitive load can be quite helpful in reaching more users (Nakov et al. 2021). This can be achieved with increased system interactivity since interactive fact-checking systems attract more attention from users without affecting their cognitive load (Shi et al. 2022). Besides, content and visualisation are among the key factors of a usable solution. Within this direction, the output presented by fact-checking systems needs to be concise (Lim and Perrault 2023).

2.6.3 Explainability

As mentioned in Section 2.3.4, users must be convinced by the output, e.g., the assessed veracity or the retrieved evidence, of the fact-checking system for an efficient solution. This requires fact-checking systems to provide users with clear explanations of how evidence is retrieved and the verdict is reached, for the sake of transparency. In addition to clarity, explanations provided by fact-checking systems are crucial to be unbiased and accountable to ethical considerations (Nakov et al. 2021), as well as to avoid overtrusting such systems which might occasionally produce inaccurate responses (Mohseni et al. 2021). While fact-checking systems can employ transparent models which generate such explanations by themselves, post-hoc algorithms can be developed to generate explanations for the outcome of fact-checking (Shih-Yi Chien and Yu 2022). Potential benefits of considering explainability in fact-checking systems include improved understandability and trust (Kandul et al. 2023). With the advancements in complex deep learning architectures, there is a growing interest in the research community of explainable AI (XAI) systems to make the inner workings of AI models more understandable and to get insights into the outcomes of the models (Dwivedi et al. 2023) although

there are a limited number of automated fact-checking systems with an explainability component (Kotonya and Toni 2020). Nonetheless, explanations revealing the inner workings of AI models might pose a risk for malicious actors to attempt to manipulate the system with adversarial inputs (Lim and Perrault 2023). Therefore, for automated fact-checking systems, example-based explanations or natural language explanations might be preferable compared to feature-based explanations.

Chapter 3

Topical Analysis of False Information Literature

Considering the scope and impacts of the false information problem, researchers from a variety of disciplines, including computer science, psychology, sociology, philosophy, economics, and political science, have been attempting to develop a better understanding of the problem and seeking more effective solutions in accordance with their respective disciplines. While an economist may examine the economics of disseminating conspiracy theories, a psychologist may investigate the motivation behind spreading conspiracy theories, and a computer scientist may seek a method to detect fake profiles. Despite such disparities in viewpoint, the problem at hand is the same. As a result, researchers from all relevant fields need to collaborate and combine their efforts to combat this global and multifaceted problem. The usage of diverse terminology across fields, however, complicates such collaboration. This reality necessitates the development of a common conceptual basis shared by stakeholders of the false information ecosystem from different research backgrounds. As mentioned in Section 2.1, the research community has striven to develop conceptual models that can provide a common understanding of the phenomena. While one of the ultimate goals for having such a basis

can be developing an exhaustive ontology of false information involving the relevant terms and the hierarchies between them, this is not straightforward due to the complexity of the problem and the lack of a consensus on the definitions of many concepts. Therefore, a better understanding of the concepts relevant to false information needs to be developed through different means of modelling and analysis.

With this respect, this chapter presents a data-driven topical analysis of the relevant literature to explore the concepts that have been studied by the research community. The identified topics can serve as a basis for the construction of a structured conceptual model of false information, e.g., a taxonomy or an ontology. To ensure reproducibility and objectivity, a well-defined methodology is followed, which considers the references of survey papers (i.e., papers that aim to review the literature) from various disciplines published between 2009-2020, including peer-reviewed research articles, preprints, and technical reports.

3.1 Related Work

In the literature, topical analyses have been leveraged to achieve different tasks regarding false information analysis, detection, and mitigation. Although using topic modelling approaches such as Latent Dirichlet allocation (LDA) and Latent semantic analysis (LSA) to perform such analyses has been quite common, other automated methods and systematic approaches have also been utilised by the research community. Topical analyses in the literature have covered different objectives, including developing conceptual understanding, analysing how false information is propagated, and detecting false information.

3.1.1 Developing Conceptual Understanding

One of the objectives of topical analysis has been obtaining relevant topics and concepts from false information datasets or relevant literature for developing a better conceptual understanding of the phenomenon. For example, Chong and Choy (2020) built a data-driven taxonomy of fake news by applying the k-means clustering algorithm to a reference news dataset. The algorithm finally generated eight clusters, which were then analysed to determine distinct types of fake news describing each cluster. More recently, Song et al. (2021) leveraged their own classification-aware neural topic modelling approach based on BERT and Variational Autoencoders to categorise COVID-19 disinformation to help identify the most harmful types of disinformation and guide policymakers. As another example, Sharma and Garg (2021) used LDA-generated categories in their Indian fake news dataset to categorise news statements covered. Finally, Kapantai et al. (2021) systematically reviewed existing taxonomies of disinformation from the literature and identified 39 disinformation types. Based on this, they presented a unified taxonomy which covers 11 distinct types of disinformation described by three dimensions – *motive*, *facticity*, and *verifiability*.

The topical analysis presented in this chapter similarly aims to develop a better conceptual understanding of false information. Nevertheless, compared to the previous studies, it offers a more comprehensive literature review, covering not only existing taxonomies but also a wider coverage of the literature. And, its focus goes beyond the classification of false information types to also consider the concepts related to different aspects of the phenomenon, such as contexts where false information can be observed and methods to combat false information.

3.1.2 Analysing False Information Propagation

Another relevant task achieved by topical analysis was analysing existing false information to gain insights into how it is propagated. Ceron, de Lima-Santos and

Quiles (2021) employed time-series analysis, topic modelling, and trend analysis to analyse data posted on Twitter by two Brazilian fact-checking organisations in 2020 during the COVID-19 pandemic. Their analysis involved topic clustering as well as exploring their current evolution in a predefined time window. In another study, Mohammadi et al. (2022) conducted a thematic analysis on 127 false COVID-19 news published in 2020, collected from 8 fact-checking websites. Consequently, they identified 4 main themes and 19 unique subthemes regarding COVID-19 misinformation. More recently, Yousuf (2023) analysed 1,530 fact-checking stories published by four mainstream news outlets in the US by utilising topic modelling, and then, compared the generated topics with the top public priority policy items between 2017 and 2019. They found that the analysed fact-check coverage mainly considered the US presidency and popular policy topics.

3.1.3 Detecting False Information

Ultimately, topical analysis has been used to obtain semantic and contextual information during fact-checking. For example, Atanasova et al. (2019) leveraged topic modelling to obtain a number of topics from political speeches and debates, which were then used as a feature for check-worthy claim detection. Casillo et al. (2021) followed a similar approach by using LDA for topic modelling as a part of the feature extraction process of detecting false information. More recently, Mohawesh et al. (2023) proposed a semantic graph-based topic modelling framework to extract structural and semantic representations of texts for multilingual fake news detection.

3.2 Methodology

3.2.1 Overall Methodology

The topical analysis was conducted on full-text versions of research papers and relevant documents. As depicted in Figure 3, survey papers, which correspond to any publication aimed at reviewing the literature on a relevant topic, such as research articles, technical reports, and books, were leveraged as seeds for constructing the actual dataset. The decision to use survey papers was made as they tend to relate to more relevant studies than any other type of publication, enabling to consider more relevant studies while analysing fewer publications. Then, the actual dataset used for topic modelling was constructed from the references of the collected survey papers. Eventually, a number of topics were identified with topic modelling. More details are presented in the following subsections.

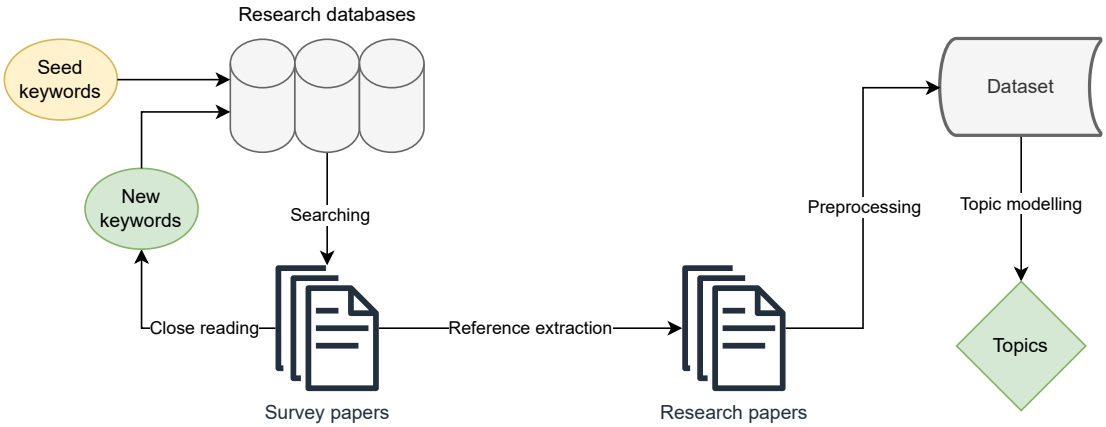


Figure 3: The overall methodology of the performed topical analysis

3.2.2 Collecting Survey Papers

Relevant survey papers were collected in four main steps. These steps were followed iteratively until a stable state was reached, i.e., no new keywords were discovered.

Step 1: Identifying keywords. The key terms used to look for survey papers were determined. However, to collect only relevant survey papers from the literature, some keywords (e.g., review, overview) were also employed to help filter the results and acquire only the survey papers.

Step 2: Searching for papers. The obtained list of keywords was used to form a search query to search for relevant survey papers in the literature. This search query was utilised to perform searches on multiple research databases. Besides, it was used to perform a web search to be able to find survey-like papers such as technical reports, that may not be covered by research databases.

Step 3: Analysing papers. This step included a close reading of the papers found in the previous step. It also covers survey papers fetched from the backward and forward citations of the papers identified before.

Step 4: Identifying new keywords. Analysing the collected survey papers resulted in the discovery of new keywords. This necessitates repeating the search procedure in Step 1 to identify any uncovered survey papers and expand the constructed dataset. As a result, if at least one new keyword was discovered in this step, the procedure was restarted from Step 1 with the updated keyword list.

In the first two steps, the snowballing technique with a number of seed keywords was utilised to search for relevant survey papers. In this manner, the terms "false information", "misinformation", "disinformation" and "fake news" were selected as the seed keywords to be used in the search query while searching for a number of relevant papers on research databases. The obtained papers were then analysed based on their content as well as backward and forward citations to extend the initial list of keywords. This process was repeated iteratively until no new keyword was added to the resulting search query. Besides, a similar iterative process was carried out to determine keywords that can be used to limit the

search results to cover only survey papers. For this part, the search process was initialised with the keyword "survey", and then, similar keywords such as "review" and "overview" were appended. During the search process, different spellings of keywords (e.g., rumour/rumor, click-bait/clickbait) were considered by making use of each research database's search syntax rules.

As a result, the following search query was obtained:

```
(trustworthiness OR believability OR credibility OR veracity OR
"fact-checking" OR "truth discovery" OR "information quality" OR
rumour OR hoax OR fraud OR "fake news" OR "false news" OR satire OR
propaganda OR clickbait OR "junk news" OR fabrication OR
"conspiracy theory" OR "fake review" OR parody OR "opinion spam" OR
"false information" OR misinformation OR disinformation OR
deception OR "misleading information" OR "malinformation" OR
"inaccurate information" OR "information disorder" OR "fake
account" OR "video forgery" OR "image tampering" OR "video
tampering" OR deepfake OR "social bot" OR "fake profile" OR "spam
profile")
AND
(survey OR review OR overview OR "current state" OR revisited OR
"meta-analysis")
```

Since the first part of the given query is quite long considering the maximum input size limitations of different search engines, it was split into smaller chunks during the search process. The searches were performed mainly in Scopus to avoid bias in the resulting publications in terms of their research area. However, other research databases, including Google Scholar, ACM Digital Library, IEEE Xplore, and ProQuest were taken into account so as not to overlook any relevant study. In addition, Google Web Search was utilised to cover survey-like papers, such as

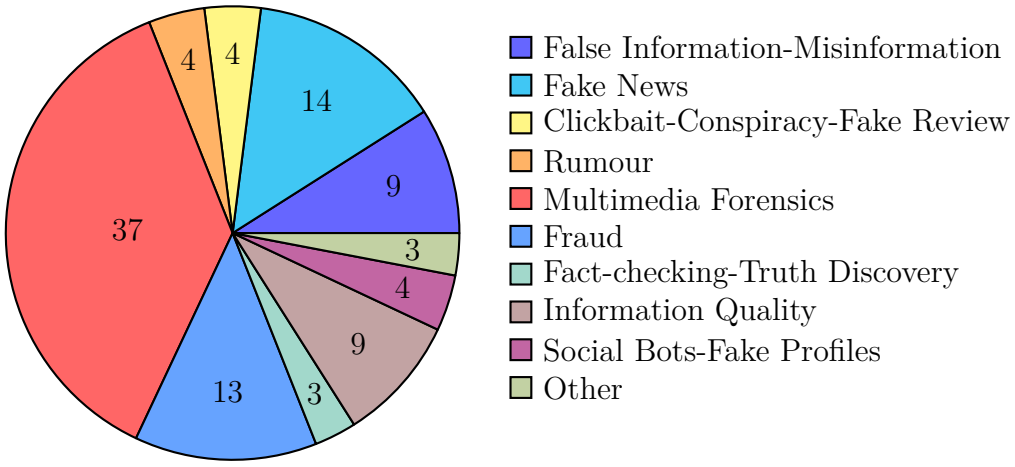


Figure 4: The distribution of the topic of the surveys collected.

technical reports, blog posts from credible sources (e.g., organisations, well-known experts), and magazine articles. Although papers obtained from this search might not be peer-reviewed, they were still included based on the observation that some well-known conceptual models were based on papers that were not peer-reviewed (e.g., (Wardle and Derakhshan 2017)).

When selecting the papers for consideration, an inclusive strategy has been followed. Considering that false information is a broad and well-researched topic, only the papers published (or became available online) between 2009 and 2020 (until 1 June) were included. Nevertheless, the papers whose incremental work (i.e., a newer version written by the same authors) had already been included and non-English papers were excluded. In consequence, 100 survey papers were identified in total, which are presented in Table 2 with their topics. As shown in Figure 4, over one-third of them were about multimedia forensics, e.g., image/video forgery and deepfakes. Then, the most frequently observed topics were fake news, fraud, false information (including mis/disinformation), and information quality, respectively.

Table 2: List of survey papers collected and their topics

Topic(s)	Papers
False Information, Mis/Disinformation	Søe (2017); Webb et al. (2016); Kumar and Shah (2018); Wardle and Derakhshan (2017); Shu et al. (2020); Lewandowsky et al. (2012); Karlova and Lee (2011); Tucker et al. (2018); Zannettou et al. (2019)
Fake News	Zhang and Ghurbani (2020); Conroy, Rubin and Chen (2015); Sharma et al. (2019); Rubin, Chen and Conroy (2015); Tandoc Jr., Lim and Ling (2018); Zhou and Zafarani (2020); Shu et al. (2017); Pierri and Ceri (2019); Saquete et al. (2020); Alemanno (2018); Cooke (2017); Blaber et al. (2019); Figueira and Oliveira (2017); Lazer et al. (2018)
Multimedia Forensics	Qureshi and Deriche (2015); Mahdian and Saic (2010); Johnston and Elyan (2019); Sharma and Dhavale (2016); Mizher et al. (2017); Parashar and Tiwari (2015); Mahmood et al. (2015); Wang, Dong and Tan (2009); Huynh et al. (2015); Zheng, Zhang and Thing (2019); Warbhe, Dharaskar and Thakare (2016); Khodabakhsh, Busch and Ramachandra (2018); Barni, Stamm and Tondi (2018); Piva (2013); Milani et al. (2012); Singhal and Gandhani (2015); Tolosana et al. (2020); Redi, Taktak and Dugelay (2011); Birajdar and Mankar (2013); Korus (2017); Zakariah, Khan and Malik (2018); Sitara and Mehtre (2016); Poisel and Tjoa (2011); Kaur and Jindal (2020); Farid (2009); Stamm, Wu and Liu (2013); Verdoliva (2020); Lian and Zhang (2010); Triphati, Ahad and Haq (2020); Hashem and Sulong (2015); Al-Qershi and Khoo (2013); Abdul Wahab et al. (2014); Tao, Jia and You (2017); Thajeel and Sulong (2013); Qazi et al. (2013); Kingra, Aggarwal and Singh (2016); Rocha et al. (2011)
Clickbait, Conspiracy Theory, Fake Review	Heydari et al. (2015); Douglas, Sutton and Cichocka (2017); Douglas et al. (2019); Chen, Conroy and Rubin (2015)
Rumour	Cao et al. (2018); Zubiaga et al. (2018); Ahsan, Kumari and Sharma (2019); Shelke and Attar (2019)
Fact-checking, Truth Discovery	Li et al. (2016); Thorne and Vlachos (2018); García Lozano et al. (2020)
Information Quality	Ghasemaghaei and Hassanein (2016); Jamil et al. (2015); Ginsca, Popescu and Lupu (2015); Viviani and Pasi (2017); Pasi and Viviani (2020); Colepicolo (2015); Adams (2010); Reuter, Kaufhold and Steinfort (2017); Jones and Moncur (2018)
Fraud	Wang (2010); Phua et al. (2012); Trompeter et al. (2012); Griffiths (2010); Amiram et al. (2018); Vivian and Hussain (2014); Free (2015); Putniņš (2012); Sabau (2012); Ngai et al. (2011); Dorminey et al. (2012); Joudaki et al. (2015); Pirrong (2017)
Social Bot, Fake Profile	Krithiga and Ilavarasan (2019); Alothali et al. (2018); Ramalingam and Chinnaiah (2018); Ferrara et al. (2016)
Other	Vartapetiance and Gillam (2014); Fanelli (2009); Kucuk and Can (2020)

3.2.3 Dataset Construction

Research items cited in the collected surveys were considered to construct the actual dataset used for the topical analysis. To begin with, references from the collected survey papers were extracted by leveraging a machine learning-based

tool, CERMINE Tkaczyk et al. (2015), which is capable of extracting structured metadata and content from a given research paper in PDF format. Based on the output provided by CERMINE, an SQL database was set up for the extracted research items to make the further steps easier and more systematic. Running CERMINE on each of the survey papers yielded 16,206 records for the database.

Nonetheless, it is unsurprising that there were a number of papers cited by multiple survey papers, meaning that the constructed database included several duplicates. Thus, a two-stage deduplication was applied to the records. The first step was to eliminate the obvious duplicates with the exact same title, author, and DOI number by applying a simple SQL query. This decreased the number of records to 7,400. In the second stage, the duplicates corresponding to the same research item with different data (e.g., containing fields with different spelling or differences due to different citation styles) were attempted to be removed. With this respect, a similarity table storing the Levenshtein similarity ratio of the *title* fields of each record was constructed. Then, the pairs having a Levenshtein similarity ratio higher than 0.8 were selected to reduce the number of records in the similarity table to a feasible level for manual investigation. Out of 3,571 records considered potential duplicates, 1,647 of them have been identified as duplicates. As a result, the number of records was reduced to 5,753 after deduplication steps.

The next step was downloading the full-text versions of as many of the research items in the database as possible to construct a dataset with a maximum size. For this purpose, the APIs provided by two major DOI registration agencies, CrossRef [1] and DataCite [2], were utilised to obtain the DOI link of the research items by searching their titles. To check whether any of the search results shown as the most relevant ones correspond to the searched research item, the Levenshtein similarity ratio of the titles of both records was ensured to be higher than 0.8.

[1] https://www.crossref.org/education/retrieve-metadata/rest-api/
[2] https://support.datacite.org/docs/api

Besides, the titles of the research items were searched on the DBLP database [3] and the DuckDuckGo [4] search engine to obtain the URLs of research items, especially those without a DOI number. For the search results listed as the most relevant by these services, the same Levenshtein similarity ratio criterion was applied to confirm the obtained result was the one being looked for. In consequence, 81% (4,676 out of 5,753) of the URLs were fetched. Note that any publisher APIs, e.g., ACM Digital Library, IEEE Xplore, were not used to search for URLs to avoid a biased dataset towards specific disciplines focused by the publishers, considering the existence of research items from different disciplines.

Having the URLs of the papers did not guarantee being able to download the research items since many of them were behind the paywall. Despite the existence of the author's institutional access for most of them, it was still an issue to automate the downloading process because of the different web designs of publishers, complicating a unified web scraping approach. Thus, Selenium WebDriver was used to handle the downloading process in a semi-automated manner, which enables simulating the manual downloading task performed through web browsers. While using Selenium WebDriver, necessary considerations were taken into account regarding the appropriate use of web scraping, e.g., adding delays between requests (Clark 2023).

For locating the download links on the publisher websites corresponding to the research items, a rule-based approach was followed. The rules were determined based on inspections of the source codes of different publisher websites and updated when an unhandled website was identified. The rules included looking for an *href* attribute that contains the string "pdf" or "epdf", specific JSON fields such as "linkToPdf" and "pdfPath", and HTML metatags, including "citation_pdf_url". Once the download links were located by applying the rules mentioned, full-text versions of the research items in PDF format were downloaded automatically.

[3] https://dblp.org/
[4] https://duckduckgo.com/

Besides, the downloading process was monitored for manual intervention when necessary, e.g., avoiding the wrong documents being downloaded and handling CAPTCHAs. As a result of this process, 3,792 of the research items were downloaded and properly processed, which also corresponds to the size of the dataset that was used to perform the topical analysis. Figure 5 shows the steps for dataset construction and the number of records remaining at the end of each step.

3.2.4 Topic Modelling

As the final step, the downloaded full-text versions of the research items were analysed to identify a number of the most frequently appearing topics. With this respect, a popular unsupervised topic model, LDA, was leveraged to perform the topical analysis.

Since the downloaded full texts were in PDF format, preprocessing was applied to prepare the data for topic modelling. Firstly, the PDF files were processed with CERMINE to obtain their XML-based structured content, and the useless parts of the texts (e.g., references and author information) were removed. Then, stopwords, non-alphabetic tokens, punctuation marks, and short tokens with less than three characters were filtered out. Furthermore, PoS-tagging was employed for eliminating tokens with specific PoS-tags, and all the tokens other than nouns, adjectives, and verbs were also filtered out. Finally, tokens exist in less than five documents or more than 90% of the documents were removed to eliminate too rare and too frequent terms. Other than general preprocessing practices, a manually derived list of words was blacklisted, which contains commonly used words in research papers such as “research”, “article”, “figure”, and “table”. This list was expanded iteratively after multiple runs of topic modelling on the constructed dataset. After the filtering process was completed, the resulting set of tokens was lemmatised and converted to lowercase.

Before running the LDA model on the constructed dataset, the k parameter,

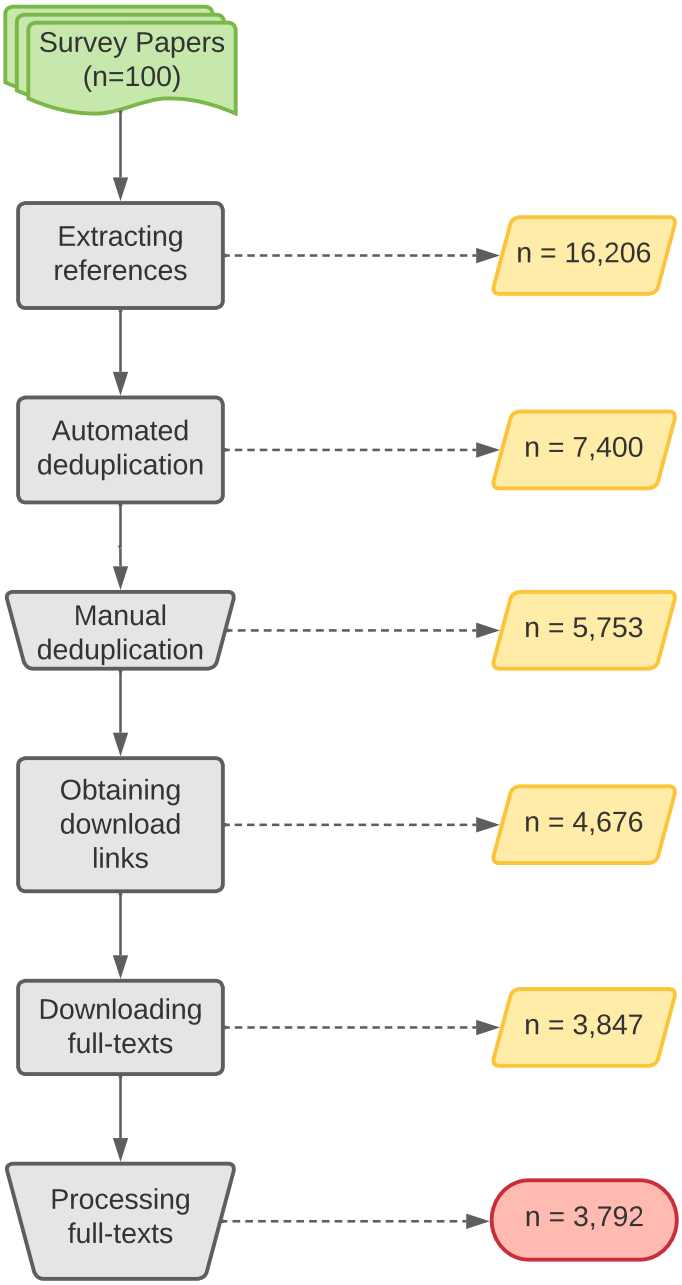


Figure 5: The steps for dataset construction and the number of records after each step

referring to the number of topics to be identified by the model, needed to be specified. Therefore, multiple coherence metrics, including CV, NPMI, and UMass, were utilised to determine the candidates for the optimum k value. Each coherence metric was computed on topic models constructed from increasing values of k from 5 to 50 topics. Based on the three plotted coherence value graphs computed with the three mentioned metrics, the optimum k value was determined by applying the elbow method to each generated graph to identify the best three candidates and then selecting the optimum value with a manual assessment of the generated topics. As a consequence of this process, the optimum k value has turned out to be 13. For the implementation of all the coherence metrics and the LDA model, the Gensim Python library (Řehůřek and Sojka 2010) was utilised. Besides, the pyLDAvis[5] Python library implementing LDAvis, an R-based tool for interactive visualisation of LDA models (Sievert and Shirley 2014), was utilised for further analysis of the generated LDA models.

Although too frequent terms were tried to be eliminated, there might still be some common terms appearing among the most relevant terms for a topic. This can make more topic-specific terms less appearing as highly ranked by the LDA model, complicating the topic identification for the most relevant terms. With this respect, the relevance parameter, λ (lambda), provided by LDAvis was utilised to rerank the most relevant terms such that topic-specific words are prioritised. Following the findings of the user study conducted by Sievert and Shirley (2014) to determine the optimum value for lambda, the value of lambda was adjusted to 0.6.

[5] `https://github.com/bmabey/pyLDAvis`

3.3 Results

The implemented LDA model generated 13 distinct sets of terms, each corresponding to a topic. The list of generated topics with their sizes and the most relevant ten terms for each topic is shown in Table 3. The topics include different applications in which false information can exist (e.g., healthcare, e-commerce, politics, and finance), specific types of false information (e.g., fake news, fake review, conspiracy theory, and forged multimedia), and different methods to combat false information (e.g., AI-based methods and graph-based methods). In terms of the sizes of the topics, healthcare applications constituted the largest portion with a size of 13.6% of the tokens in the corpus. However, terms for the topics relevant to multimedia forensics, i.e., image and video forensics, collectively corresponded to 19.3% of the tokens, which was quite unsurprising due to the dominance of multimedia forensics papers in the dataset.

For further analysis of the generated topics, LDAvis was leveraged to generate the intertopic distance map computed with Jensen-Shannon divergence and scaled with Principal Components, as well as the top 30 words with the highest saliency, i.e., a measure to rank terms considering term frequency and distinctiveness (Chuang, Manning and Heer 2012), as shown in Figure 6. According to the intertopic distance map, some circles are overlapping, indicating closely tied topics. For instance, the closeness of the circles for "Image Forensics", "Video Forensics", "AI-based Methods", and "Graph-based Methods" implies that multimedia forensics research highly benefits from AI-based and graph-based methods. Besides, the circles for Conspiracy Theory and Political Misinformation, as well as the circles for "Fraud" and "Market Manipulation" are overlapping, consistently to the common sense as they frequently share similar contexts – politics and finance, respectively. Finally, trust-related topics referring to different contexts, i.e., "Credibility/Deception", "Online Consumer Trust", and "Healthcare Applications", were closely tied. When it comes to the most salient terms, the

Table 3: Topics identified in the collected survey papers

Topic(s)	**Size (%)**	**Most Relevant Terms**
Healthcare Applications	13.6	health, technology, practice, knowledge, organization, patient, development, project, field, medical
Image Forensics	11.2	image, feature, region, detection, block, pixel, forgery, detect, algorithm, color
Online Consumer Trust	9.9	web, site, quality, customer, service, online, consumer, website, product, trust
AI-based Methods	9.7	datum, model, training, layer, input, learn, network, train, neural, dataset
Video Forensics	8.1	video, frame, signal, watermark, camera, forensic, bit, compression, fingerprint, attack
Fake News/Rumour	8.0	user, tweet, social, twitter, news, rumor, medium, fake, post, content
Graph-based Methods	7.2	node, graph, network, source, probability, edge, algorithm, time, tree, model
Fake Review Detection	7.2	review, feature, stance, text, sentence, document, sentiment, spam, topic, classification
Credibility/Deception	5.7	participant, deception, condition, cue, subject, statement, credibility, truth, source, effect
Fraud	5.6	firm, fraud, financial, earning, accounting, restatement, audit, auditor, ceo, variable
Political Misinformation	5.5	political, news, party, partisan, medium, respondent, blog, public, election, exposure
Market Manipulation	4.2	price, market, trading, trade, stock, law, plaintiff, security, settlement, court
Conspiracy Theory	4.2	conspiracy, belief, theory, vaccine, people, conspiratorial, thinking, group, believe, individual

figure shows terms whose saliency measure was more affected by their high frequency, e.g., “image”, “user”, “feature”, and “model”, as well as those with a higher distinctiveness, such as “conspiracy”, “fraud”, “tweet”, and “belief”.

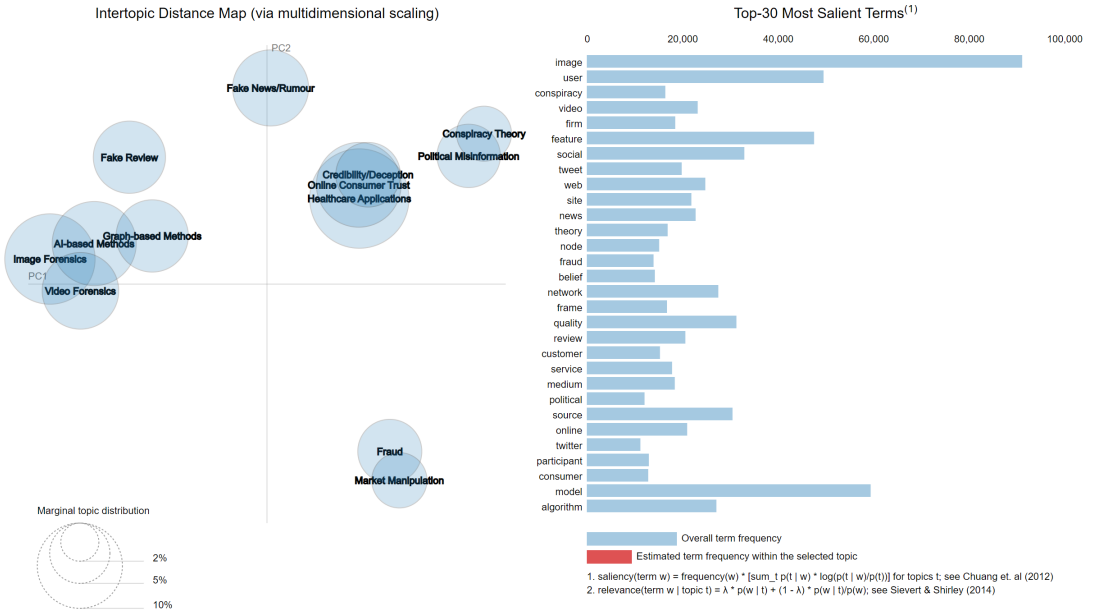


Figure 6: Intertopic distance map and the top 30 most salient terms, generated by LDAvis

Considering the relatively high percentage (37%) of papers on multimedia forensics in the initial set of survey papers, the generated topics may have been highly influenced by those papers, which might have avoided the identification of many other topics. Thus, another topic model trained on a dataset excluding multimedia forensics papers was constructed. To exclude papers on multimedia forensics for generating the second dataset, the topic distribution provided by the first model was utilised. From the topics generated by the first model, shown in Table 3, those corresponding to multimedia forensics, i.e., “Image Forensics” and “Video Forensics”, were selected. Then, the documents assigned to at least one of the two topics with a probability over 0.3 were removed from the dataset. This led to the construction of another dataset less biased towards multimedia forensics research items with a size of 2,674, i.e., 70.52% of the original dataset. Using

this dataset and following the same methodology mentioned in Section 3.2.4, the second LDA model was generated for a recomputed optimum k value of 13.

The list of generated topics by the second model is depicted in Table 4, ordered by their sizes. Out of the generated 13 topics, 8[6] of them were common with those generated by the first model. Nevertheless, the second model generated some new topics, including "News Media", "Health Misinformation", "Fake Account", "Data Mining Approaches", and "Game-Theoretic Methods/Agent-based Systems". Moreover, new close intertopic relationships appeared, as shown in Figure 7. For example, the overlapping circles for Conspiracy Theory and Health Misinformation imply the relatively high amount of research around health-related conspiracy theories in the dataset. Another closely connected group of topics involves "Fake Account" and "Fake Review Detection", which is unsurprising since fake accounts are used to post fake reviews in many cases.

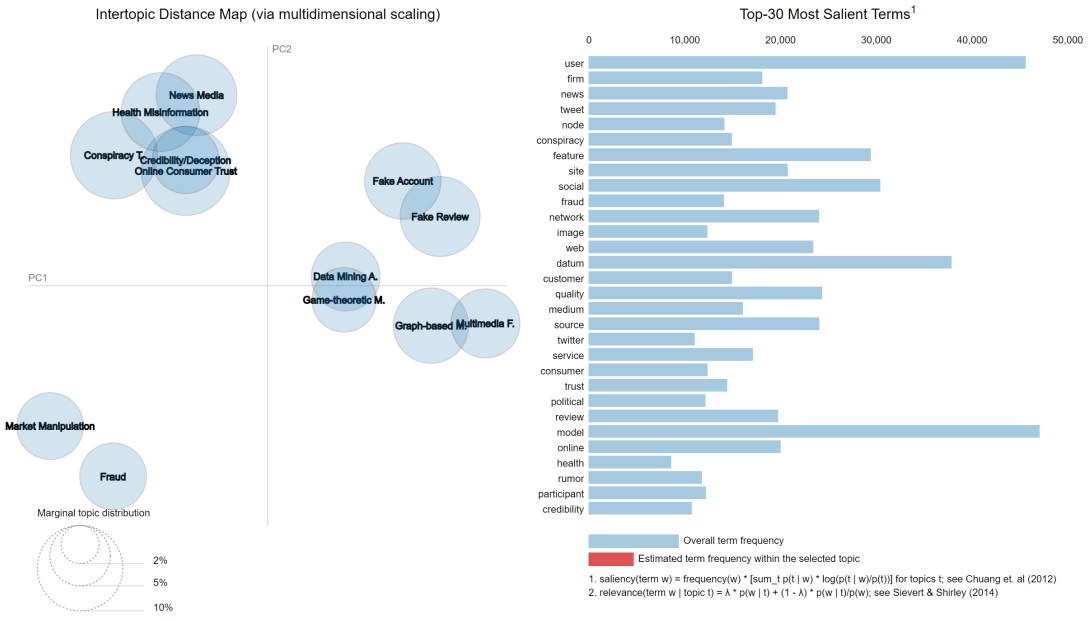


Figure 7: LDAvis-generated intertopic distance map and the top 30 most salient terms when multimedia forensics papers were not considered

[6]Multimedia Forensics is considered as common since Image and Video Forensics are types of Multimedia Forensics.

Table 4: Topics identified when the papers on multimedia forensics were not considered

Topic(s)	**Size (%)**	**Most Relevant Terms**
Online Consumer Trust	10.8	quality, web, site, customer, service, consumer, product, online, satisfaction, website
Conspiracy Theory	10.4	conspiracy, political, belief, theory, party, partisan, people, respondent, group, believe
News Media	8.9	news, medium, social, blog, journalist, story, public, media, content, journalism
Fake Review Detection	8.7	review, stance, feature, topic, text, sentence, sentiment, opinion, label, rumour
Health Misinformation	8.5	trust, health, credibility, individual, student, patient, participant, source, group, organizational
Fake Account	8.0	user, tweet, twitter, fake, spam, social, feature, account, spammer, post
Graph-based Methods	7.7	node, source, network, graph, algorithm, probability, edge, model, random, distribution
Multimedia Forensics	6.5	image, training, video, layer, train, model, input, neural, face, learn
Data Mining Approaches	6.5	datum, mining, rule, cluster, database, application, record, algorithm, fuzzy, claim
Credibility/Deception	6.1	deception, participant, lie, condition, statement, misinformation, truth, cue, vaccine, rumor
Market Manipulation	6.1	market, price, trading, law, trade, action, stock, enforcement, plaintiff, fund
Fraud	6.1	firm, fraud, earning, accounting, financial, restatement, auditor, ceo, audit, variable
Game-theoretic Methods, Agent-based Systems	5.7	agent, game, player, model, query, function, signal, strategy, equilibrium, answer

3.4 Summary

As a complex phenomenon with many different aspects, false information has been studied by several disciplines for decades. This has made it challenging for researchers from different disciplines to come up with a common conceptual understanding of false information and relevant concepts. Therefore, this chapter aimed to contribute to the interdisciplinary efforts for constructing a common understanding of the phenomenon by analysing the recent research in terms of the topics they focused on. Another objective of this chapter was to explore what topics have been more studied by the research community. This can facilitate the identification of potential research gaps and guide researchers on the potential research directions.

Topic models can be considered as simple conceptual models that can serve as an initial step for building more complex conceptual models. When the topics generated by both topic models are collectively considered, the findings of the chapter are supportive. The labels assigned for the topics can be grouped into different categories – each referring to a different aspect of false information research. These categories include different applications of false information (e.g., "Health Misinformation", "Political Misinformation", "Market Manipulation", and "Fraud"), different types of false information (e.g., "Fake News/Rumour", "Fake Review", "Conspiracy Theory", "Multimedia Forensics"), different methods used in false information research (e.g., "AI-based Methods", "Graph-based Methods", "Game-theoretic Methods", and "Data Mining Approaches"), and contexts where false information can exist (e.g., "News Media"). In other words, the assigned labels can be used as a first step to construct an ontology of false information, considering concepts referring to different aspects of the phenomenon. With this respect, to obtain a more complete ontology, a separate LDA model for each generated topic can be constructed recursively to find more relevant concepts related to each topic, provided that there are sufficient amount of training data for each

topic considered.

The work presented in this chapter has some limitations. Firstly, only the research items cited by the survey papers published between 2009 and 2020 were taken into account, meaning that the constructed dataset does not contain research from the last three years. This can be expected to have a significant impact on the presented findings, especially due to the COVID-19 pandemic that emerged in early 2020, which led to an unprecedented volume of pandemic-related publications (Nane et al. 2023), including COVID-19 misinformation. As another limitation, the research items collected for the training dataset of the topic models were limited to those that can be reached through institutional access. This has resulted in the research items from some publishers and research databases being excluded due to lack of institutional access, such as the APA PsychInfo[7] database which contains over 5 million records covering behavioural and social sciences.

[7] https://www.apa.org/pubs/databases/psycinfo

Chapter 4

Entity-Relationship Modelling of the False Information and Fact-Checking Ecosystem

The processes of detecting false information and investigating the veracity of information are complex procedures with various actors involved. Therefore, conceptual models covering different aspects of false information and fact-checking, as well as the roles of different actors involved, need to be modelled for a better understanding of the ecosystem. Much effort has been placed by the research community to shed light on different aspects of false information online, including the behaviours of different actors in response to false information, how false information spreads, and how to more effectively identify and mitigate it using technical and socio-technical methods (Guo et al. 2020; Kumar and Shah 2018). In addition, the process of fact-checking has garnered much attention among researchers more recently (Juneja and Mitra 2022; Neumann, De-Arteaga and Fazelpour 2022; Das et al. 2023). While such previous studies have offered high-level overviews of the ecosystem, there is a lack of entity-relationship (ER) based conceptual models thoroughly describing different actors and their roles in the complicated

ecosystem of false information online fact-checking. This chapter aims to address this research gap through a typical enhanced entity-relationship (EER) model of the false information and fact-checking ecosystem. Unlike previously presented models, this new model covers multiple processes in the ecosystem, including information generation, fact-checking, and lawmaking and regulation, with their corresponding actors involved. The proposed model can be used in many ways, such as for conducting more systematic literature reviews and meta-analyses, automatically building a data-driven computational ontology, and automatically categorising real-world scenarios of false information and fact-checking. To better explain its usefulness, two example applications for the proposed model are introduced in the last two sections. The work presented in this chapter benefits from an archived research paper co-authored by the author of this thesis (Yuan et al. 2022).

4.1 Related Work

Conceptual models based on entities and relationships are helpful tools for exploring different stakeholders of the false information and fact-checking ecosystem, and understanding the relationships between various actors in the ecosystem. Furthermore, such models have several practical applications, such as deriving data models to support knowledge graphs (Tchechmedjiev et al. 2019) and systematic categorisation of incidents. In this manner, the literature involves several studies systematically exploring false information and fact-checking ecosystems in terms of the roles of key players in the propagation and detection of false information. From a general perspective, Sharma et al. (2019) categorised false information actors as *adversary*, i.e., sources and promoters of false information, *fact-checker* combating false information, and *susceptible* affected by false information.

More specifically, Tucker et al. (2018) reviewed the actors that play a role in

producing false information: *trolls*, *bots*, *fake-news websites*, *conspiracy theorists*, *politicians*, *highly partisan media outlets*, *the mainstream media*, and *foreign governments*. Similarly, Zannettou et al. (2019) described different types of actors involved in false information propagation, including *bots*, *criminal/terrorist organisations*, *activist or political organisations*, *governments*, *hidden paid posters and state-sponsored trolls*, *journalists*, *useful idiots*, *true believers and conspiracy theorists*, *individuals that benefit from false information*, and *trolls*.

When it comes to false information detection, Neumann, De-Arteaga and Fazelpour (2022) categorised relevant stakeholders in fact-checking into four groups: *seekers of information*, *sources of information*, *subjects of information*, and *sources of evidence*. Seekers of information are those who might be impacted by the veracity of claims and can retrieve evidence that supports or refutes the claims. Examples of them include average users and professionals, e.g., journalists, who use a fact-checking system for veracity assessment. Sources of information are the ones who make the claims to be fact-checked. Authors of articles, content creators generating or posting a claim, and politicians asserting a claim are among the members of this group. Subjects of information, however, comprise individuals, groups, or conceptual topics mentioned in a claim. Finally, sources of evidence refer to individuals generating evidence for fact-checking. These involve domain experts commenting on a relevant claim, users flagging a false claim on social media, and fact-checkers that analyse a suspicious claim.

Based on the actors involved in false information and fact-checking ecosystems, there has been growing interest in modelling the relationships between different actors. To exemplify, Shu, Bernard and Liu (2019) leveraged network analysis to study fake news detection and mitigation. They examined the relationships between different entities of the news dissemination ecosystem, such as publishers, news spreaders, consumers, news pieces, and comments, in three dimensions – *content*, *social*, and *temporal dimensions*. As another example, Zhou and Zafarani

(2020) presented the life cycle of fake news, involving its creation, publication, and propagation, to explain different fake news detection perspectives. More recently, Mirza et al. (2023) derived a cyber security-inspired framework for modelling disinformation and its mitigation. The framework was based on the findings obtained from interviews with fact-checkers, journalists, trust and safety specialists, researchers, and analysts. In the framework, disinformation events and campaigns were characterised by *threat actors*, *attack patterns*, *attack channels*, and *target audience*. Finally, Zhao and Hu (2023) studied the community-driven content moderation system of Weibo, a Chinese social media platform, with eleven million public moderation cases and decision data from 2012 to 2021. They examined the three actors (i.e., *platform*, *juror*, and *reporter*) involved in the moderation pipeline.

The EER-based conceptual model presented in this chapter differs from the previous models regarding multiple aspects. Initially, the use of EER enables the model to support a standard representation, facilitating the computational applications (e.g., developing a computational ontology and automated incident reporting) using the model. Moreover, the proposed model offers a more comprehensive representation of the ecosystem and covers various actors overlooked by the previous models, such as regulators, accounts, and associations.

4.2 Methodology

Due to the complexity of the false information and fact-checking ecosystem, benefitting from the methodologies for systematically developing more complex conceptual models, such as ontologies, can be quite helpful. Therefore, during the construction of the model, the Ontology Development 101 methodology (Noy, McGuinness et al. 2001) has been followed to systematically identify entity types, and relationships relevant to the ecosystem of false information and fact-checking.

Ontology Development 101 is a simple guide that outlines 7 major steps based on iterative design to give developers a fundamental understanding of ontology development and assist them in constructing an ontology. These steps were adapted for the model development as follows:

Step 1: Determine the domain and scope. The domain of the ontology, the types of questions it should address, and its potential users were all determined at the initial stage of the development process. The main goal was to develop a thorough understanding of the ecosystem of false information and fact-checking, including its structure, scale, and connections among the actors involved, to provide insightful information on how false information can be produced and disseminated, as well as the extent to which fact-checking can function as an effective countermeasure. The proposed model is expected to address a wide range of audiences, including researchers, practitioners, journalists, media outlets, and policymakers.

Step 2: Consider reusing existing ontologies. To the best of found knowledge, no work has been done corresponding to a comprehensive ER-based model to examine the entire ecosystem of false information and fact-checking. There have been, however, some works from other fields that were comparable. The entity-type graph-based semantic model proposed by Lu and Li (2022) is an intriguing example that motivated the construction of the proposed model in this chapter. It illustrates the disclosure of users' personal data to other entities and the advantages of such data-sharing activities. Several entity types (such as 'Person', 'Data', 'Service', 'Organisation', and 'Online group') and a number of edges between entities were used in this semantic model to represent the privacy-benefit trade-off of personal data. The 'Data' entity type can be compared to the information in the false information and fact-checking ecosystem, and the trade-off

between privacy and benefit is comparable to the competition between false information spreading and fact-checking activities in the ecosystem. As a result, it shares some structural and scale similarities with the ecosystem for false information and fact-checking. It is also possible to map more entity types in the semantic model to represent additional ecosystem actors.

Step 3: Enumerate important terms. Specifying a list of terms used in the ontology together with their definitions and attributes was the goal of this step. As mentioned in Section 4.1, Shu, Bernard and Liu (2019) defined a number of entity types including 'publisher', 'news spreader', 'consumer', 'news pieces', and 'comments' to study fake news detection and prevention. To better model disinformation and fake news, Shu et al. (2020) focused on understanding the relationships between entity types such as 'publisher', 'news', and 'users' in their later study. Likewise, Zhou and Zafarani (2020) concentrated on the life cycle of fake news and offered many relevant entity types, including 'news article', 'news creator', 'publisher', and 'users'. More recently, Mirza et al. (2023) presented a disinformation threat framework and included the concept of 'fact-checkers' in the framework. Zhao and Hu (2023) viewed the three primary actors for content moderation on social media platforms to be the platform, juror, and reporter. These examples from the literature indicate that some entity types share the same definition. Reviewing these entity types and the semantic model proposed by Lu and Li (2022), which was introduced in Step 2, the entity types that have a similar definition have been attempted to be grouped and merged, as shown in Figure 8. This yielded the initial list of key terms which contains the following: 'Person', 'Media outlet', 'Fact-checking outlet', 'Service', 'News', 'Comments', 'Account', 'Organisation', and 'Regulator'.

Step 4: Define the classes and the hierarchy. As Figure 8 shows, some key terms can be derived from multiple entity types, e.g., the term 'Person' from

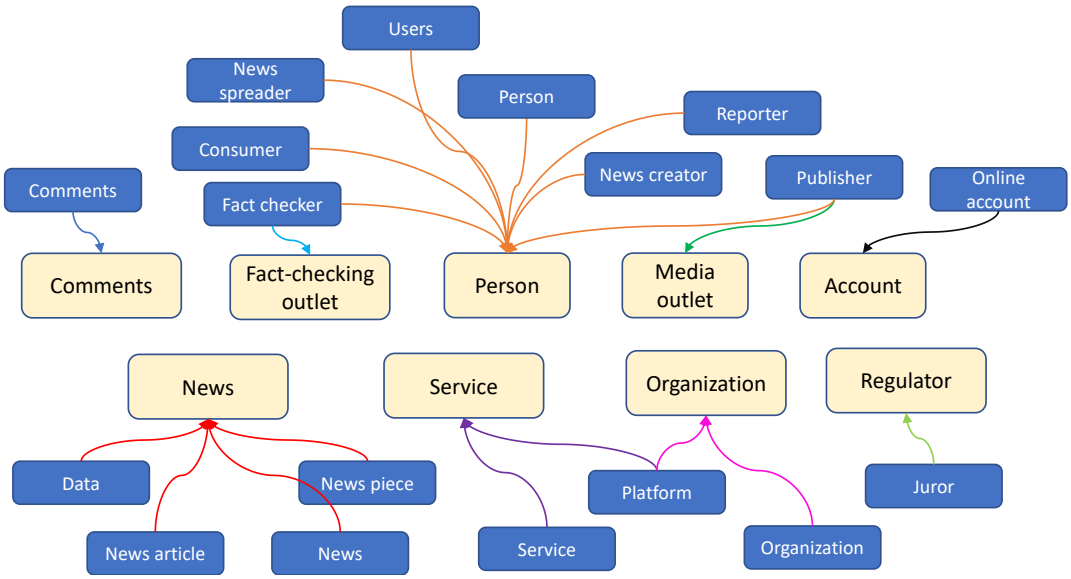


Figure 8: An illustration of the method for selecting the important concepts for the proposed model. The entity types in blue boxes are those that have been introduced in previous work, and the entity types in yellow boxes are those that were determined for the proposed model. The relevant entity types are connected with the edges.

the entity types 'User', 'Consumer', and 'Reporter'. This was used to group the entity types with similar meanings into the same class. The generated classes were then expanded by following a bottom-up approach suggested by Noy, McGuinness et al. (2001). By reviewing and analysing a variety of example scenarios related to the false information and fact-checking ecosystem, the hierarchy and relationships for the entity types mentioned in Step 3 were defined. Besides, new entity types that are closely related to existing ones and their relationships were identified to expand the model.

Step 5 & Step 6: Define the attributes of classes-slots & Define the facets of the slots. These closely linked steps correspond to adding necessary attributes and information to entity types and obtaining the final version of the ontology, i.e., an EER model in this case. With this respect, new superclass entity types were introduced by utilising shared attributes among different entity types.

Step 7: Create instances. This last step aimed to evaluate the effectiveness and usefulness of the generated model through individual instances created. In this manner, a number of real-world incidents and scenarios were modelled as some representative subsets of the entire EER model describing false information and fact-checking ecosystem. More details on the modelled incidents and scenarios are provided in Section 4.4.

4.3 EER Model of False Information and Fact-Checking Ecosystem

Following the methodology presented in Section 4.2, a typical EER model of false information and fact-checking ecosystem was constructed. The proposed model can be considered as a first step towards developing a complete epistemic model of the phenomena.

The proposed EER model involves various entity types and relationships as illustrated in Figure 9. The entity types are shown with rectangles while diamond shapes represent semantic relationships between them with cardinalities on both sides of the edges. Besides, they are coloured differently depending on their type. Entity types representing individual actors, organisational actors, and those produced by actors are coloured green, red, and blue, respectively. To simplify the presented diagram by merging some common relationships, a superclass entity type covering all the actors in the model, called *Actor*, is added and coloured in purple. The attributes of entity types are shown with the labels in the bottom half of the rectangles, ending with three dots to emphasise that more attributes might exist, e.g., *timestamp* and *date* attributes can be added for temporal information. Finally, superclass-subclass hierarchies are shown with the notation adapted from enhanced ER modelling, i.e., small circles with a letter *o* implying that the connected subclasses are overlapping entity types.

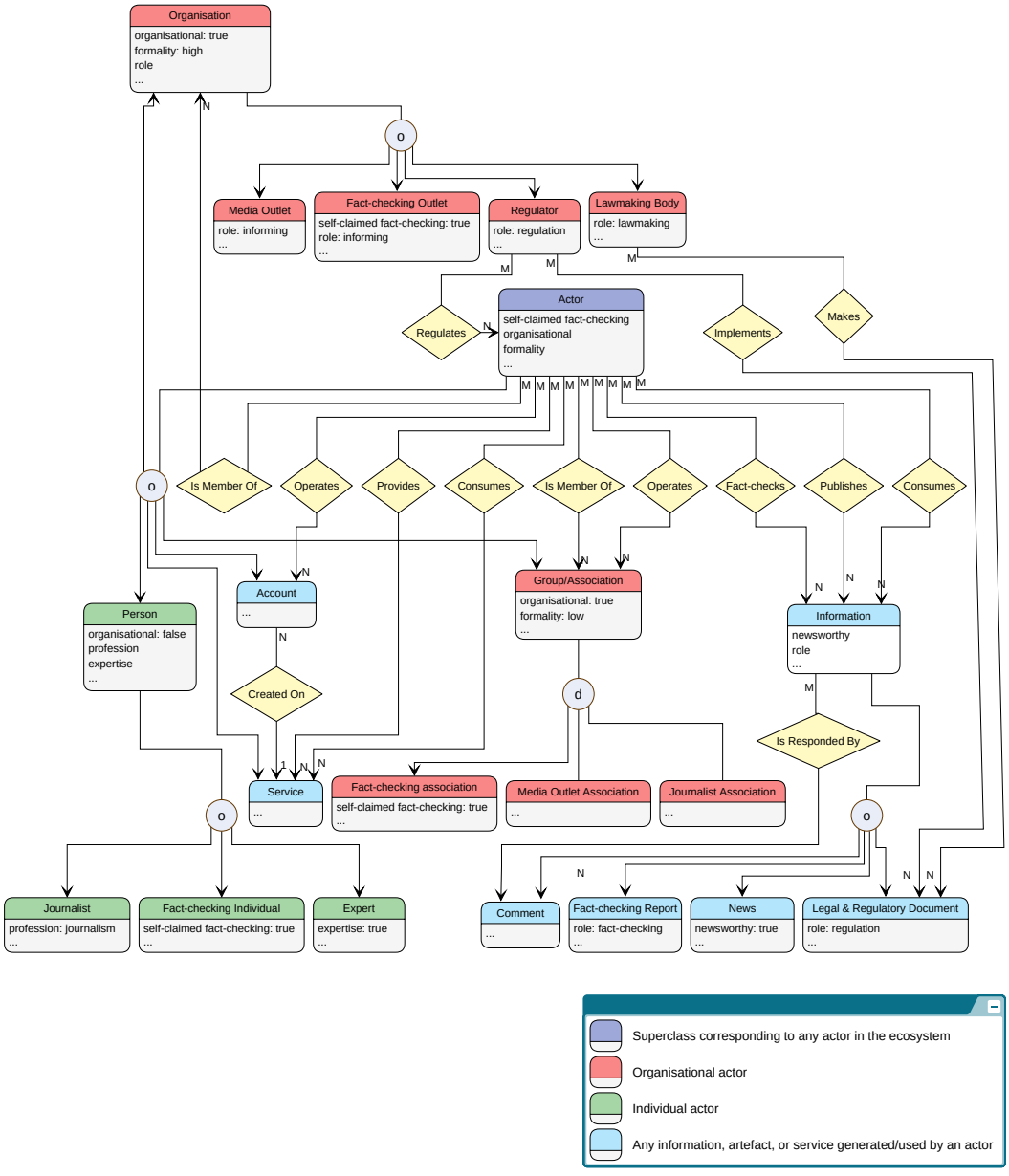


Figure 9: The proposed entity-relationship model of the false information and fact-checking ecosystem

4.3.1 Entity Types

The proposed model contains $N = 21$ different entity types, which are defined below. Those that can have instances representing individuals or organisations that claim themselves as fact-checkers have an attribute of *self-claimed fact-checking*. In addition, different attributes are defined to distinguish the entity types, including those specific to the subclass entity types and the ones inherited from the superclass, e.g., the *Actor* entity type.

Actor: A regulated individual, organisation, or account involved in the ecosystem of false information and fact-checking. This entity type is added to simplify the proposed model by combining some common relationships that might cause redundancy and connecting to a single entity type.

Person: This corresponds to an individual in the physical world. Depending on the role in the ecosystem, it also covers different subclasses shown as separate entity types in Figure 9, including *Journalist* who produces and reviews newsworthy content, *Expert* who contributes to veracity investigation and provides evidence for fact-checking, and *Fact-checking individual* who drives the fact-checking process and produce fact-checking reports. Besides, there can be many other subclasses, e.g., *Naïve User*, *Politician*, or *Legislator*, that are not shown explicitly on the model for the sake of simplicity.

Group/Association: A number of individuals or organisations that share common interests, commitments, or purposes.

Media Outlet Association: An organisation that serves the shared interests of media outlets to protect the general interests of its members in different aspects such as political, regulatory, and legal matters. Examples include News Media

Association[1] in the UK and News Media Alliance[2] in the US.

Journalist Association: A journalism organisation or journalist union that is dedicated to encouraging responsible reporting and ethical behaviour, e.g., British Association of Journalists[3] and National Union of Journalists[4] in the UK, and Society of Professional Journalists[5] in the US.

Fact-checking Association: An association that has members of fact-checking outlets that share the same commitment and purposes to promote fact-checking services and duties. Examples of this entity type include IFCN[6] and EFCSN[7].

Organisation: An organised group of people with a particular purpose.

Media Outlet: A broadcasting/press channel that can provide information in different formats to the public by different paths, including newspaper, television, radio, and social media. This entity type can overlap with the *Fact-checking outlet* entity type when its *self-claimed fact-checking* attribute (inherited from the *Actor* entity type) is true.

Fact-checking Outlet: An organisation dedicated to performing fact-checking or offering fact-checking services, such as Full Fact[8] from the UK and Snopes[9] from the US.

[1] `http://www.newsmediauk.org`
[2] `https://www.newsmediaalliance.org/`
[3] `https://bajunion.org.uk/`
[4] `https://www.nuj.org.uk`
[5] `https://www.spj.org/`
[6] `https://www.poynter.org/ifcn/`
[7] `https://efcsn.com/`
[8] `https://fullfact.org/`
[9] `https://www.snopes.com/`

Lawmaking Body: A governmental body with the power to make, amend, and repeal laws.

Regulator: A governing body that regulates other services, e.g., Ofcom[10] in the UK.

Account: Virtual identities of individuals or organisations using online services, such as social media accounts.

Service: Platforms, tools, or datasets provided by an actor to the public for getting or verifying information. This has a wide coverage, including social media platforms, news websites, and services used for fact-checking, e.g., periodic newsletters about recent fact-checks, databases of previously fact-checked articles, and fact-checking tools (e.g., Logically Browser Extension[11]).

Information: Knowledge or details about someone or something, conveyed in the form of text and/or multimedia.

Legal & Regulatory Document: Laws issued by administrative authorities as well as regulations made to implement principles of laws to bring them into effect (Kosti, Levi-Faur and Mor 2019).

News: A report of recent events or information that attracts public interest, published by a media outlet. It can be physical or digital assets in different formats, such as text, image, video, and audio.

Fact-checking Report: The outcome of a veracity investigation for a piece of information, which can be published by a fact-checker or a fact-checking outlet in different formats such as a report, a blog, or a news item.

[10] `https://www.ofcom.org.uk/`
[11] `https://www.logically.ai/products/browser-extension`

Comment: An actor's opinion, feedback, or other information in response to published news or posted content.

4.3.2 Relationships

As previously mentioned, a textual label on a relationship indicates a semantic connection between two entities. It is a somewhat sophisticated network with a wide variety of relationships between entities, as shown in Figure 9. To better explain the model, it is examined concerning the key roles of the actors through the edges connected to the *Actor* entity type. These roles include information generation, fact-checking, and lawmaking and regulation.

Information Generation

As can be seen in Figure 9, information is generated by (i.e., the *Publishes* relationship) the *Actor* entity type which covers all the actors in the ecosystem. This indicates that information can be directly published by each actor through different channels. Besides, the connection between the *Account* and *Actor* entity types implies that those individual and organisational entity types can also have accounts which they use to publish information. More precisely, the proposed model suggests four main directions for information generation:

1. A *Person* can generate information and release it using different channels. This covers many cases, such as a journalist publishing a news story, a user commenting on a news article, and a domain expert sharing their views on a subject. Furthermore, a *Person* can use an *Account* to publish user-generated content, e.g., a post shared on a social media platform or a product review left by a consumer on Amazon.

2. An *Organisation*, e.g., a media outlet, a fact-checking outlet, or a regulator,

can generate information. This can be exemplified by scenarios such as media outlets publishing news, institutions publishing reports, and regulators making statements.

3. Although a *Group/Association* can be considered as an organisation in many cases, it needs a separate connection to the *Actor* entity type since it could be an information source without being necessarily an established organisation. Examples of such cases include online groups on messaging applications, e.g., WhatsApp and Telegram, sharing posts and social groups, where a physical group of people are socially connected, commenting on a subject. To distinguish *Group/Association* and *Organisation* entity types, the *formality* attribute is used, assuming that organisations are more formal.

4. Although the entity types *Account* and *Service* can be expected to be created and used by the other actors in the ecosystem, there might be cases in which they drive information generation. For instance, a social bot can publish information on social media without any intervention from its owner. Or, an automated service, such as a search engine or an automated fact-checking tool, can generate information on its own.

Everyone can now readily generate and share user-generated content, such as their stories, videos, and images, thanks to the proliferation of services, platforms, and service providers that have been made possible by the growth of internet technology. This makes it easier to develop new kinds of media outlets that are distinct from the established ones. The most representative example is *We Media*. It was (Bowman and Willis 2003) who originally used the phrase to describe news produced by grassroots media.

Although *We Media* has the potential to become a large and influential media outlet (e.g., accounts of celebrities and social media influencers) in terms of subscribers, its management still centres mostly on a single person. Considering the

proposed model, a *We Media* can be represented as a *Person* entity type and an *Account* entity type. The interaction between these entity types and the others can then be further investigated in order to mimic various motivations (such as monetary, political, and/or religious) of *We Media*.

Fact-checking

Similar to information generation, the fact-checking process is specified by the *Actor* entity type through the *Fact-checks* relationship. To distinguish actors playing a central role in fact-checking, the *self-claimed fact-checking* attribute is defined for the *Actor* entity type.

A *Person* entity type with a *true* value for the *self-claimed fact-checking* attribute denotes that the person considers themselves as a fact-checker. Additionally, a *Person* can be a part of a *Fact-checking outlet*. A *fact-checking report* can be published as a result of fact-checking by both an independent fact-checker (as a *Person*) and a fact-checking outlet (as an *Organisation*).

In addition to fact-checkers and fact-checking outlets, media outlets can also perform fact-checking via specific units dedicated to fact-checking, such as *FactCheck*[12] of Channel 4 News and *BBC Verify*[13] (formerly BBC Reality Check). Such fact-checking services often have an advantage over independent fact-checkers and fact-checking outlets in that they can reach wider audiences (Graves and Cherubini 2016).

It is important to note that the *Actor* entity type is connected to the *Service* entity type via *Provides* and *Consumes* relationships, signifying that *Person*, *Group/Association*, and *Organisation* entity types can produce various services in different formats, including platforms, tools, and datasets to share with the fact-checking community. A *Service* can therefore be used by other actors to

[12] `https://www.channel4.com/news/factcheck`
[13] `https://www.bbc.com/news/reality_check`

cross-reference, fact-check, and do additional analysis. Besides, it is also considered as a subclass of *Actor*, implying that a *Service* can also perform fact-checking on its own, e.g., when it is an automated fact-checking tool.

Lawmaking and Regulation

Regulators must apply different regulations and/or laws issued by lawmaking bodies in various circumstances. The relationships between the *Lawmaking Body* and *Regulator* entity types, and the *Legal & Regulatory Document* entity type have been implemented in the model to describe this. The *Lawmaking Body* entity type is responsible for issuing laws necessary to regulate the ecosystem. The *Regulator* entity type, however, regulates the actors involved in the ecosystem, such as the *Person*, *Organisation*, and *Group/Association* entity types, by implementing the issued laws and regulations. For instance, the Charity Commission[14] (i.e., the *Regulator* entity type in the model) regulates nonprofit organisations like the fact-checking outlet *Full Fact* (i.e., the *Fact-checking Outlet* entity type), and its powers and procedures are outlined in the Charities Act 2011 (i.e., the *Legal & Regulatory Document* entity type). The Charity Commission also publishes a variety of guidelines (i.e., the *Service* entity type), such as "Trustee role and board" and "Money, tax, and accounts" to assist in the establishment and administration of a charitable organisation.

4.4 Application 1: Modelling Real-World Scenarios

This section aims to demonstrate how the proposed EER model can be applied to examine relevant real-world scenarios, which can help the identification of the actors involved and more systematical incident reporting through the categorisation

[14] `https://www.gov.uk/government/organisations/charity-commission`

of incidents based on shared entities and/or relationships. The three examples of real-world scenarios provided can guide researchers, practitioners, and policymakers on how to analyse real-world incidents and help them reveal the actual scale and scope of the incidents. They are all represented graphically using the same colour scheme as the model shown in Figure 9, and the scenarios are chosen to depict a sub-ecosystem that is a representative portion of the entire model.

4.4.1 BBC Breakfast Incident

A clip of military aircraft was broadcast on BBC Breakfast on February 25, 2022, with the claim that it was the Russian military entering Ukraine after the invasion of Ukraine. A journalist and Full Fact fact-checker Sarah Turnnidge investigated this video clip and produced a fact-checking report, which Full Fact then published (Turnnidge 2022). According to the report, the BBC Breakfast video was actually of military parade preparations scheduled to take place outside Moscow in May 2020 rather than the alleged invasion. Interestingly, the video clip that was presented on the BBC Breakfast Show was initially derived from a widely shared video clip on X (formerly Twitter), which has been fact-checked by many fact-checkers working for fact-checking outlets, including Amy Sood from AFP Fact Check (Sood 2022) and Abbas Panjwani from Full Fact (Panjwani 2022). The model of the incident based on the proposed EER model is depicted in Figure 10.

4.4.2 Trump's Twitter Account Suspension Incident

The former US President of the United States, Donald Trump, tweeted a video message on January 6, 2021, claiming that the presidential election had been rigged and urging his followers to take action (Time 2021). Trump's followers subsequently assembled and assaulted the US Capitol, leading to the deaths of five people. Due to "the risk of further incitement of violence", Twitter permanently suspended Trump's Twitter account on January 8, 2021, following its *Civic*

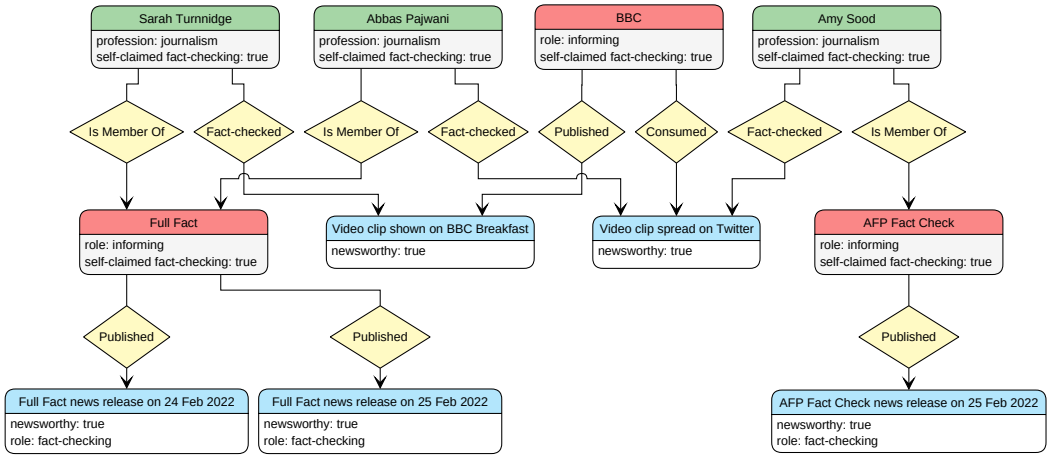


Figure 10: Modelling a real-world scenario that involves BBC and multiple fact-checking outlets

Integrity Policy. And, Twitter's judgement was not contested by the US communications regulator, the Federal Communications Commission (FCC) (Reuters 2021). Figure 11 demonstrates the corresponding model covering the actors involved in the incident.

4.4.3 Fact-checking Fact-checkers

Shortly after the beginning of the Russian invasion of Ukraine, a newly established pro-Russian so-called "fact-checking service," *War on Fakes*, published a piece via their website and Telegram channel for debunking the claim that Ukrainians were not spreading disinformation about Russians. However, PolitiFact, an IFCN-verified fact-checking outlet, examined over 380 so-called "fact-checks" that War on Fakes had released in English and found that they were essentially Russian propaganda (Romero 2022). Figure 12 shows the model generated based on the incident. This real-world scenario demonstrates how fact-checking is a dynamic process that occasionally calls for fact-checking fact-checkers.

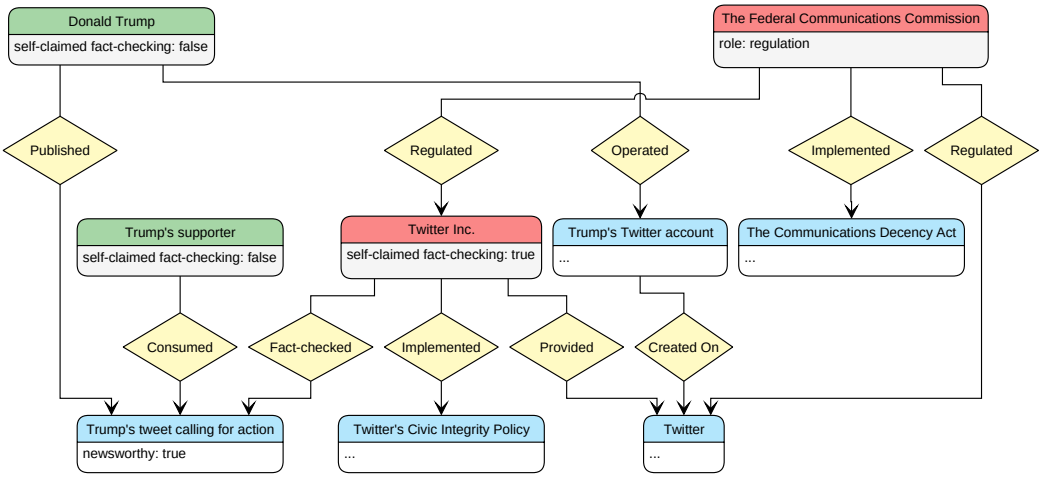


Figure 11: Modelling the suspension of Trump's Twitter account

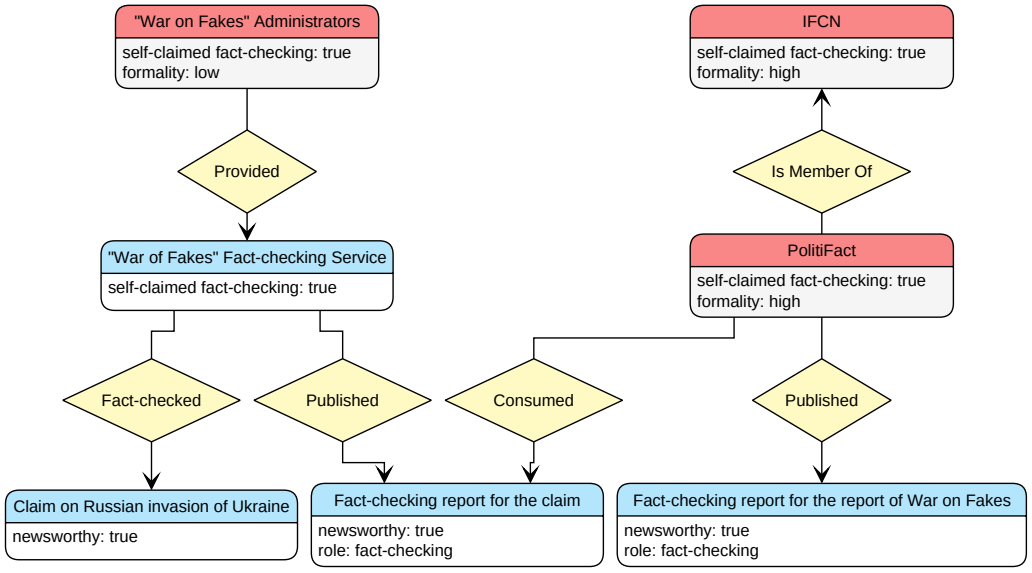


Figure 12: Modelling the War on Fakes incident

4.5 Application 2: Reviewing Recent Literature

As another example application, this section is intended to illustrate the value of the presented EER model through a review of a limited number of recent, relevant studies based on the model. The methodology followed for the literature review can be utilised to better understand the current landscape and to identify overstudied and understudied topics.

4.5.1 Methodology

The following search query was used to seek relevant publications on Google Scholar, limiting the results to those that had been published in 2023.

```
"misinformation" OR "disinformation" OR "false information" OR "fake news" OR "fact-checking"
```

The top 10 peer-reviewed English articles were selected from the search results after being sorted by relevance. However, one publication that showed up in the search results was excluded since it appeared only because it used an included keyword in an irrelevant context. A wide range of academic fields, including Communication Studies (van der Meer, Hameleers and Ohme 2023; Freiling et al. 2023; Hameleers 2023), Psychology (Altay, Berriche and Acerbi 2023; Lawson, Anand and Kakkar 2023; Modirrousta-Galian and Higham 2023), Computer Science (Roy et al. 2023), Marketing (Dholakia, Ozgun and Atik 2023), Economics (Azzimonti and Fernandes 2023), and Political Science (Erlich and Garner 2023), were represented in the papers that were ultimately chosen. All selected papers were qualitatively analysed after the paper selection step to find topics relevant to the entity types and relationships in the proposed model and then encoded with incrementally specified tags in the format "$Entity_1 Relationship Entity_2$". Such a systematical generation of tags can facilitate the tagging process of typical literature reviews. Table 5 shows the list of tags with the corresponding papers,

identified in consequence of the analysis. These tags were then used to review the literature in terms of different aspects of the ecosystem by grouping the papers containing discussions on the same or similar relationships covered by the EER model.

Table 5: The list of generated tags and relevant papers for each tag

Tag	**Relevant Paper(s)**
UserConsumesInformation	Freiling et al. (2023); Dholakia, Ozgun and Atik (2023); Azzimonti and Fernandes (2023); van der Meer, Hameleers and Ohme (2023); Modirrousta-Galian and Higham (2023)
UserPublishesInformation	Freiling et al. (2023); Lawson, Anand and Kakkar (2023)
UserFactchecksInformation	Erlich and Garner (2023); Altay, Berriche and Acerbi (2023); Roy et al. (2023)
GroupPublishesInformation	Lawson, Anand and Kakkar (2023)
OrganisationPublishesInformation	Hameleers (2023)
MediaOutletPublishesInformation	Dholakia, Ozgun and Atik (2023)
AccountPublishesInformation	Azzimonti and Fernandes (2023)
ExpertProvidesService	Roy et al. (2023); Modirrousta-Galian and Higham (2023)
RegulatorRegulatesService	Altay, Berriche and Acerbi (2023); Hameleers (2023); Dholakia, Ozgun and Atik (2023)

4.5.2 Literature Review

In general terms, the tags shown in Table 5 indicate that user-information relationships were more studied than other portions of the EER model. Besides, rather than fact-checking, most studies focused on information generation and consumption.

More precisely, the literature involves attempts to develop a better understanding of the factors and motivations behind believing and sharing false information. Freiling et al. (2023) found that political beliefs, anxiety, and their interactions can affect whether people believe and share false claims. Moreover, Lawson, Anand and Kakkar (2023) illuminated the role of conformity pressure and social costs within online groups people belong to when it comes to false information dissemination. Apart from being ideologically and socially motivated, Hameleers (2023) pointed out that false information can also be disseminated in the form of disinformation with political motivations such as disinformation campaigns originating from the Russian Internet Research Agency (IRA).

As examined by Dholakia, Ozgun and Atik (2023), the current economic organisation of media outlets also contribute to the generation and dissemination of false information. With the emergence of social media platforms, user data has become another profitable product for media outlets. Such platforms offer a "phatic communication" environment to their users, lacking the structural mechanisms and editorial processes of the news media outlets that can avoid false information. In addition, bots on social media platforms are among the spreaders of false information (Azzimonti and Fernandes 2023).

When it comes to dealing with false information, recent research also covers different debunking and prebunking strategies and their consequences. As Erlich and Garner (2023) demonstrated with their study conducted with Ukrainians, people are capable of distinguishing false information from true information in many cases. They follow various strategies to identify false information, from cross-checking and following recent fact-checks to different digital tools such as search engines and online encyclopaedias. However, Altay, Berriche and Acerbi (2023) pointed out that more research is needed on the uses and misuse of such digital tools. In addition, there are efforts to develop fully automated systems to detect false information. For instance, Roy et al. (2023) proposed a deep stacked

LSTM model for false information prediction on social platforms to identify it as soon as posted on the social platform.

As remarked by van der Meer, Hameleers and Ohme (2023), attempts to combat false information, such as fact-checking and misinformation warnings, can also have the unintended consequence of a decrease in general trust in factual correct information. Therefore, it is crucial not only to fight against false information but also to reestablish trust in legitimate news. Modirrousta-Galian and Higham (2023) came to a similar conclusion in their overview of gamified inoculation interventions which aim to increase people's resistance against false information by exposing them to common false information generation techniques. Their findings suggest that such interventions did not improve discrimination between true and false news, but increased false responses to all news items.

Finally, the collected papers discussed the existing ecosystem in terms of its regulation. With the digital media ecosystem, including social media, traditional gatekeepers no longer control the media landscape, facilitating the dissemination of false information alongside true information, according to Altay, Berriche and Acerbi (2023). Therefore, as Hameleers (2023) underlined, social media platforms and big tech companies need to play a central role in regulating this digital ecosystem to protect their users from false information online. With this respect, Dholakia, Ozgun and Atik (2023) suggested the adoption of new legal frameworks toward holding social media networks accountable for limiting the spread of false information on their platforms, based on the success of legal frameworks concerning privacy protection.

4.6 Summary

The false information and fact-checking ecosystem involves many types of entities with different roles. As shown in Chapter 3 with the identified topics and the

intertopic distance maps, these are closely connected, which suggests exploring the relationships between them to develop a better understanding of the complicated ecosystem. This motivated the construction of an enhanced entity-relationship (EER) model of the ecosystem, covering different actors with their roles. The proposed model can be utilised for accomplishing a wide range of tasks, two of which, i.e., modelling real-world scenarios and reviewing recent literature, were covered in this chapter to show its usefulness.

Apart from the two applications provided in this chapter, the presented model can be leveraged for different purposes. For instance, empirical studies can be designed by benefitting from the proposed model to identify relevant actors for a more representative recruitment of participants. Furthermore, modelling and simulating the false information and fact-checking ecosystem, particularly with agent-based models, can be achieved based on the EER model, as another application. Ultimately, in a variety of false information and fact-checking scenarios, different types of supportive and conflicting relationships between various entities can also assist researchers in applying different theoretical frameworks, such as game theory and epistemic logic, to rigorously study the dissemination of false information, the behavioural changes of different entities, and the overall evolution of such ecosystems.

Similar to the generated topic models described in Chapter 3, the presented EER model can also serve as a basis for a computational ontology to be developed. Such an ontology derived from the EER model can have several applications. For example, large-scale Internet measurement studies can be conducted based on the ontology to study the dissemination of false information and the roles of different actors. Besides, it can be leveraged to identify new phenomena associated with false information and fact-checking, through automated reasoning. Other than that, enhancing existing fact-checking solutions with the ontology derived from the EER model, e.g., pinpointing key entities and relationships to determine the

optimal communication strategy, to increase their awareness of the ecosystem can be another potential application. Finally, such an ontology can be utilised to enhance the existing knowledge graphs, such as ClaimsKG (Tchechmedjiev et al. 2019) which is based on a data model limited to the entities for claims, their authors, and the fact-checking organisation reviewing them, to support more entity types and relationships.

Chapter 5

Human Perceptions and Attitudes towards Fact-checking Tools

Although (semi-)automated fact-checking highly involves technical research, the consideration of human aspects is quite important for the practical adoption of developed fact-checking tools. However, as discussed in Section 2.6, existing fact-checking solutions generally overlooked human aspects such as trust and explainability (Das et al. 2023). Considering the research objectives of this thesis mentioned in Section 1.2, it is crucial to identify how better consideration of such aspects can support the design of a new fact-checking tool. Therefore, this chapter explores key principles that should be followed when developing fact-checking tools containing automated components with an increased consideration of human aspects. For this purpose, initially, an overview of existing fact-checking tools is presented through a new taxonomy to identify the scope of the current landscape. Informed by the proposed taxonomy, an online user survey is conducted to investigate people's level of trust towards different stakeholders of the information ecosystem, their attitudes towards existing fact-checking tools, and

their fact-checking behaviours. While the main focus of this chapter is addressing RQ3 from the research objectives, more precisely, it seeks answers to the following research sub-questions:

- **RQ3.1:** What information and information providers do users feel the need for fact-checking?
- **RQ3.2:** What stakeholders (including the user itself) do users trust as reliable information providers for fact-checking and fact-checkers?
- **RQ3.3:** To what extent are users familiar with fact-checking tools and what are their usage patterns?
- **RQ3.4:** To what extent do users trust fact-checking tools with a level of automation?
- **RQ3.5:** How should fact-checking tools be designed to enhance explainability for non-expert users?

5.1 Related Work

5.1.1 Trust in Fact-checking Tools and Different Actors

According to the Edelman 2023 Trust Barometer (Edelman 2023), an annual online survey with over 32,000 respondents in 28 countries, the most trusted news sources appeared as search engines and traditional media. When it comes to individual actors, however, scientists appeared to be the most trusted (76%). Similar results were obtained by two KFF[1] polls on COVID-19 misinformation conducted in the US with 1,519 and 1,888 participants. No news media sources, including traditional and social media, were trusted by the majority of the public (Hamel et al. 2021b), and the public had the highest level of trust (up to 85%) in health

[1] `https://www.kff.org/`

experts (Hamel et al. 2021a). Despite the overall findings of the Edelman 2023 Trust Barometer and KFF polls, there might be differences in most trusted actors across countries. For instance, Gruzd and Mai (2020) surveyed 1,500 adults in Canada focusing on COVID-19 misinformation. They reported that participants most trusted public service and governmental bodies (61%) when getting news about COVID-19.

Since there has been a growing interest in the research community to utilise AI in fact-checking tools, trust in AI needs to be considered. Ada Lovelace Institute and The Alan Turing Institute (2023) conducted a survey of 4,010 respondents in the UK regarding their attitudes towards AI. The results underlined the importance of explainability for the trustworthiness of AI-based systems, even if this may require concurring with lower accuracy. Moreover, participants expressed the idea of humans making ultimate decisions. Besides the general public, experts also featured the importance of trust in AI for the adoption of AI-based fact-checking systems, according to a study conducted by Seböck, Biron and Lampoltshammer (2023) with over 100 communication experts. Regarding the use of AI in fact-checking tools, Shin and Chan-Olmsted (2023) examined the human and software-related factors influencing users' level of trust in AI-based fact-checking systems through a survey with 1,127 participants. They reported that those who are more accustomed to using fact-checking tools, more familiar with AI and have higher levels of overall trust in AI tend to trust AI-based fact-checking tools more conveniently.

5.1.2 Attitude towards Fact-checking

Regarding the attitudes towards fact-checking, there are two sides of the coin – the attitudes of professionals, e.g., journalists and fact-checkers, and the attitudes of the general public. The existing literature mostly focuses on the former (Nieminen and Rapeli 2019), suggesting the need for more research on the latter.

Mena (2019) surveyed 61 professional journalists based in the US regarding the principles and boundaries of fact-checking. The respondents reported strong levels of agreement with several fact-checking principles, such as transparency of sources and boundaries with partisanship and activism. A similar study was conducted by Rodríguez-Pérez et al. (2022) – their survey of 122 fact-checkers from 17 countries in Ibero-America indicated that less-experienced fact-checkers were more likely to consider activism a purpose of fact-checking while younger fact-checkers were more likely to strive to uphold the ideals of journalism and serve as a commitment to information transparency.

Petter Bae Brandtzaeg and Ángeles Chaparro Domínguez (2018) interviewed 32 young journalists and performed a content analysis of 595 posts from social media users to explore how journalists and social media users perceived online fact-checking services. The results showed that young journalists considered such services helpful in the investigative reporting process although they were unwilling to rely on these for fact-checking. Similarly, while some users emphasised the value of these services, others expressed strong distrust. Apart from trust in fact-checking services, researchers have also been concerned about the visibility of such services. Kyriakidou et al. (2022) conducted 14 focus groups with 52 participants from the UK as well as two surveys with 1,065 and 542 participants, respectively. Their findings indicated that the majority of the participants were unaware of active fact-checkers in the UK. Nevertheless, the participants reported their practices to verify information, illustrating the need for fact-checking for daily news consumption.

5.1.3 Fact-checking Practices

Understanding fact-checking practices of people can help the development of more effective fact-checking solutions and inspire fact-checking tool designs (Yin et al. 2023). For this purpose, Francis et al. (2023) investigated the news verification

practices of US users through an online survey of 400 participants and 19 follow-up interviews. The participants reported many fact-checking strategies they applied, including checking the news publisher, article's source(s), sharer(s) of the article, publish date, corrections to the article, embedded content, and Google search results about the content. Based on the findings, they underlined the perceived need for usable fact-checking tools and suggested several design considerations for developing such tools. Poynter Institute for Media Studies, MediaWise and YouGov Inc. (2022) conducted another survey of 8,585 respondents in seven countries on generational behaviours and concerns around false information. The results showed that respondents were most likely to verify information they saw on search engines and instant messaging apps. Furthermore, the most common methods for fact-checking were checking the source and date of the post and using a search engine. During the veracity assessment, what respondents found most important was whether verdicts were supported by sources and facts, indicating the importance of explainability. Nonetheless, it is worth noting that there were differences in fact-checking behaviours across generations. Whereas younger generations were more confident in identifying false information than older generations, they tended to use more advanced techniques when using search engines.

Sakhnini and Chattopadhyay (2022) provided further evidence on generational fact-checking behaviours. They systematically reviewed 45 smartphone fact-checking apps and performed a semi-structured interview with 11 older adults to shed light on the types of fact-checking apps available and the fact-checking practices of older adults. The findings suggested that older adults were not among the users of fact-checking apps despite their other fact-checking practices, e.g., asking friends or family members with domain expertise, or searching on Google. Finally, Bouleimen et al. (2024) asked 261 students to search online to determine the accuracy of six different news items. Their findings implied that the students were more likely to verify true news than to refute fake news during online

searching. Furthermore, most participants acquired information mostly online, and those who got their knowledge through books and online searching were the most accurate at determining the veracity of a news item.

5.1.4 Effectiveness of Fact-checking Tools

A number of surveys and user studies explored the effectiveness of different fact-checking tools and strategies. In their meta-analysis covering 30 studies with a total sample size of 20,963, Walter et al. (2020) examined the effectiveness of fact-checking in eradicating political misinformation. When employing truth scales, refuting certain sections of a claim, and fact-checking campaign-related claims, the effects gradually reduce, although fact-checking has a significantly positive overall effect on political beliefs.

Regarding the effectiveness of specific tools, Saling et al. (2021) investigated the willingness of fact-checking newsletter subscribers to spread potentially false information through a survey of 1,397 Australian participants. They discovered that fact-checking newsletters are not sufficiently effective in reducing the spread of false information. Another user study was conducted by Ibrahim, Safieddine and Pourghomi (2023) with 117 university students in the UK. The participants were provided with a semi-automated fact-checking tool in which users can right-click on a news item to check if it had been reported previously. The results showed that most participants failed to identify false information. Nevertheless, the participants expressed a need for an effective tool assisting people's fact-checking practices despite some respondents expressing distrust in third-party verification tools.

The findings of studies about the effectiveness of fact-checking indicate the need for research on key principles of designing fact-checking tools. With this respect, Bhuiyan et al. (2021) conducted interviews with 15 journalists and 16 news consumers regarding a scenario demonstrating three transparency features

– source characteristics, crucial details of the event, and reporting style. The participants suggested several design considerations for improved transparency, including providing objectivity and evidence indicators. Despite journalists and news consumers agreeing on many suggestions, some of their suggestions were also conflicting.

5.2 A Taxonomy of Fact-checking Tools

Before investigating people's attitudes and behaviours regarding fact-checking tools, it is crucial to explore what kind of fact-checking tools exist in the current landscape. This can also ensure that the conducted survey provides a wider coverage. With this respect, the literature covers several studies that overview existing fact-checking tools. For example, Nakov et al. (2021) reviewed the available technologies that can assist human experts in different fact-checking steps, such as finding check-worthy claims, identifying previously fact-checked claims, evidence retrieval, and veracity assessment. More recently, Westlund et al. (2022) offered a taxonomy of digital technologies used in fact-checking, leveraging a socio-technical framework that consists of *social actors*, *technological actants*, *audiences*, and *activities*. It is based on a mapping of 136 existing fact-checking technologies according to the corresponding general fact-checking stage where they are made use of, which covers *identification*, *verification*, and *distribution*. Despite such efforts to provide an overview of existing tools, they heavily focused on different stages of fact-checking for categorisation, overlooking different aspects of such tools, e.g., to what extent they are automated or where they are deployed.

To fill the above-mentioned research gap, this section introduces a new taxonomy of fact-checking tools based on existing tools in use as well as approaches proposed in the literature. To construct the taxonomy, the available tools and approaches were first collected from existing taxonomies and survey papers, and

then, more of them were searched over the Web and research databases. Once a pool of tools and approaches was obtained, the main skeleton of the taxonomy was shaped by the author through several attributes used for generating different categories. When the attributes were selected, those less mentioned in the literature and more useful for the thesis objectives were preferred. Then, the author and his primary supervisor iteratively refined the taxonomy by adding new nodes and removing or updating the existing ones.

As shown in Figure 13, the proposed taxonomy involves a categorisation of fact-checking tools in terms of the extent they are automated (i.e., *Level of automation*), their contribution to fact-checking (i.e., *Method*), and the platforms they support (i.e., *Platform*). Note that the entities included are not necessarily distinct from each other, meaning that a tool can belong to multiple entities.

5.2.1 Level of Automation

As discussed in Section 2.4, fact-checking can be fully automated, semi-automated, or fully manual. Likewise, fact-checking tools include fully automated models mostly based on blackbox or whitebox AI models, semi-automated tools joining the forces of automation (e.g., based on some AI methods) and human actors (e.g., human fact-checkers, domain experts, or the crowd), and manual tools operated by human actors only without any automation[2].

5.2.2 Method

Fact-checking tools can contribute to different stages of fact-checking by facilitating different tasks. To begin with, analysing metadata is a useful way of getting

[2]Despite any digital tool can be expected to contain a certain level of automation, the focus here is if the fact-checking itself uses automation, ignoring other automation elements of the tool

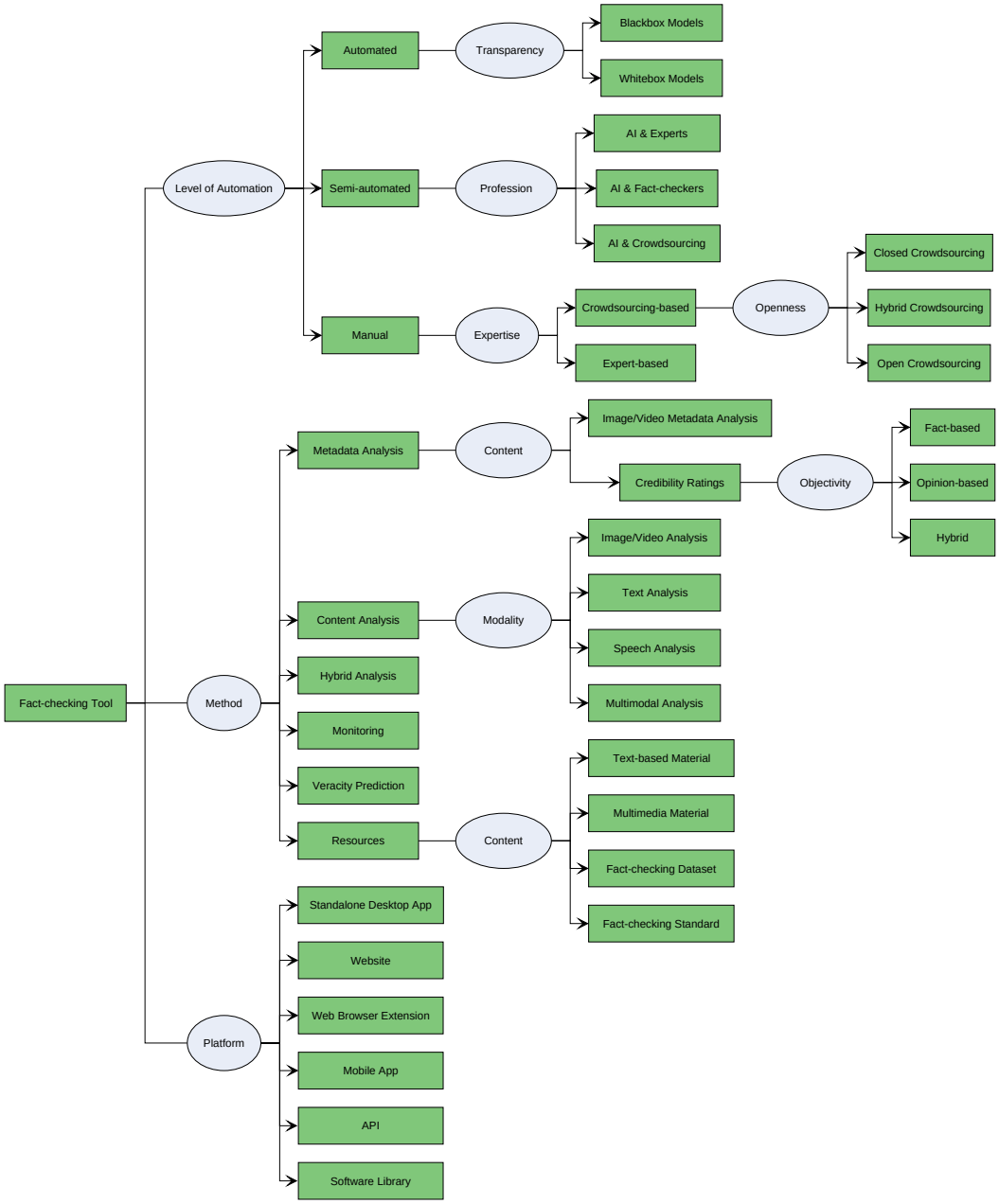


Figure 13: The proposed taxonomy of fact-checking tools

insights about the veracity of information. One common metadata analysis approach is to analyse image/video metadata (e.g., EXIF data for images) to verify the authenticity of the multimedia embedded in the content. With this respect,

InVID (Teyssou 2019) is quite a popular tool among fact-checkers for accomplishing such analyses. In addition, credibility ratings can be considered another type of metadata analysis. As the information sources, such as news outlets, websites, authors, or politicians, that have shared false information previously are likely to share false information again, tools based on credibility ratings with different levels of automation (from manually labelled ones (Chen and Freire 2020) to those using LLMs (Yang and Menczer 2023)) and objectivity can be leveraged to track potential false information before going viral.

Another group of fact-checking tools is tools for analysing the content itself. While content analysis tools can focus on a single modality, such as text, image, video, and audio, there also exists tools supporting multimodal analysis, such as Dataminr[3]. Especially in the context of social media which provides fast communication facilitating the rapid dissemination of false information, monitoring and web analytics tools are also quite helpful in identifying false information and analysing reader engagement for fact-checking reports (Liu et al. 2023). Examples of such tools include Google Analytics[4], Botometer (Yang, Ferrara and Menczer 2022), and the Fact-checking Observatory (Burel and Alani 2023). Beyond assisting users with fact-checking, there also exists fact-checking tools aiming to predict the veracity of information accurately based on prior data they have been trained on and/or relevant information provided by their human components, such as users or experts. Finally, fact-checking resources, including relevant written (e.g., handbooks, newsletters, and guides) and multimedia materials (e.g., visual aids, training videos, and tutorials), datasets, and standards (e.g., C2PA (Rosenthol 2022) and JPEG Trust (Temmermans et al. 2023)) are covered in the taxonomy with separate entities.

[3] `https://www.dataminr.com/`
[4] `https://analytics.google.com/`

5.2.3 Platform

The final main categorisation of fact-checking tools in the taxonomy is based on the platform they support. While existing tools are commonly served in the forms of standalone desktop applications and websites, web browser extensions and mobile applications are drawing a growing interest due to their practicality and usability. Besides, APIs and software libraries are among the common ways of making the working prototypes of new tools and approaches publicly available for demonstration, testing, and reproducibility purposes. Note that some fact-checking tools can support multiple platforms.

5.3 Methodology

Since no past surveys covering the aspects within the interests of this thesis have been found, an online survey was conducted to explore people's perceptions, attitudes, and practices regarding fact-checking tools. Participants were recruited through the Prolific platform[5].

5.3.1 Survey Procedure

The survey procedure was approved by the Central Research Ethics Advisory Group at the University of Kent (Application No. CREAG134-09-23). The constructed taxonomy of fact-checking tools was leveraged when developing the survey to cover different types of fact-checking tools currently available. For instance, the options of the survey questions were ensured to include different levels of automation (i.e., manual, semi-automated, and automated approaches) and different stakeholders commonly used in existing tools (e.g., AI, experts, fact-checkers, and crowdsourcing workers) when appropriate. Besides, existing surveys on similar topics (e.g., (Francis et al. 2023)) were reviewed for inspiration about potential

[5] `https://www.prolific.com`

survey questions to be added. To make sure that the survey questions were understood correctly and that the survey coverage was sufficient for the needs of the study, two pilot surveys were conducted, each with 30 participants recruited through Prolific. The survey was developed using an internal LimeSurvey-based survey system managed by the Institute of Cyber Security for Society (iCSS) at the University of Kent[6].

The survey questionnaire consists of 37 questions in six sections, including single- and multiple-choice questions, five-point Likert scales, drop-down questions, and free-response text boxes. While the first section aims to collect general information about the participants, such as demographics, each of the other sections maps to the research questions from RQ3.1 to RQ3.5, respectively. The full questionnaire can be seen in Appendix A. The sections of the questionnaire were as follows:

1. General questions
2. Information and information providers that require fact-checking
3. Reliable information providers for fact-checking and fact-checkers
4. Familiarity with and usage patterns of fact-checking tools and services
5. Trust in fact-checking tools with a level of automation
6. Explainability in fact-checking tools

The questionnaire started with a cover page providing the participant information sheet and the consent form. The participants were able to proceed to the next page to participate in the actual survey only after they provided their consent.

[6] `https://cyber.kent.ac.uk/survey/`

5.3.2 Recruitment, Screening, and Data Collection

To estimate the required number of participants, a power analysis was performed by using the G*Power tool (Faul et al. 2009). As an example scenario for the analysis of the collected responses, one-way ANOVA can be utilised to determine whether there is any statistical difference in trust in fact-checking tools for users with different levels of education, i.e., less than a Bachelor's degree, a Bachelor's degree, and a postgraduate degree. With an effect size of 0.25, α error probability of 0.05, power of 0.95, and three groups, the estimated sample size turned out to be 259. Inspired by this obtained sample size and considering many such scenarios, the target number of participants for the survey was determined to be 300, taking into account potential responses to be discarded due to a number of reasons, such as incomplete data and outliers.

The participants were recruited anonymously by using Prolific and no personal or sensitive data was collected for the study. By making use of Prolific's prescreening features, the participants were limited to adults who were at least 18 years old and fluent in English. No other requirements were applied to capture a more representative sample of the general population. Each participant received financial compensation at the rate of £6/hour for their time spent on taking the survey, which corresponded to £1.5 for the conducted 15-minute survey. After the pilot surveys, 242 more participants were recruited for the main survey. The average completion time for the recruited participants was 14 minutes and 53 seconds. The survey questionnaire included necessary attention checks, and the responses of those who failed them were disregarded during the data analysis.

Since the pilot studies did not suggest a significant modification to the questionnaire, the responses collected in the pilot surveys have also been considered for the actual survey instead of recruiting more participants to reach the intended number of participants. In total, 313 people participated in the study, and the responses of 11 of them were excluded either because they failed the attention

check or completed the survey too quickly. Thus, the responses from 302 participants have been considered when analysing the responses. For analysing the survey data, IBM SPSS Statistics (version 29) was utilised to run different types of statistical tests, including the chi-squared, one-way ANOVA, Friedman, and Wilcoxon signed-rank tests. For the statistical tests, the sample size N was considered 302. However, for comparing specific groups which constitute the majority of the participants, some outlier responses were disregarded when necessary, e.g., to explore differences in the preferences of the male and female participants, only the male and female participants ($N = 293$) were considered.

5.3.3 Participant Demographics

The demographic information about the participants is shown in Table 6. 49% of the participants were male while 48% of them were female. Young adults were dominant among the participants, which was unsurprising due to the dominance of young respondents in Prolific (Francis et al. 2023). Native languages of the participants were diverse, including 28 distinct languages although the most commonly observed ones were among European languages. Almost two-thirds of the participants were holding at least a Bachelor's degree or equivalent whereas one-third of them were high school graduates. However, the majority (58%) of the participants reported having an advanced or professional level of computer knowledge. Besides, most participants (58.9%) reported spending more than 5 hours online per day.

Table 6: The demographics of the 302 survey participants

Characteristics	n	%	**Characteristics**	n	%
Gender			**Education**		
Male	148	49.0	Less than high school or eq.	2	0.7
Female	145	48.0	High school or eq.	99	32.8
Other	7	2.3	Bachelor's degree or eq.	153	50.7
Prefer Not to Say	2	0.7	Master's degree or eq.	38	12.6
			Doctoral degree or eq.	5	1.7
			Other	5	1.7
Age			**Computer Knowledge**		
18-29	204	67.5	Basic	20	6.6
30-49	76	25.2	Intermediate	107	35.4
50-64	22	7.3	Advanced	124	41.1
Over 65	0	0.0	Professional	51	16.9
Native Language			**Time Spent Online Daily**		
English	87	28.8	More than 5 hours	178	58.9
Polish	54	17.9	3-5 hours	83	27.5
Spanish/Castilian	47	15.6	1-3 hours	38	12.6
Portuguese	35	11.6	Less than 1 hour	3	1.0
Hungarian	12	4.0			
Other	67	22.2			

5.4 Findings

5.4.1 Information Source Preferences

The majority of participants (around 90%) reported that they use websites, computers/laptops, and smartphones as platforms to get information more than once

a day. While 64.2% of the participants prefer to use a desktop or laptop computer to get information, 71.2% of them expressed their preferred devices as mobile devices such as smartphones or tablets. Furthermore, over half of them get information from TV and/or radio. Regarding age differences in the frequent usage of platforms, when the participants below 30 were compared to those over 30, smartphones ($\chi^2(1, N = 302) = 8.61$, $p = .003$) were significantly more preferred among the participants below 30, on the contrary to TV/radio ($\chi^2(1, N = 302) = 6.93$, $p = .008$), newspapers ($\chi^2(1, N = 302) = 4.73$, $p = .030$), and tablets ($\chi^2(1, N = 302) = 5.75$, $p = .016$), which were more popular among the older participants.

Regarding regularly followed information sources, video streaming platforms such as YouTube were the most commonly followed sources (76.8%) for all age groups. Then, online forums such as Reddit (57.6%), mainstream news outlets (54%), and online encyclopaedias including Wikipedia (48%) were among the most popular sources to obtain information. However, the preference of some sources correlated to the gender and age of the participants. While female participants preferred news aggregators such as Google News ($\chi^2(1, N = 293) = 6.88$, $p = .009$) and local news outlets ($\chi^2(1, N = 293) = 11.50$, $p < .001$) significantly more compared to male participants, significantly more male participants reported following scientific publications ($\chi^2(1, N = 293) = 4.49$, $p = .034$) and online forums ($\chi^2(1, N = 293) = 18.43$, $p < .001$). Nonetheless, informative social media accounts ($\chi^2(1, N = 302) = 5.94$, $p = .015$) appeared as more preferred among those below 30 while significantly more participants over 30 reported following mainstream news outlets ($\chi^2(1, N = 302) = 12.10$, $p < .001$), independent journalists ($\chi^2(1, N = 302) = 6.52$, $p = .011$), and periodic email newsletters ($\chi^2(1, N = 302) = 4.69$, $p = .030$).

To search for specific information, over 95% of the participants preferred search engines whereas three-fourths used video streaming platforms. Besides, online

encyclopaedias (52%) and social media platforms (47.7%) were commonly used for searching information. Nonetheless, the significant percentage differences in online forums and audio/podcast streaming platforms indicate that only some participants regularly following these sources preferred to use their search functionalities. Regarding demographic differences in searching behaviours, the female participants utilised social media platforms for searching significantly more than the male participants ($\chi^2(1, N = 293) = 7.51$, $p = .007$). Furthermore, the participants below 30 reported using research databases for searching significantly more, compared to the participants over 30 ($\chi^2(1, N = 302) = 6.36$, $p = .012$). Finally, there was a positive correlation between the level of education (i.e., high school or less, Bachelor's degree, or postgraduate degree) and the percentage of those using research databases for searching information ($F(2, N = 297) = 3.71$, $p = .026$). Table 7 provides a breakdown of the information source preferences of the participants.

5.4.2 RQ3.1: Information and Information Providers Users Feel the Need for Fact-checking

Most participants reported encountering false or inaccurate information online at least every week and agreed that it is a serious threat to society. Besides, they stated that they feel the need to fact-check suspicious claims at least every week.

The participants were asked to what extent they perceived several individual and organisational actors as credible when it comes to providing true and accurate information, through a five-point Likert scale. As shown in Figure 14, among organisational actors, the vast majority of the participants (84.8%) considered academic institutions as credible. Furthermore, almost two-thirds of the participants reported that local news outlets provide true and accurate information. For over half of the participants, mainstream news outlets (54.3%), governmental bodies (51%) and non-governmental organisations (NGOs) (56.9%) were credible

Table 7: The participants' preference of information sources when getting and searching for information

Regularly Followed Sources	%	**Sources for Searching**	%
Video streaming platforms	76.8	Search engines	95.4
Online forums	57.6	Video streaming platforms	75.8
Mainstream news outlets	54.0	Online encyclopaedias	52.0
Online encyclopaedias	48.0	Social media platforms	47.7
Informative social media accounts	38.4	Online forums	38.4
Audio/Podcast streaming platforms	33.8	Research databases	34.4
Scientific publications	33.1	Academic institution websites	28.1
Local news outlets	32.1	Government websites	20.9
News aggregators	31.5	Audio/Podcast streaming platforms	17.2
Independent journalists	21.9		
Science/Tech magazines	18.5		
Blogging websites	17.9		
Periodic email newsletters	11.6		

although a considerable portion of them disagreed on this for mainstream news outlets (23.8%) and governmental bodies (20.2%). Finally, only 16.9% of the participants considered social media platforms as credible while 48.7% of them reported the opposite.

When it comes to individual stakeholders, Figure 15 depicts the distribution of levels of trust in different actors. 85.8% of the participants reported that domain experts (e.g., scientists, engineers, or doctors) are credible. Among journalists, independent journalists are perceived as the most credible (61.6%), which is followed by those working for a local news outlet (56.3%), and those working for a mainstream news outlet (47.1%). The vast majority of the participants agreed that politicians (66.6%) and celebrities/influencers (71.8%) are not credible information sources. Besides, only 12.3% of the participants reported that unknown

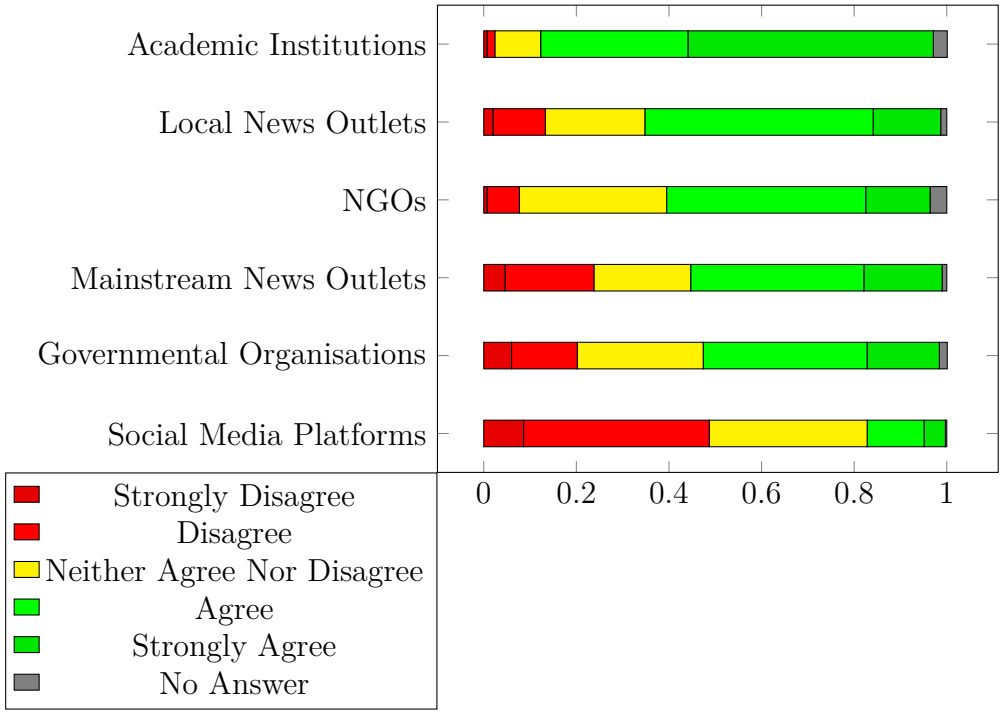


Figure 14: The percentages of participants' agreement/disagreement on the credibility of different organisational actors.

social media users are credible whereas 48.4% of them disagreed. Lastly, the participants seemed unsure about the credibility of activists as well as their family, friends, and acquaintances. While 25.5% of them feel activists credible sources, 30.8% think they are not credible. Regarding the latter, 35.4% of the participants considered them as credible whereas 17.9% expressed the opposite.

5.4.3 RQ3.2: Stakeholders Trusted as Reliable Information Providers for Fact-checking and Fact-checkers

The participants reported different practices for investigating suspicious claims. The most common practices that appeared in the responses were searching for relevant information via search engines (80.8%), checking source credibility (76%), looking for signs showing the website is not a reputable source (e.g., bad web design, text in all caps, excessive punctuation usage) (64.6%), checking the URL

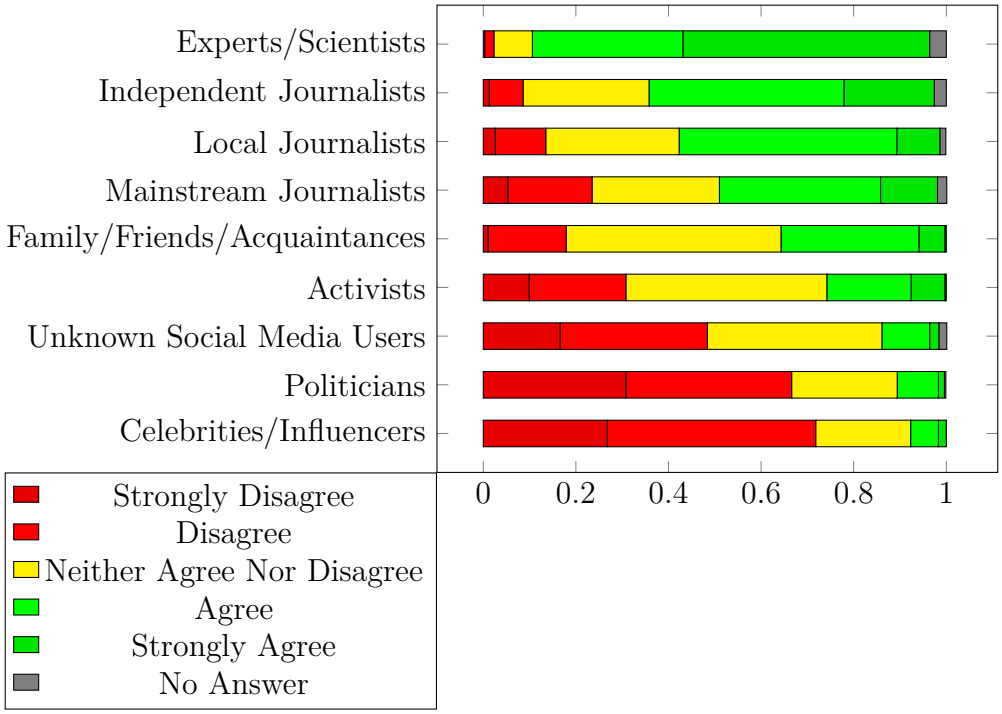


Figure 15: The percentages of participants' agreement/disagreement on the credibility of different individual actors.

to see if it is legitimate (63.2%), and cross-checking from other sources (62.6%). However, seeking previously investigated claims from the websites of fact-checking organisations (25.8%), asking someone trusted (23.8%), and using fact-checking software (6.3%) were not among common practices. This suggests the need for fact-checking tools and services to be improved so that they can become a part of daily lives. Other than the practices provided, the participants mentioned other practices involving checking references and using common sense and logic.

When it comes to relying on external information for fact-checking purposes, the participants were asked what organisational and individual stakeholders they trust via a five-point Likert scale. The responses indicate that academic institutions and fact-checking organisations were perceived as credible organisations by the vast majority, i.e., 85.1% and 61.2% of the participants, respectively. Among

news organisations, local news outlets were treated as credible by 53.6% of the participants whereas this percentage appeared as 51% for mainstream news outlets. Besides, 45.4% of the participants perceive that news associations and NGOs are credible while 40.8% of them trust regulators as credible sources. However, only 15.6% of the participants consider anonymous informative social media accounts as credible. Finally, the participants were divided in terms of the credibility of governmental bodies and tech companies. 42% of them think that governmental bodies are credible, which is disagreed by 27.2% of the participants. For tech companies, 31.8% of the participants reported as credible while 29.8% think the opposite. Figure 16 shows the distribution of the responses for each organisational actor covered.

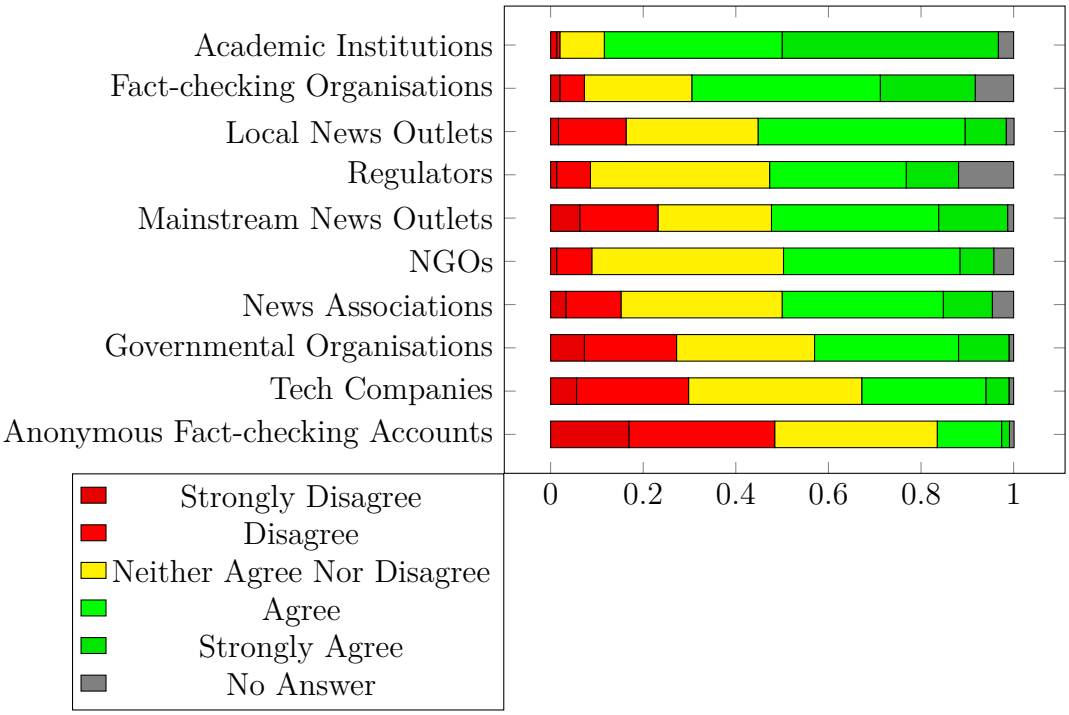


Figure 16: The percentages of participants' agreement/disagreement on the credibility of different organisational actors when fact-checking.

The findings regarding the individual stakeholders were similar, as shown in Figure 17. Domain experts and fact-checkers were reported as credible by the vast

majority, i.e., 85.1% and 66.2%, respectively. They were followed by journalists, including those working for a local (57.6%) or mainstream (48%) news outlet and those working independently (56%). Over one-third of the participants feel their family, friends, and acquaintances are credible sources while over one-fourth of them consider activists credible. The participants mostly agreed that celebrities and influencers (71.8%), politicians (61.6%), and unknown social media users (54.3%) are not credible sources for fact-checking. Interestingly, only 13.9% of the participants reported that they think crowdsourcing workers are credible while 29.8% of them reported the opposite. This might be a cue to explore people's level of trust in crowdsourcing-based fact-checking approaches.

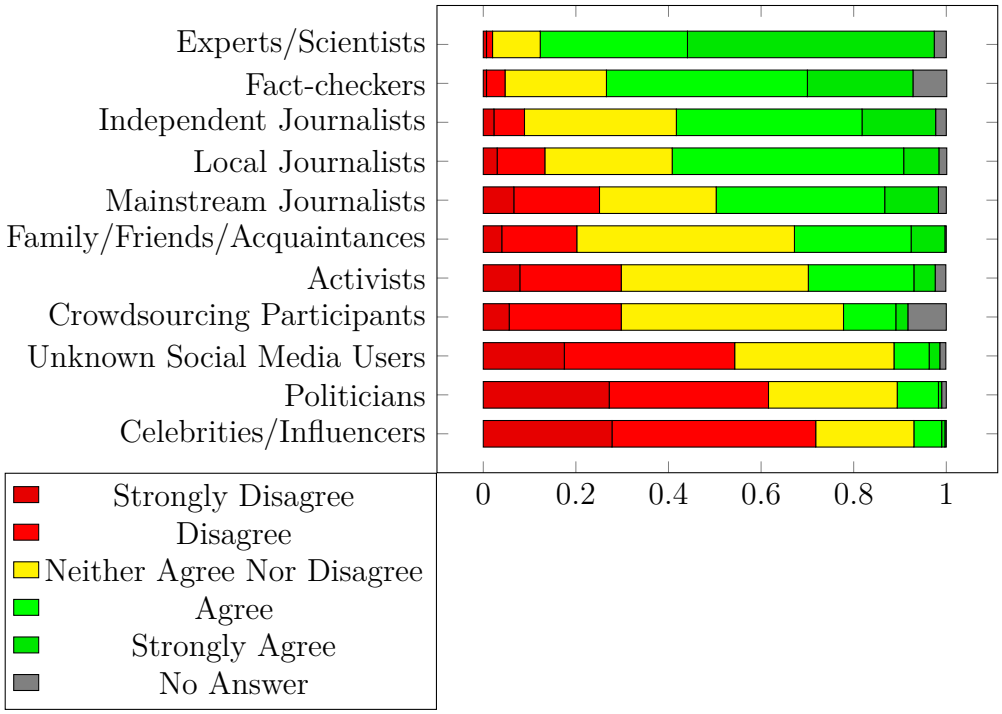


Figure 17: The percentages of participants' agreement/disagreement on the credibility of different individual actors when fact-checking.

5.4.4 RQ3.3: Familiarity with and Usage Patterns of Fact-checking Tools

The participants were provided with the definitions of *fact-checking organisation* (i.e., an organisation dedicated to investigating suspicious claims to verify or debunk) and *fact-checking tool* (i.e., any software, artefact, or resource that can be used to investigate the accuracy of suspicious claims or to seek findings of previous investigations of suspicious claims), and asked whether they heard of any organisation or tool that might answer the given definitions. The majority of the participants have never heard about fact-checking organisations or fact-checking tools – 64.9% and 78.8%, respectively. While 21.9% of the participants could name at least one fact-checking organisation, only 13.2% of them were able to provide at least one fact-checking tool. In addition, 13.2% of the participants were unsure when naming fact-checking organisations whereas this percentage was 7.9% for fact-checking tools. Although they were unsure, many of those participants reported that they came across a fact-checking website, tool, or TV program but never paid attention to their names. These findings suggest that both fact-checking organisations and tools need more visibility.

Fact-checking organisations named by the participants span 40 different organisations, including networks and hubs (e.g., IFCN[7], EDMO[8], and CEDMO[9]) as well as 29 IFCN-verified and 8 non-verified fact-checking organisations. The most commonly mentioned organisations were Snopes[10] (26 participants), FactCheck.org[11] (14 participants), Full Fact[12] (12 participants), PolitiFact[13] (9 participants), and

[7] https://www.poynter.org/ifcn/
[8] https://edmo.eu/
[9] https://cedmohub.eu/
[10] https://www.snopes.com/
[11] https://www.factcheck.org/
[12] https://fullfact.org/
[13] https://www.politifact.com/

BBC Reality Check[14] (8 participants) – all originated from English-speaking countries. It is worth noting that one of the reasons behind this could be that the most frequent native language among the participants is English.

The participants seemed more unaware of specific fact-checking tools. Only a few tools dedicated to fact-checking were named, e.g., InVID, Google Fact Check Explorer, Hoaxy, and ClaimReview. Nevertheless, several tools that can support fact-checking were mentioned, including search engines (e.g., TinEye, Google Reverse Image Search, Google Web, and Bing), research databases (e.g., Google Scholar), content analysis tools (e.g., findExif), generative AI-based services (e.g., ChatGPT and Google Bard), and social media platforms (e.g., Reddit and TikTok).

The responses indicate that the majority of the participants regularly use search engines as well as online written and multimedia materials (e.g., handbooks, video series, visual aids) for fact-checking or following recent fact-checks, at least every month. In addition, most participants reported that they use research databases and content analysis tools (including reverse image search engines) at least once a year. Nevertheless, the majority of the participants have never used social media monitoring tools such as Botometer (77.2%), web browser extensions for fact-checking (76.8%), mobile fact-checking applications (72.8%), messaging applications such as WhatsApp chatbots (64.6%), web-based fact-checking applications (60.9%), content moderation tools on social media (58.3%), or social media accounts of fact-checking organisations (54.6%) for fact-checking purposes. Furthermore, almost half of the participants (49.7%) have never used the websites of fact-checking organisations to follow recent fact-checks. Considering the reported computer knowledge and time spent online daily exhibited in Table 6, these results imply the need for more usable fact-checking tools and services with increased variety and visibility.

[14] `https://www.bbc.com/news/reality_check`

5.4.5 RQ3.4: Trust in Fact-checking Tools with Automation

The participants were asked to rank fact-checking approaches with different automation levels, i.e., human-only, human-AI teaming, and AI-only, in terms of their perceived trust[15]. A Friedman test was carried out to compare the rankings for the three main approaches. The mean ranks (from most trusted to least trusted) obtained from the test indicate that the participants' preference was human-AI teaming approach ($M = 1.47$), human-only approach ($M = 1.97$), and AI-only ($M = 2.56$) approach, respectively, with a significant overall difference ($\chi^2(2, N = 302) = 178,44$, $p < .001$). The participants were in moderate agreement according to the calculated Kendall's coefficient of concordance, $W = .295$. To identify which approaches, in particular, differ from each other, a post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < .017$. The analysis implied a significant difference between each approach, i.e., human-only/human-AI teaming ($Z = -6,789$, $p < .001$), AI-only/human-AI teaming ($Z = -12,358$, $p < .001$), and human-only/AI-only ($Z = -7,055$, $p < .001$). These findings indicate that the participants considered semi-automated fact-checking solutions combining human intelligence and AI as more trustworthy and credible.

To further explore the design cues that would make a fact-checking solution trustworthy for users, the participants were then asked to rate a number of fact-checking tool workflows regarding their perceived credibility. As shown in Figure 18, the provided workflows included different potential components of a typical fact-checking system contributing to different stages of fact-checking. The results

[15]In the questionnaire, the term "AI" was preferred instead of "automation" for the sake of simplicity since AI is a more popular term understandable for a wider audience.

show that most participants found semi-automated solutions involving AI and humans to be credible provided that the humans involved have expertise, e.g., fact-checkers, or are the users themselves. However, considerable distrust appeared for the semi-automated solutions involving crowdsourcing, in line with the previous findings presented in Section 5.4.3 regarding low trust in crowdsourcing workers. Furthermore, the participants indicated their scepticism towards fully automated AI-based systems, validating previous research (Juneja and Mitra 2022; Das et al. 2023).

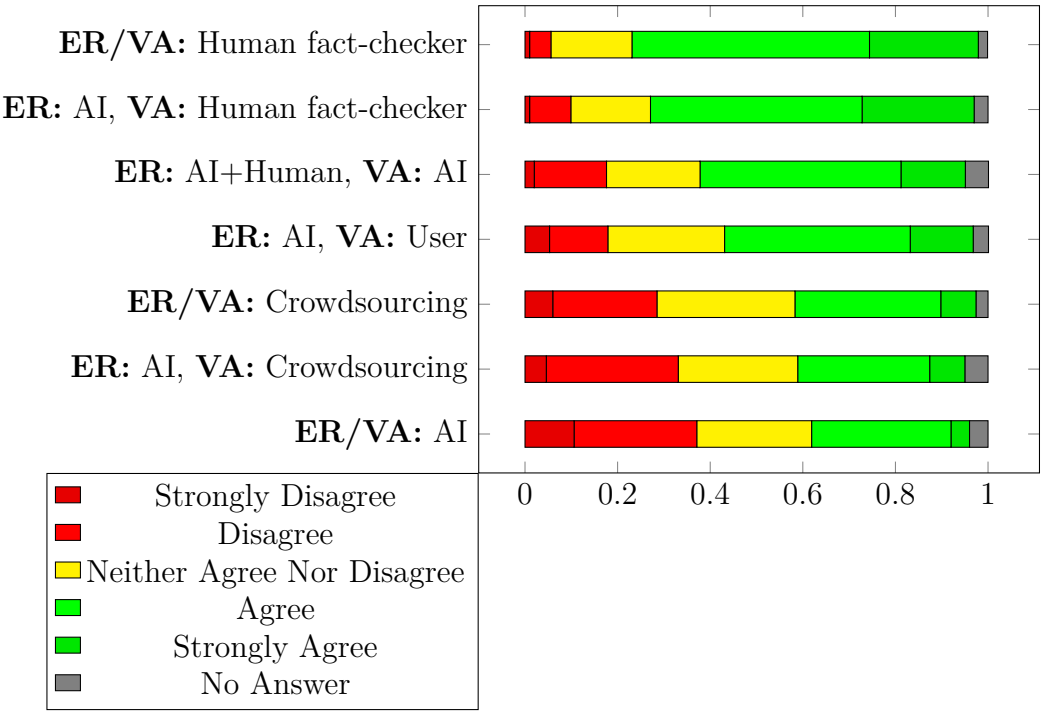


Figure 18: The percentages of participants' agreement/disagreement on the credibility of different fact-checking approaches. (**ER:** Evidence Retrieval, **VA:** Veracity Assessment)

Finally, given the stakeholders perceived by the participants as more credible, they were given some human-AI teaming approaches and asked to rate them regarding their perceived credibility. As depicted in Figure 19, all the given approaches led by the human component were perceived as credible by the vast

majority. However, the most trusted approach appeared as tools supporting humans (fact-checkers or users) with a level of automation to facilitate and accelerate their veracity investigation process. Regarding the automation component of semi-automated solutions, the vast majority of the participants prioritised the automation of certain tasks, including monitoring potential false information online (84.1%), source identification (80.8%) and verification (75.5%), claim determination (78.8%) and prioritisation (75.5%), embedded multimedia verification (77.8%), and evidence retrieval (71.8%). Moreover, veracity assessment (57.2%) with explanations (59.9%) appeared as important to be automated for over half of the participants.

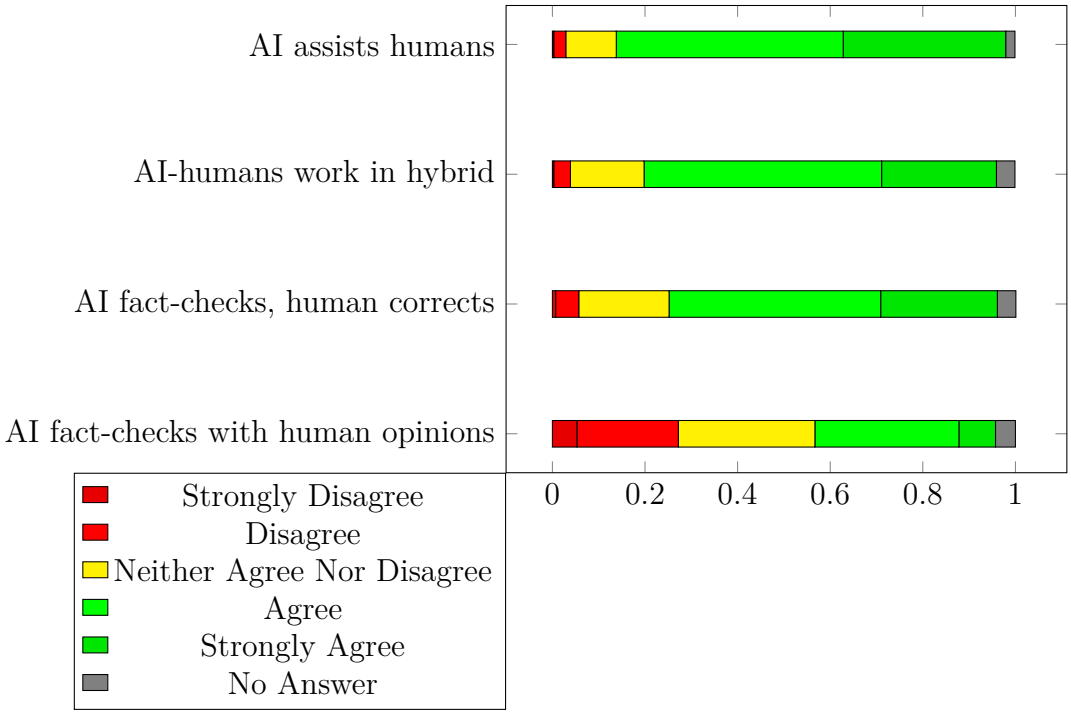


Figure 19: The percentages of participants' agreement/disagreement on the credibility of different semi-automated fact-checking approaches.

5.4.6 RQ3.5: Explainability of Fact-checking Tools

64.9% of the participants reported that it is insufficient for fully automated fact-checking tools to output only veracity labels, e.g., *True* or *False*, without providing any explanation justifying the final verdict. With this respect, for 56% of the participants, explainability appeared as a key factor for a fully automated fact-checking tool to be trustworthy. Regarding increasing explainability, the participants underlined the importance of different design cues that can be considered in fully and semi-automated solutions. For example, 64.9% of the participants reported that such solutions should provide saliency explanations, i.e., highlighted parts of the input showing which words or sentences have more contribution to the reached verdict. Besides, 75.8% of the participants reported that a clear explanation of how a semi-automated system's automated component(s) works is needed to ensure transparency and interpretability, even if the human component reaches the final verdict. Lastly, the vast majority of the participants considered the credibility (76.8%) and transparency (82.1%) of sources used for evidence retrieval quite significant to be exhibited.

5.5 Summary

Considering the EER model of false information and fact-checking ecosystem presented in Chapter 4, this chapter magnified user perspectives on fact-checking and explores human attitudes and behaviours regarding information consumption and verification. More precisely, this chapter explored specific needs for fact-checking tools (which are covered by the *Service* and *Information* entity types in the EER model) to increase their trustworthiness and usability for users.

As important as achieving its technical objectives for a fact-checking solution is its ability to convince users that the solution involves the presented functionality. Thus, especially for the tools containing an automated component which tends to

develop a distrust, taking trustworthiness into account is necessary when designing such solutions for their practical adoption. This involves many considerations in a typical fact-checking process, e.g., utilising trusted sources during evidence retrieval and providing explanations of how the tool reached a verdict. To shed light on how such considerations should be implemented in (semi)-automated fact-checking, this chapter offered an online survey with 302 participants, which revealed some design cues that can help develop more trustworthy fact-checking solutions when followed and guide the design of a new fact-checking tool presented in the next chapter. These cues and their contribution to the design of the new tool can be summarised as follows:

1. **Integrate common information sources.** Search engines and video streaming platforms were two key sources of seeking specific information for the participants. Bearing this in mind, fact-checking tools can facilitate users' searching process by utilising such sources for evidence retrieval, provided that they also offer a filtering mechanism to avoid records from untrusted sources in the search results. Alternatively, information provided by mainstream news outlets and online encyclopedias can be used since these were also key information sources for the majority.

2. **Utilise audiovisual communication.** Recent research indicated a growing interest towards following online news in multimedia rather than reading it (Newman et al. 2023). The survey results provide further evidence as the vast majority reported following video streaming platforms to seek particular information. Hence, fact-checking solutions can leverage visual communication strategies to reach wider audiences.

3. **Consider science communication.** According to the survey results, experts and academic institutions were the most trusted actors in the information ecosystem. On the contrary, the percentage of the participants who

followed relevant information sources, such as research databases, scientific publications, and technology magazines, was much lower. Due to the positive correlation between education level and the use of research databases, this can be explained by the absence of sufficiently effective science communication strategies which can help people with different educational backgrounds access credible information. Therefore, fact-checking tools can consider implementing an enhanced science communication strategy to present scientific information in a more understandable format, e.g., automatic simplification of scientific texts (Ermakova et al. 2023) or using scientific news articles.

4. **Join the forces of human expertise and automation.** In terms of the utilisation of automation, including AI, the most trusted options shown in Figures 18 and 19 suggest two main criteria for a trustworthy solution – human component of the system needs to have sufficient expertise, and the final verdict needs to be reached by the human component. In light of this, solutions leveraging crowdsourcing for scalability should also consider additional mechanisms to resolve trust issues. Regarding the automated component, however, a crucial factor in increasing trustworthiness for users appeared as explainability, which can bring more transparency to fact-checking processes. This can be considered in fact-checking tools by utilising xAI systems, as well as following explainable user interface principles (Purificato, Shahania and De Luca 2022). Besides, when the role of automation is limited to assisting the human component in different tasks (e.g., monitoring, claim determination, or evidence retrieval) or, at least, when its output generated as a result of fact-checking is corrected by the human component, it can develop greater trust.

Another key finding of the presented work is that the vast majority of the participants were unaware of existing fact-checking organisations and tools although

most participants considered fact-checking organisations as credible. This suggests the need for increasing the visibility of such organisations and tools so that they can become more prominent in daily information verification practices. Other than advertising, improving the usability of services provided by fact-checking organisations and the developed tools can also positively impact their visibility.

The presented study in this chapter is not free from limitations. There were no participants over 65, and a vast majority of them were below 30. This implies that the survey may not have been capable of capturing age-based differences in behaviours and attitudes sufficiently. In addition, the sample size of the study was chosen close to the minimum number of participants required, obtained from the power analysis, due to financial concerns. Therefore, the findings of this study can be enhanced by considerably increasing the sample size and applying sample balancing.

Chapter 6

aedFaCT: Expert Discovery-Based Scientific Fact-Checking

Inspired by the findings obtained in the previous chapters, this chapter presents an example application of stakeholder-based semi-automated false information detection. It is based on two published/archived works of the author (Altuncu et al. 2023, 2022).

As discussed in Section 4.1, experts, especially scientific experts, are among the sources of evidence in a typical fact-checking process. They play a crucial role in society by enlightening people with truths and facts through different communication media, e.g., news, social, and printed media. According to the 2022 Ipsos Mori Veracity Index, scientists and professors were among the most trusted professions in Britain (Ipsos MORI 2022).

Especially when there is a chaotic atmosphere impacting most people due to several unknowns, scientific experts attract all the attention in society. They provide expert comments, give interviews to news outlets, and support journalists in

reporting online false information to bridge gaps in their contextual understanding and methodological expertise (McClure Haughey, Povolo and Starbird 2022). To illustrate, in the first days of the COVID-19 pandemic, the public was seeking information on how to protect and recover from COVID-19 by following experts, especially those specialised in virology, through multiple communication channels, especially news media. Until the official organisations came up with some useful guidelines based on the most recent scientific evidence, people were trying to obtain reliable information taking multiple scientific experts into account. This indicates that in society experts are frequently considered as a credible information source and a bridge that brings non-experts and true information together. As a result, experts can be quite useful in scientific fact-checking.

As a crucial element of scientific fact-checking, much effort has been made to engage with experts during information verification. For example, Science Media Centre[1] meets scientific experts with journalists so that the media covers more accurate and evidence-based information. Besides, Meedan, which is a technology not-for-profit organisation, introduced Digital Health Lab[2] to support evidence-based responses to health misinformation with a team of scientists, content moderation experts, and journalists. Other than active engagement with experts, there have also been efforts to develop automatic expert recommendation systems to find credible evidence on a specific news topic (Zhang et al. 2023).

In spite of the efforts for expert-journalist engagement against false information, this collaboration might produce some side effects. For instance, outlier experts who hold a different view than the majority could mislead non-expert people. In addition, the selection of experts with incompatible expertise is likely to misdirect the public (Palmer 2020). Other than these, the challenges in scientific communication dilute expert-journalist collaboration in fact-checking (Bucchi 2017). As a result, one can argue that the best practice for leveraging expert opinions for

[1] https://www.sciencemediacentre.org/

[2] https://meedan.com/programs/digital-health-lab

fact-checking would be to cover the views of multiple experts as evidence. Therefore, this chapter introduces a new web browser extension for Chrome browsers, namely *aedFaCT* (i.e., **a**utomatic **e**xpert **d**iscovery-based **Fa**ct-**C**hecking **T**ool), which aims to assist professional users (e.g., journalists and fact-checkers) and common readers during scientific fact-checking of online news articles. Given a news article, aedFaCT searches for relevant expert opinions in the shape of expert comments covered in news media and peer-reviewed scientific publications by following a semi-automated approach.

6.1 Related Work

6.1.1 Human-Machine Teaming for Fact-checking

Scalable automated fact-checking is a difficult task. Therefore, as broadly mentioned in Section 2.4.3, current research offers hybrid solutions based on human-machine teaming that automate certain steps of fact-checking for fact-checkers and the general public. Nguyen et al. (2018), for instance, proposed a mixed-initiative approach to fact-checking where the system predicts the truthfulness of a claim based on relevant articles with their stance towards the truthfulness of the claim and the reputation of each source. The users' role in this design is to alter each article's source reputation and stance for a more accurate prediction. More recently, Gupta et al. (2021) established an evidence retrieval strategy to look for semantically similar news articles in order to aid users in validating news items. The proposed system leaves the final verdict to the user. Furthermore, La Barbera, Roitero and Mizzaro (2022) suggested a hybrid human-in-the-loop approach that relies on three key components – AI, crowdsourcing, and experts, for the assessment of claim veracity. If any component yields a prediction with a high confidence score, the truthfulness of the claim is taken into consideration as having been accurately classified. If not, the claim is passed on to the following

component. As another example of human-in-the-loop AI systems, *HAMLET* is a conceptual framework that uses AI-expert teaming in a number of fact-checking tasks, including the collection of expert data annotations and expert feedback, AI system performance monitoring, and life cycle management (Bandhakavi, Hoffmann and Lear 2022). Last but not least, Arroyo et al. (2023) introduced a toolbox called *Ms.W* that combines a number of freely accessible services and tools to assist users in fact-checking and determining the credibility of sources.

Several fact-checking tools utilise crowd intelligence at various phases as an additional method of human-machine teaming. Vo and Lee (2018), for example, developed a fact-checking URL recommendation model to encourage guardians – social media users who correct false information by referencing fact-checking URLs – to engage in fact-checking activities more frequently. Additionally, social media platforms enable users to flag posts that include false information and, if there are enough flags, send them to fact-checkers for additional investigation. X (formerly Twitter) recently introduced Community Notes[3] (formerly known as Birdwatch), where users can provide context for tweets to help prevent false information from being spread on the platform.

6.1.2 Web Browser Extensions for Fact-Checking

Web browser extensions are quite useful for fact-checking, especially for web-based documents and articles, since they minimise user effort. For instance, users can automatically verify the accuracy of a news article or a snippet from an article using the tool *BRENDA* (Botnevik, Sakariassen and Setty 2020). It identifies check-worthy claims, classifies them using a deep neural network, and then displays the findings to the user together with the supporting data obtained from the top-10 Google Search results. *FADE* is another automated tool that finds several sources containing the same news item and performs automated fact-checking in

[3] `https://help.twitter.com/en/using-twitter/community-notes`

accordance with the trustworthiness of the news sources and the sources that are cited in the article (Jabiyev et al. 2021). Apart from these, The Factual[4], a tool that automatically rates news items based on a number of factors, including their source quality and bias, author expertise, and tone, provides an alternative to the solutions developed in academia.

Additionally, there are Web browser plugins that assist users in evidence retrieval and content analysis for fact-checking. One such tool is *InVID*, which uses a range of tools to assist users in verifying videos and images (Teyssou 2019). Another example is *News2PubMed*, which, given a news article, retrieves relevant health research papers (Wang and Yu 2021). In order to help users assess the credibility of the source and content, another tool called *News Scan* displays several aspects of the source and content of news articles, such as source popularity, sentiment, objectivity, and bias (Kevin et al. 2018). Finally, NewsGuard[5] shows manually assigned source credibility ratings next to links on search engines and social media platforms.

6.1.3 Key Differences of aedFaCT and Previous Tools

Compared to existing fact-checking tools, aedFaCT is different in several aspects. To begin with, although there have been studies using experts in different stages of fact-checking (La Barbera, Roitero and Mizzaro 2022; Bandhakavi, Hoffmann and Lear 2022; Wang and Yu 2021), to the best of our knowledge, aedFaCT is the first fact-checking tool that is entirely based on expert opinion discovery. A further advantage of aedFaCT is that it retrieves both news articles and scientific papers, making it suitable for both professional and general users. Depending on their degree of knowledge and expertise, this allows users to consider information sources they are more familiar with. In other words, aedFaCT aims to leverage

[4] `https://www.thefactual.com/`
[5] `https://www.newsguardtech.com/`

enhanced science communication for users with different educational backgrounds to transmit trustworthy information to wider audiences. In addition, it provides users with evidence from a variety of sources and experts, which is a beneficial approach for helping users leave their echo chambers. Third, aedFaCT combines human expertise with the speed of automation by considering user input when confirming the extracted keywords and, more importantly, leaving it up to users to interpret the collected evidence to reach a verdict. Because of users' scepticism towards automation (Juneja and Mitra 2022), this can make it easier for aedFaCT to build trust among users than many fact-checking tools that have a completely automated decision-making mechanism and a black-box architecture. This is also supported by the empirical evidence presented in Chapter 5 since most design cues for enhanced trustworthiness mentioned align with the aedFaCT design. Ultimately, aedFaCT's general workflow is consistent with the standard procedures followed by human fact-checkers, wherein engaging with experts is a key element (Juneja and Mitra 2022; Micallef et al. 2022; Procter et al. 2023). This means that aedFaCT can be useful for fact-checkers to accelerate their claim investigation processes. More specifically, Table 8 provides an overview of different characteristics of aedFaCT and other existing tools. In consequence, aedFaCT has the potential to be a part of standard fact-checking processes performed by both human fact-checkers and common readers.

6.2 System Design Overview

aedFaCT focuses on fact-checking online news articles. Therefore, it is designed as a web browser extension for Chrome-based browsers for more usability. In addition, it follows a semi-automated approach which involves automated elements with user feedback to minimise the errors likely to be produced by the automation. More precisely, Figure 20 demonstrates the overall workflow of aedFaCT. The

Table 8: The comparison between aedFaCT and other existing web browser extensions for fact-checking

Tool	**Approach**	**Task**	**Output**	**Domain**
BRENDA	Auto	Veracity Prediction	News articles	Any
FADE	Auto	Veracity Prediction	News articles	Any
The Factual	Auto	Credibility Assessment	Source/Content credibility	Any
InVID	Semi-auto	Content Analysis	Content information	Any
News2PubMed	Auto	Evidence Retrieval	Research papers	Health only
NewsGuard	Auto	Credibility Assessment	Source credibility	Any
News Scan	Semi-auto	Credibility Assessment	Source/Content credibility	Any
aedFaCT	Semi-auto	Evidence Retrieval	News articles Research papers Researchers	Any

system involves two main parts – keyword extraction and selection, and evidence retrieval. Given an online scientific news article, it first extracts some descriptive keywords automatically from the news article. Then, user feedback is received to shortlist the initial list of keywords and cover any missing keywords. The final selection of keywords constitutes a search query to be used for searching. Then, the generated search query is used to search for relevant news articles and peer-reviewed publications. From there, expert comments covered in news media and abstracts of the research papers are extracted and shown to users. The details of each step in the workflow will be covered in the upcoming sections.

6.3 Keyword Extraction and Selection

As briefly discussed in Section 2.3.1, extracting descriptive keywords from text-based documents can be a way of claim determination in fact-checking. With this respect, this section intends to shed light on the current status of automatic keyword extraction (AKE) research to explore the sufficiency of existing algorithms

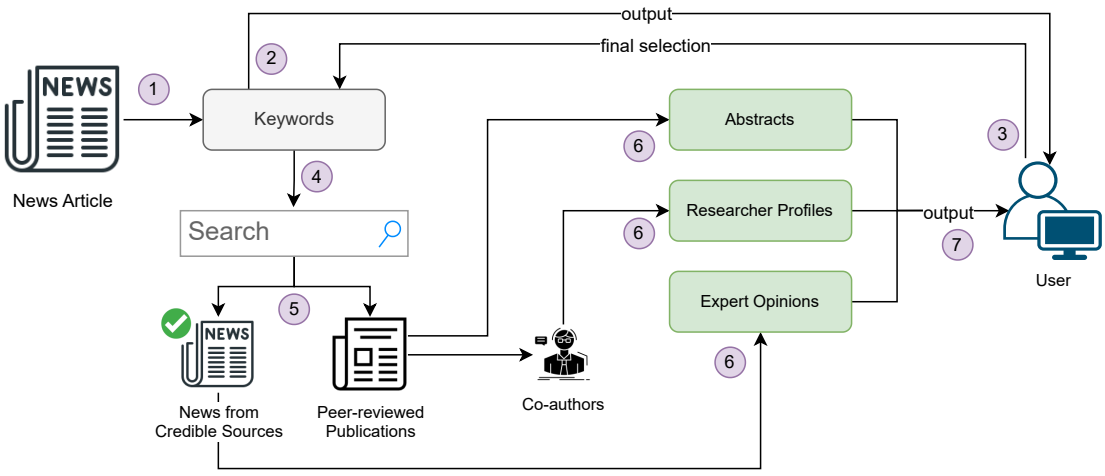


Figure 20: The architecture of aedFaCT

and provide more details on how AKE is incorporated into the implementation of aedFaCT as a claim determination component.

6.3.1 Automatic Keyword Extraction (AKE)

Automatic keyword extraction (AKE) is a useful information retrieval task to obtain some insights about the context of a text document. With the increasing amount of digital textual data processed by digital systems, the AKE task has gained more importance. With this respect, the scientific literature contains a vast amount of studies around AKE, especially through different applications in natural language processing (NLP) and information retrieval (IR). Despite these efforts, AKE has been demonstrated to be a challenging task, and AKE methods with very high performance have yet to be discovered (Papagiannopoulou and Tsoumakas 2020). The absence of a clear definition of the AKE task and the lack of uniform performance evaluation metrics and benchmarks are the two main challenges (Merrouni, Frikh and Ouhbi 2020). Due to the lack of a consensus on the definition and characteristics of a *keyword*, keyword extraction datasets curated by researchers show diverse characteristics, e.g., the minimum/average/maximum

numbers of keywords, if absent keywords (human-annotated keywords that do not exist in the text) are allowed, and what part-of-speech (PoS) tags are considered as valid keywords. This makes comparing and evaluating the effectiveness of AKE approaches more challenging.

AKE techniques described in the literature can be categorised into supervised and unsupervised techniques depending on whether a labelled training set is employed. While supervised approaches utilise either conventional or deep machine learning models, unsupervised methods use statistical, graph-based, embedding-based, and/or language model-based techniques (Papagiannopoulou and Tsoumakas 2020). Unexpectedly, the majority of AKE algorithms have not taken into account semantic information to match the returned keywords with the semantic context of the input content although there exist AKE methods considering how to extract more semantically meaningful keywords through different semantic elements, such as PoS-tags and knowledge bases. For instance, with the usage of PoS-tagging, the retrieved keywords are limited to a predefined set of PoS-tag patterns, such as noun phrases only (Hulth 2003; Pay 2016; Zervanou 2010). Furthermore, some methods make use of external knowledge to give relevant contextual information that helps extract more semantically meaningful keywords. Gazendam, Wartena and Brussee (2010) proposed using semantic relations between thesaurus terms for ranking candidate keywords without a reference corpus, and Li and Wang (2014) proposed an AKE method that benefits from domain knowledge by using author-assigned keywords of scientific publications. To enhance the effectiveness of AKE approaches, thesaurus relations have also been coupled with machine learning techniques (Hulth et al. 2001; Medelyan and Witten 2006). More recently, Sheoran, Jadhav and Sarkar (2022) used domain-specific ontologies for aspect assignment of candidate keywords retrieved from opinionated texts, ensuring that the chosen candidates cover a maximum number of aspects.

Wikipedia, a helpful source of semantic data, has also been used by several

AKE algorithms. Shi et al. (2008) utilised Wikipedia to retrieve semantic information from candidate keywords. Based on the hierarchical relationships derived from Wikipedia, their technique constructs a semantic network connecting candidate keywords to document topics, and semantic feature weights are then assigned to candidate keywords using a link analysis algorithm. Another AKE technique that makes use of Wikipedia is WikiRank (Yu and Ng 2018). It constructs a semantic network by using the TAGME annotator (Ferragina and Scaiella 2010) to connect sensible word sequences in the input document to Wikipedia concepts. In order to find the keyword set that has the best coverage of the identified concepts, it then transforms the keyword extraction task into an optimisation problem on the graph. Ultimately, many embedding-based AKE methods leverage Wikipedia for pre-training and/or fine-tuning their underlying embedding methods (Bennani-Smires et al. 2018a; Papagiannopoulou and Tsoumakas 2018).

Despite the usefulness of AKE in a wide range of applications and the existence of several AKE approaches, existing AKE methods are still far from providing sufficiently high performance (Papagiannopoulou and Tsoumakas 2020). Thus, before being used in fact-checking processes for claim determination, current AKE methods require additional steps for a more satisfying level of accuracy, including post-processing. More precisely, despite the existence of AKE methods leveraging semantic elements, there is still a need for a universal approach to improve existing AKE methods in terms of semantic and morphosyntactic awareness. This can enable benefiting from state-of-the-art AKE methods in claim determination even if they have considered semantic information insufficiently. As a result, this research gap needs to be addressed before integrating a state-of-the-art AKE method into aedFaCT.

6.3.2 Analysis of Ground Truth Keywords

Since the evaluation of AKE methods is currently based on to what extent the generated keywords match with the human-generated ones, a thorough analysis of the ground truth ("golden") keywords generated by human annotators is needed to be able to identify the shortcomings of existing methods and understand how humans generate keywords. Besides, this analysis needs to cover as many datasets as possible since keyword generation is a highly subjective task lacking a standard approach. Hence, 17 English AKE datasets covering multiple contexts (e.g., agriculture, computer science and health) and several types of documents (e.g., scientific papers, news, theses and abstracts) were collected for inspection. Since aedFaCT employs AKE on news articles with generally considerable length, the datasets containing short texts, such as tweets, were excluded. Table 9 shows further details about each dataset.

Observation 1: PoS-Tag Patterns

Although the linguistic properties of different multi-word expression (MWE) types, including collocations (Smadja 1993) and technical terms (Justeson and Katz 1995), are well-known for quite a long time, these are insufficient to explain the linguistic properties of human-generated keywords due to the lack of linguistic standards for keyword generation. Thus, the first step of the analysis was reviewing the distribution of different PoS-tag patterns in the inspected ground-truth keywords. The PoS tags were extracted by using the NLTK library's PoS-tagger. The distribution shown in Table 10 indicates that nine of the top ten PoS-tag patterns correspond to either noun or gerund phrases. The remaining pattern in the table belongs to a single adjective (JJ), with an average percentage of 6.85%, and the top ten patterns cover 80% of all identified patterns.

Table 9: Basic information about the 17 datasets

Dataset	Content	Context	Size	Avg. #(Keys)	Abs. Keys	Annotators[1]
KPCrowd (Marujo et al. 2013)	News	Misc.	500	48.92	13.5%	Readers
citeulike180 (Medelyan, Frank and Witten 2009)	Paper	Misc.	183	18.42	32.2%	Readers
DUC-2001 (Wan and Xiao 2008)	News	Misc.	308	8.1	3.7%	Readers
fao30 (Medelyan and Witten 2008)	Paper	Agr.	30	33.23	41.7%	Experts
fao780 (Medelyan and Witten 2008)	Paper	Agr.	779	8.97	36.1%	Experts
Inspec (Hulth 2003)	Abstract	CS	2,000	14.62	37.7%	Experts
KDD (Das Gollapalli and Caragea 2014)	Abstract	CS	755	5.07	53.2%	Authors
KPTimes (test) (Gallina, Boudin and Daille 2019)	News	Misc.	20,000	5.0	54.7%	Editors
Krapivin2009 (Krapivin, Autaeu and Marchese 2009)	Paper	CS	2,304	6.34	15.3%	Authors
Nguyen2007 (Nguyen and Kan 2007)	Paper	CS	209	11.33	17.8%	Authors & Readers
PubMed (Gay, Kayaalp and Aronson 2005)	Paper	Health	500	15.24	60.2%	Authors
Schutz2008 (Schutz 2008)	Paper	Health	1,231	44.69	13.6%	Authors
SemEval2010 (Kim et al. 2010)	Paper	CS	243	16.47	11.3%	Authors & Readers
SemEval2017 (Augenstein et al. 2017)	Paragr	Misc.	493	18.19	0.0%	Experts & Readers
theses100[2]	Thesis	Misc.	100	7.67	47.6%	Unknown
wiki20 (Medelyan, Witten and Milne 2008)	Report	CS	20	36.50	51.2%	Readers
WWW (Das Gollapalli and Caragea 2014)	Abstracts	CS	1,330	5.80	55.0%	Authors

[1] Experts: Professional indexers assigned for annotation, Readers: People recruited for annotation regardless of their expertise, Authors: The authors of the document annotated
[2] `https://github.com/LIAAD/KeywordExtractor-Datasets#theses100`

Observation 2: n-Gram Size

Many AKE methods include a parameter for the maximum n-gram size, referring to the maximum number of words a generated keyword can have. Although it is quite well-known that bigrams ($n = 2$) and trigrams ($n = 3$) are most common among English MWEs (Choueka 1988), this is unable to explain how human

Table 10: Percentages of top 10 PoS-tag patterns across 17 datasets. PoS tags: NN – noun (singular), NNS – noun (plural), JJ – adjective, VBG – verb gerund.

Dataset	(NN)	(NN NN)	(JJ NN)	(NNS)	(JJ)	(JJ NNS)	(NN NNS)	(JJ NN NN)	(VBG)	(NN NN NN)
KPCrowd	31.38	2.18	3.29	11.65	10.13	0.95	0.95	0.26	5.27	0.17
citeulike180	48.71	7.03	4.78	12.93	12.74	1.61	1.56	0.15	1.95	0.05
DUC-2001	19.13	15.90	15.28	10.49	1.80	8.73	10.16	3.65	0.28	1.52
fao30	32.60	14.68	7.92	15.84	5.06	6.62	9.35	0.00	0.78	0.26
fao780	29.56	14.11	9.11	15.18	3.78	6.02	10.88	0.06	1.21	0.04
Inspec	19.05	12.57	12.49	6.64	3.85	8.11	5.95	4.35	1.11	2.50
KDD	27.93	13.49	9.06	5.89	9.25	5.13	3.55	2.22	4.81	0.76
KPTimes	15.32	16.65	15.67	4.27	2.83	8.62	6.26	2.92	1.76	1.51
Krapivin2009	35.15	4.70	4.06	14.14	5.67	2.17	1.47	0.27	0.95	0.17
Nguyen2007	20.85	19.83	11.31	4.84	2.53	4.79	3.37	3.06	1.51	2.66
PubMed	30.88	9.23	3.87	15.43	12.01	3.51	5.50	0.77	0.56	2.03
Schutz2008	30.15	6.20	10.61	18.63	10.91	5.04	3.19	1.61	0.31	0.66
SemEval2010	19.45	21.74	21.54	0.08	3.20	0.17	0.06	6.40	0.42	3.15
SemEval2017	14.57	8.73	9.00	7.23	2.12	5.95	4.46	3.31	0.66	1.62
theses100	27.88	8.55	5.39	9.48	15.24	6.13	4.28	0.00	1.30	0.19
wiki20	41.91	18.65	11.06	1.49	6.60	0.50	1.82	2.81	2.81	0.99
WWW	32.33	13.44	8.98	5.41	8.74	3.88	3.88	1.63	2.86	1.05
Average (%)	*28.05*	*12.22*	*9.61*	*9.39*	*6.85*	*4.59*	*4.51*	*1.97*	*1.68*	*1.13*

annotators of the 17 datasets consider the n-gram size if they do. Thus, the distribution of the n-gram size across the datasets was computed, as shown in Table 11. The distribution indicates that bigrams ($n = 2$) were the most common with a rate of 45.55% on average while 36.45% of the golden keywords were unigrams ($n = 1$) and 12.73% of them were trigrams ($n = 3$). Moreover, keywords containing more than three words constitute only 5.12% of the golden keywords. These imply that human annotators largely used two or three as the maximum n-gram size, consistent with previous findings in the literature and covering 82.01% and

94.74% of the golden keywords across the datasets, respectively.

Table 11: n-gram distributions of the 17 datasets

Dataset	$n = 1$	$n = 2$	$n = 3$	$n \geq 4$	$n = 1, 2$	$1 \leq n \leq 3$
KPCrowd	**73.78**	18.47	4.90	2.83	92.25	97.15
citeulike180	**77.10**	19.98	2.79	0.09	97.08	99.87
DUC-2001	17.32	**61.29**	17.73	3.66	78.61	96.34
fao30	43.02	**52.74**	3.41	0.83	95.76	99.17
fao780	42.32	**53.72**	3.62	0.34	96.04	99.66
Inspec	16.44	**53.68**	23.05	6.84	70.12	93.17
KDD	25.48	**56.32**	13.97	4.24	81.80	95.77
KPTimes	**46.68**	34.39	12.55	6.38	81.07	93.62
Krapivin2009	18.95	**61.61**	15.74	3.70	80.56	96.30
Nguyen2007	27.53	**49.96**	15.42	6.97	77.49	92.91
PubMed	35.79	**43.74**	15.90	4.58	79.53	95.43
Schutz2008	**57.83**	30.22	8.15	1.67	88.05	96.20
SemEval2010	20.05	**52.97**	20.66	6.31	73.02	93.68
SemEval2017	25.23	**33.74**	17.19	23.84	58.97	76.16
theses100	31.63	**50.37**	11.09	6.90	82.00	93.09
wiki20	26.20	**53.52**	18.17	2.11	79.72	97.89
WWW	34.36	**47.71**	12.15	5.78	82.07	94.22
Average (%)	*36.45*	***45.55***	*12.73*	*5.12*	*82.01*	*94.74*

Observation 3: Semantic Information

The final step of the analysis aimed to discover whether human annotators had explicitly or implicitly relied upon semantic information while generating keywords. For this purpose, the percentage of ground-truth keywords covered by Wikipedia

named entities (i.e., titles of Wikipedia articles) was calculated. As shown in Table 12, 64.39% of the golden keywords were Wikipedia named entities, implying that Wikipedia can be pretty useful as a diverse and rich knowledge base for AKE methods as it covers the majority of human-annotated golden keywords.

Apart from the Wikipedia coverage of golden keywords, it has been observed based on a manual inspection that many collected datasets contain domain-specific keywords. This suggests that considering domain-specific terms can also potentially be useful to improve the performance of AKE methods.

Table 12: The percentages of golden keywords covered by Wikipedia.

Dataset	**%**	**Dataset**	**%**
KPCrowd	71.77	Nguyen2007	52.19
citeulike180	83.78	PubMed	81.28
DUC-2001	51.05	Schutz2008	67.43
fao30	80.97	SemEval2010	41.27
fao780	79.00	SemEval2017	31.02
Inspec	39.08	theses100	68.82
KDD	62.92	wiki20	89.01
KPTimes	79.09	WWW	63.83
Krapivin2009	52.12		

6.3.3 Improving AKE Performance with Post-Processing

AKE algorithms usually involve assigning a numerical score s_i to each keyword candidate w_i to enable ranking and then shortlisting the top n keywords with the highest scores (or with the lowest scores for the AKE methods assigning lower scores for more preferred keywords). Based on this and inspired by the observations mentioned in Section 6.3.2, a three-step post-processing approach that can be applied to any AKE method having such a scoring system was proposed. The approach aims to improve existing AKE methods by re-adjusting the computed

scores so that the scores of true positive keywords increase while the scores of true negative keywords decrease. As shown in Figure 21, the proposed post-processing approach includes the following steps:

1. Filtering out candidate keywords with an unlikely PoS-tag pattern by zeroing its score ($s_i = 0$)

2. Using one or more context-aware (i.e., domain-specific) thesauri to prioritise important candidate keywords for the target domain ($s_i = c_i s_i$, where c_i is an amplifying factor larger than 1)

3. Prioritising candidate keywords covered by Wikipedia as named entities ($s_i = d_i s_i$, where d_i is another amplifying factor larger than 1)

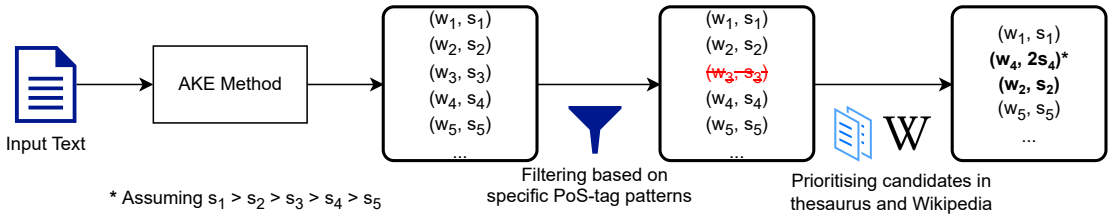


Figure 21: The overview of the proposed post-processing approach.

Filtering Specific PoS-Tag Patterns

PoS-tagging is quite helpful in considering morpho-syntactic features in AKE methods. Based on the observations mentioned in Section 6.3.2, the first post-processing step includes removing candidate keywords corresponding to unlikely PoS-tag patterns by utilising the NLTK library's PoS-tagger and regular expressions for pattern matching. More precisely, the following PoS-tag patterns were identified as the inclusion criteria for candidate keywords:

1. *Simple nouns and noun phrases* – one or more nouns that come after zero or more adjectives

2. Two or more *simple nouns and/or noun phrases* connected by one or more prepositions or conjunctions[6]

3. A single adjective

Note that in the abovementioned PoS-tag patterns, *nouns* and *adjectives* represent any PoS-tags capable of providing the corresponding functionality in a sentence. Hence, *nouns* and *adjectives* also cover gerunds and past participle verbs, respectively. Considering this, the provided inclusion criteria correspond to over 90% of the observed patterns across all the collected datasets mentioned in Section 6.3.2. Nevertheless, the proposed patterns can still be refined to reflect domain-specific needs. For instance, gerunds can be disregarded in the health domain if preferred, as they were observed to be quite uncommon.

Prioritising Domain-Specific Terms

Domain-specific terms are one of the clearest signs of understanding the context of a document. Likewise, many keywords in AKE datasets are relevant to the context of the document, as reported in Section 6.3.2. Therefore, taking domain-specific terms into account can be useful to increase semantic and context awareness in AKE. While such terms can be covered in a more structured vocabulary, such as a thesaurus or an ontology, a simple word list can often be sufficient for AKE. With this respect, this step leverages one or more domain-specific thesauri, containing terms particular to a target context, to prioritise candidate keywords covered by such thesauri. Implementation-wise, candidate keywords to be prioritised were weighted doubly, although the actual weight increase can be a parameter to be determined empirically based on some training data or qualitative evidence observed. To identify the existence of a candidate keyword in a given

[6]Examples include "quality of service" and "buyer and seller" from the SemEval2010 dataset. Although none of the possible PoS-tag patterns conforming to this criterion is among the most common patterns presented in Table 10 individually, they collectively constitute 1.3% of all patterns across the 17 datasets.

thesaurus, exact matching with lemmatisation was applied. Although using stemming with exact matching is a more common practice in AKE (Papagiannopoulou and Tsoumakas 2020), lemmatisation was preferred due to its context awareness. In the experiments, thesauri with a single context were focused, though using multiple contexts in a single thesaurus is also possible. Regarding integrating relevant thesauri, the following two different approaches were experimented with:

1. **Manual Context Consideration:** This approach can be more effective when the context of the input documents to be processed by an AKE method is known. It leverages one or more thesauri containing a list of terms relevant to the context to assign a higher weight to them for prioritisation by the AKE method.

2. **Automatic Context Identification:** In terms of usability, manually providing a thesaurus for each input document may not be ideal. Therefore, this approach aims to identify the context of an input document and a corresponding thesaurus automatically. This can be achieved with a context classifier, predicting the context or a context-specific thesaurus of an input document. Once the classifier predicts the context, the corresponding thesaurus is identified through a context-to-thesaurus look-up table. Since this automatic approach enables to use a different thesaurus for each document, it is more applicable to real-world scenarios where the processed documents can belong to multiple contexts.

Prioritising Wikipedia Named Entities

Based on the last observation mentioned in Section 6.3.2 indicating how widely Wikipedia covers golden keywords in the collected datasets, this step considers Wikipedia as a *context-independent* thesaurus to be utilised similarly to the previous step. With this respect, candidate keywords covered by Wikipedia as an

entry were prioritised by increasing their weights. To determine if a candidate keyword is a Wikipedia named entity, exact matching with lemmatisation was applied. Furthermore, as Wikipedia is an online encyclopaedia also containing a vast amount of entries with general semantic meanings (e.g., unigrams such as 'father', 'school' and 'table' which are less likely to represent a specific context), the NLTK's *words* corpus including common English dictionary words was leveraged to remove such unigrams from the retrieved list of Wikipedia entities. For the Wikipedia named entities, the 2021-10-01 version of the English Wikipedia dump[7] was used. The dump data was initially cleaned by removing the disambiguation tags[8] added next to the Wikipedia page title. Then, the data was normalised with lemmatisation and lower-casing by following the common practice.

6.3.4 Experiments and Results

For the evaluation of the proposed post-processing methods, the collected 17 datasets were used as the test dataset. The evaluation metrics employed were precision, recall, and F1 score at the top ten keywords, which have been commonly used in AKE evaluation (Papagiannopoulou and Tsoumakas 2020). Micro-averaging and exact matching with stemming were adopted when calculating the scores. Note that the obtained scores were consistently too low for some of the datasets due to their characteristics shown in Table 9. For instance, PubMed has an absent keyword percentage of 60.2%, which makes achieving higher accuracy challenging for AKE methods that only consider keyword candidates mentioned in the input text. In addition, for the datasets having an average number of keywords much higher than ten (e.g., KPCrowd and Schutz2008), recall scores at the top ten keywords were unsurprisingly too low, affecting the F1 scores as well. In consequence, the main focus of the evaluation is exploring the improvement rates

[7] `https://dumps.wikimedia.org/enwiki/20211001/`

[8] `https://en.wikipedia.org/wiki/Wikipedia:Disambiguation#Naming_the_disambiguation_page`

of the selected AKE methods under different settings and identifying the optimum setting, rather than seeking high accuracy scores.

To demonstrate the effectiveness and generalisability of the proposed methods and identify the optimum setting (i.e., the selection of an existing AKE method with one or more post-processing methods) to maximise aedFaCT's claim determination accuracy, some representative AKE algorithms with diverse characteristics were determined for experimentation. Existing AKE algorithms were reviewed based on various factors such as recency, ease of reconfiguration, and utilisation of the proposed methods by any means. The selection of algorithms is presented in Table 13. These methods were chosen for their having open-source implementations, applicability to any document type, being validated on multiple datasets, and absence of training requirements (i.e., unsupervised so that it is more convenient to use and less likely to have generalisation problems). Among the chosen methods, two statistical methods (KP-Miner and YAKE!), two graph-based methods (RaKUn and LexRank), and one embedding-based method (SIFRank+) were used as baseline methods. Due to computational constraints, SIFRank+ was evaluated on a subset of datasets with shorter documents, i.e., KPCrowd, DUC-2001, Inspec, KDD, KPTimes, SemEval2017, and WWW.

For the implementations of the selected algorithms, the PKE (Boudin 2016) library was utilised for KP-Miner while the original implementations of the other four methods were employed. Default parameters were used for all methods, except for the maximum n-gram size parameter, which was set to 3 based on the observation that n-gram sizes were mostly limited to 3 (as mentioned in Section 6.3.2).

PoS-Tag Patterns

The proposed PoS-tagging-based post-processing approach was applied to the selected AKE methods, and the evaluation was performed on all datasets. Results

Table 13: An overview of some existing open-source unsupervised AKE methods, showing a number of key characteristics.

Method	Easy to Reconfigure	PoS-tagging	Thesaurus	Wikipedia
Statistical Methods				
KP-Miner (El-Beltagy and Rafea 2009)	✓	–	–	–
YAKE! (Campos et al. 2020)	✓	–	–	–
LexSpec (Ushio, Liberatore and Camacho-Collados 2021)	✓	✓	–	–
Graph-based Methods				
TextRank (Mihalcea and Tarau 2004)	✓	✓	–	–
SingleRank (Wan and Xiao 2008)	✓	✓	–	–
RAKE (Rose et al. 2010)	✓	–	–	–
RaKUn (Škrlj, Repar and Pollak 2019)	✓	–	–	–
LexRank (Ushio, Liberatore and Camacho-Collados 2021)	✓	✓	–	–
TFIDFRank (Ushio, Liberatore and Camacho-Collados 2021)	✓	✓	–	–
Embeddings-based Methods				
EmbedRank (Bennani-Smires et al. 2018b)	✓	✓	–	✓
SIFRank (Sun et al. 2020)	✓	✓	–	✓
SIFRank+ (Sun et al. 2020)	✓	✓	–	✓
MDERank (Zhang et al. 2022)	–	✓	–	✓

showed that the proposed approach improved all methods on average, except for SIFRank+, in terms of precision, recall, and F1 score. KP-Miner showed better performance for 14 out of 17 datasets, with an average F1 score improvement of 6.08% while RaKUn exhibited a cross-dataset average F1 score improvement of 4.46% for 14 out of 17 datasets. YAKE! demonstrated the most significant improvement, with an average F1 score increase of 18.05% for 16 out of 17 datasets. This is possibly because YAKE! does not benefit from linguistic features as a more language-independent (multilingual) approach. Finally, LexRank showed a limited improvement (0.84% on average) for 12 out of 17 datasets, while SIFRank+ experienced a slight decrease in performance, likely due to their preexisting PoS-tagging-based filtering. As examples, detailed results for YAKE! and SIFRank+ can be found in Tables 14 and 15, respectively. These results provide further

evidence of the effectiveness of PoS-tagging in AKE algorithms and indicate the potential for further improvements in the utilisation of PoS-tagging in various AKE methods.

Table 14: Comparison of the precision, recall, and F1 score of the original YAKE! and the one utilising PoS-tagging, at 10 extracted keywords

Dataset	**YAKE!**			**YAKE!+PoS**		
	P%	**R%**	**F1%**	**P%**	**R%**	**F1%**
KPCrowd	24.20	4.92	8.17	**33.98**	**6.90**	**11.47**
citeulike180	23.11	13.27	16.86	**25.68**	**14.74**	**18.73**
DUC-2001	12.01	14.87	13.29	**17.44**	**21.58**	**19.29**
fao30	22.00	6.83	10.42	**25.33**	**7.86**	**12.00**
fao780	11.93	14.95	13.27	**13.18**	**16.52**	**14.67**
Inspec	19.82	14.05	16.44	**24.57**	**17.41**	**20.38**
KDD	**6.01**	**14.68**	**8.53**	5.83	14.23	8.27
KPTimes	7.97	15.83	10.61	**11.37**	**22.58**	**15.12**
Krapivin2009	9.54	17.88	12.44	**9.93**	**18.61**	**12.95**
Nguyen2007	19.00	15.82	17.26	**19.19**	**15.98**	**17.43**
PubMed	7.28	5.11	6.01	**8.66**	**6.08**	**7.15**
Schutz2008	37.29	8.06	13.26	**47.63**	**10.30**	**16.93**
SemEval2010	20.37	13.08	15.93	**20.82**	**13.37**	**16.28**
SemEval2017	20.61	11.91	15.10	**29.41**	**17.00**	**21.55**
theses100	9.40	14.09	11.28	**10.50**	**15.74**	**12.60**
wiki20	19.50	5.49	8.57	**22.00**	**6.20**	**9.67**
WWW	6.49	13.47	8.76	**6.58**	**13.66**	**8.88**
Avg. Score (%)	16.27	12.02	12.13	**19.54**	**14.04**	**14.32**
Improvement (%)				20.10	16.81	18.05

Additionally, the impact of tailoring PoS-tag patterns according to domain-specific needs on AKE method performance was investigated. The example discussed in Section 6.3.3 was considered, which revealed that gerunds are rarely used as keywords in the health domain. The health datasets (PubMed and Schutz2008)

Table 15: Comparison of the precision, recall, and F1 score of the original SIFRank+ and the one utilising PoS-tagging, at 10 extracted keywords

Dataset	**SIFRank+**			**SIFRank+ + PoS**		
	P%	**R%**	**F1%**	**P%**	**R%**	**F1%**
KPCrowd	26.08	5.30	8.81	**26.20**	**5.32**	**8.85**
DUC-2001	**28.34**	**35.09**	**31.36**	27.86	34.49	30.82
Inspec	**35.68**	**25.29**	**29.60**	35.10	24.88	29.12
KDD	**5.68**	**13.87**	**8.06**	4.42	10.80	6.28
KPTimes	**7.92**	**15.74**	**10.54**	7.74	15.37	10.30
SemEval2017	**41.66**	**24.08**	**30.52**	40.16	23.21	29.42
WWW	**6.59**	**13.69**	**8.90**	5.26	10.93	7.10
Avg. Score (%)	**21.71**	**19.01**	**18.26**	20.96	17.86	17.41
Improvement (%)				-3.45	-6.05	-4.65

from the collection were selected for this experiment, and a tailored PoS-tag-based filtering approach that disregards gerunds was applied. YAKE! was used for this experiment, as it is a language-independent algorithm that is more sensitive to linguistic-based improvements. Table 16 shows that the tailored filtering approach provided slight improvements in precision, recall, and F1 score compared to the original filtering proposal. The limited improvement can be attributed to the low percentage of gerunds as candidate keywords.

Domain-Specific Terms

Ten datasets mentioned in Section 6.3.2 with specific contexts were selected for this step, including agriculture (fao30 and fao780), health (PubMed), and computer science (Inspec, Krapivin2009, Nguyen2007, SemEval2010, KDD, Wiki20, and WWW). Additionally, a context-specific dataset of 3,258 samples, KPTimes-Econ, was constructed by extracting economy-related news articles from the KPTimes dataset. This extraction task includes considering the records that involve the

Table 16: Comparison of the precision, recall, and F1 score of YAKE! when the original (PoS) and the tailored (PoS*) filtering approaches are used, at 10 extracted keywords

Dataset	**YAKE!+PoS**			**YAKE!+PoS***		
	P%	**R%**	**F1%**	**P%**	**R%**	**F1%**
PubMed	8.66	6.08	7.15	**8.70**	**6.11**	**7.18**
Schutz2008	47.63	10.30	16.93	**47.80**	**10.34**	**17.00**
Avg. Score (%)	28.15	8.19	12.04	**28.25**	**8.23**	**12.09**
Improvement (%)				0.36	0.49	0.42

term "economy" in the *keyword* and/or *categories* field(s). Then, a thesaurus (or something similar, e.g., dictionary, ontology, or wordlist) specific to each context was collected based on these 11 datasets. More precisely, the following thesauri were used:

1. AGROVOC 2021-07 (Caracciolo et al. 2013): A multilingual controlled vocabulary constructed by the Food and Agriculture Organization of the United Nations (FAO), with 844,000 agriculture-related terms including 50,163 English ones

2. Medical Subject Headings (MeSH) 2021 (Lipscomb 2000): A thesaurus covering biomedical and health-related terms produced by the National Library of Medicine (NLM), with over 1.4 million terms in English

3. Computer Science Ontology (CSO) v3.3 (Salatino et al. 2018): A large-scale computer science ontology automatically produced by Klink-2 (Osborne and Motta 2015) algorithm from 16 million computer science publications, with 14,000 terms

4. STW v9.10 (Kempf and Neubert 2016): A bilingual thesaurus (in English and German) for economics produced by the Leibniz Information Center for

Economics (ZBW), with over 20,000 terms including 6,217 English ones.

For the initial step aiming to experiment with manual integration, each of the baseline methods was fed with each of the datasets and their corresponding thesaurus depending on the context. As in the previous experiment, SIFRank+ was evaluated on only the datasets with shorter documents, i.e., Inspec, KDD, WWW, and KP-Times-Econ in this case. According to the experiment results, the manual integration of context-aware thesauri improved the precision, recall, and F1 score of all five AKE methods significantly across all datasets. The average improvement in F1 score was observed as 29.03% for RaKUn, 23.88% for LexRank, 12.85% for YAKE! 13.19% for KP-Miner, and 7.09% for SIFRank+. As examples, detailed results for LexRank and SIFRank+ can be found in Tables 17 and 18, respectively. These experiments provide strong evidence for the effectiveness of utilising context-aware thesauri in improving the performance of AKE methods.

In the subsequent step, the automated thesauri integration process was experimented with. Focusing on the datasets containing mainly scientific papers, a classifier was trained on a dataset containing metadata of over 1.7M arXiv.org preprints[9] to classify a given article's title and abstract into their respective disciplines. Before the training stage, the arXiv.org dataset was filtered by the main discipline reflected by its *categories* field so as to include only the following three disciplines:

1. cs (Computer Science, e.g., cs.AI)
2. bio (Biology, e.g., q-bio)
3. fin (Finance, e.g., q-fin.CP) and econ (Economics, e.g., econ.EM)

After this filtering, a dataset of 583,796 samples (551,443 computer science, 20,110 biology, and 12,243 finance/economics samples) was obtained. As the resulting dataset is highly imbalanced, random downsampling was applied to equate

[9] `https://www.kaggle.com/Cornell-University/arxiv`

Table 17: Comparison of precision, recall, and F1 score of the original LexRank and its enhanced versions with manual (M) and automatic (A) thesaurus integration, at 10 extracted keywords

Dataset	Context	LexRank			LexRank+T (M)			LexRank+T (A)		
		P%	**R%**	**F1%**	**P%**	**R%**	**F1%**	**P%**	**R%**	**F1%**
fao30	Agr.	20.33	6.31	9.63	**30.33**	**9.41**	**14.36**	—	—	—
fao780	Agr.	8.55	10.72	9.51	**13.04**	**16.35**	**14.51**	—	—	—
Inspec	CS	30.49	21.61	25.29	**31.10**	**22.04**	**25.79**	30.97	21.95	25.69
KDD	CS	6.07	14.81	8.61	6.23	15.20	8.83	**6.25**	**15.26**	**8.87**
Krapivin2009	CS	7.01	13.14	9.15	**8.79**	**16.48**	**11.47**	8.74	16.37	11.39
Nguyen2007	CS	13.25	11.04	12.04	**15.69**	**13.07**	**14.26**	15.45	12.87	14.04
SemEval2010	CS	13.13	8.43	10.27	**15.10**	**9.70**	**11.81**	**15.10**	**9.70**	**11.81**
wiki20	CS	14.00	3.94	6.15	**23.00**	**6.48**	**10.11**	**23.00**	**6.48**	**10.11**
WWW	CS	6.66	13.83	8.99	**6.95**	**14.43**	**9.38**	6.93	14.40	9.36
PubMed	Health	4.22	2.96	3.48	**8.98**	**6.31**	**7.41**	8.92	6.26	7.36
Schutz2008	Health	28.32	6.12	10.07	**34.35**	**7.43**	**12.21**	34.00	7.35	12.09
KPTimes-Econ	Econ.	3.27	7.03	4.46	**4.09**	**8.80**	**5.59**	4.09	8.79	5.58
Avg. Score (%)		12.94	9.99	9.80	**16.47**	**12.14**	**12.14**	15.35	11.94	11.63
Improvement (%)					27.28	21.52	23.88	21.44	16.03	18.07

the number of samples from each discipline to the size of the smallest class, 12,243, which made the final size of the training set 36,729. The classifier built utilised the TF-IDF vectoriser for feature extraction. Besides, the calibrated linear support vector classifier (SVC) with the default parameters and the one-vs-rest setting was preferred, rather than a multi-class classification method or more advanced feature extraction methods such as BERT, to demonstrate the sufficiency of a lightweight classifier for the automatic context detection task. The classifier was evaluated with a stratified 5-fold cross-validation. The testing accuracies[10] of computer science, biology and finance/economics models were 93.2%, 94.9%, and

[10]The fraction of the number of correct predictions with respect to the total number of predictions

Table 18: Comparison of precision, recall, and F1 score of the original SIFRank+ and its enhanced versions with manual (M) and automatic (A) thesaurus integration, at 10 extracted keywords

Dataset	Context	SIFRank+			SIFRank+ + T (M)			SIFRank+ + T (A)		
		P%	**R%**	**F1%**	**P%**	**R%**	**F1%**	**P%**	**R%**	**F1%**
Inspec	CS	35.68	25.29	29.60	**36.62**	**25.95**	**30.37**	36.03	25.53	29.88
KDD	CS	5.68	13.87	8.06	**5.97**	**14.58**	**8.48**	5.95	14.52	8.44
WWW	CS	6.59	13.69	8.90	**7.32**	**15.19**	**9.88**	7.27	15.10	9.81
KPTimes-Econ	Econ.	3.49	7.50	4.76	**4.56**	**9.81**	**6.23**	**4.56**	**9.81**	**6.23**
Avg. Score (%)		12.86	15.09	12.83	**13.62**	**16.38**	**13.74**	13.45	16.24	13.59
Improvement (%)					5.91	8.55	7.09	4.59	7.62	5.92

97.0%, respectively. The classifier can also be extended to support multiple contexts for a single article, although it was not considered in this experiment for the sake of simplicity and clarity. To implement all of the mentioned components, the Scikit-learn library (Pedregosa et al. 2011) was used.

Since the training set of the classifier does not cover agriculture preprints, and a proper agriculture dataset could not be found for training, the agriculture context and the corresponding datasets, fao30 and fao780, were excluded for this part of the experiments. According to the experiment results, the automatic thesaurus integration approach provided similar performance to the manual integration approach in spite of a negligible performance decrease. More precisely, the F1 score was improved by an average of 23.23% for RaKUn, 18.07% for LexRank, 9.60% for YAKE! 11.27% for KP-Miner, and 5.92% for SIFRank+, compared to the baseline scores. As examples, more detailed results for LexRank and SIFRank+ are shown in Table 17 and Table 18, respectively. The obtained results suggest that automatic integration can be generalised to cover more contexts and thesauri, which can be quite useful in real-world AKE applications.

Wikipedia Named Entities

This part of the experiments covers the entire collection of datasets. The findings indicated that the performance of KP-Miner and RaKUn was improved for 16 of the datasets while the performance of YAKE! and LexRank was improved for all the datasets, in terms of all the evaluation metrics. To be more precise, the average improvement rates of the F1 score were observed as 18.83% for RaKUn, 11.11% for LexRank, 10.96% for YAKE! and 10.11% for KP-Miner. Nevertheless, a slight decrease was observed in the average F1 score of SIFRank+, although it provided better performance for most (5 out of 7) of the datasets, which can be attributed to its underlying sentence embedding approach, SIF (Arora, Liang and Ma 2017), already leveraging Wikipedia for pre-training and fine-tuning. Table 19 and Table 20 show more elaborative results for RaKUn and SIFRank+ as examples.

Combining Post-Processing Steps

The final part of the experiments involves combining multiple post-processing steps to examine the possibility of further performance improvement. In this manner, different combinations of the proposed enhancements were applied, and a heatmap showing the average improvements in F1 scores as well as percentages of the improved cases was generated for each AKE method. As can be seen in Figure 22, the results indicate that applying all the proposed post-processing enhancements produced the best F1 scores for YAKE!, RaKUn, and KP-Miner. However, the best combination for LexRank and SIFRank+ was integrating context-aware thesaurus and Wikipedia since they already benefited from PoS-tagging-based filtering. Besides, the applied post-processing enhancements significantly improved the baselines, with improvement rates as high as 23.7% for YAKE!, 21.3% for KP-Miner, 53.8% for RaKUn, 20.1% for LexRank, and 10.2% for SIFRank+.

Table 19: Comparison of precision, recall, and F1 score of the original RaKUn and its enhanced versions with Wikipedia, at 10 extracted keywords

Dataset	**RaKUn**			**RaKUn+Wiki**		
	P%	**R%**	**F1%**	**P%**	**R%**	**F1%**
KPCrowd	42.52	8.64	14.36	**42.64**	**8.66**	**14.40**
citeulike180	16.56	9.50	12.08	**17.92**	**10.29**	**13.07**
DUC-2001	5.68	7.03	6.29	**6.17**	**7.64**	**6.82**
fao30	15.00	4.65	7.10	**18.67**	**5.79**	**8.84**
fao780	6.50	8.14	7.23	**7.64**	**9.57**	**8.50**
Inspec	6.54	4.64	5.43	**6.74**	**4.77**	**5.59**
KDD	**3.66**	**8.92**	**5.19**	3.63	8.86	5.15
KPTimes	8.07	16.03	10.74	**8.15**	**16.18**	**10.84**
Krapivin2009	2.77	5.20	3.62	**4.94**	**9.26**	**6.44**
Nguyen2007	6.79	5.66	6.17	**9.67**	**8.05**	**8.78**
PubMed	4.30	3.02	3.55	**6.58**	**4.62**	**5.43**
Schutz2008	33.14	7.16	11.78	**40.09**	**8.67**	**14.25**
SemEval2010	6.75	4.33	5.28	**10.04**	**6.45**	**7.85**
SemEval2017	11.42	6.60	8.37	**11.74**	**6.79**	**8.60**
theses100	3.90	5.85	4.68	**4.80**	**7.20**	**5.76**
wiki20	9.50	2.68	4.18	**19.50**	**5.49**	**8.57**
WWW	4.32	8.98	5.84	**4.39**	**9.12**	**5.93**
Avg. Score (%)	11.02	6.88	7.17	**13.14**	**8.08**	**8.52**
Improvement (%)				19.24	17.44	18.83

Importantly, these improvements were consistently observed, as at least one combination of the post-processing enhancements resulted in better performance for each method across all the datasets. These results clearly demonstrate that even for more recent AKE methods, there is still ample opportunity for improvement by employing straightforward post-processing techniques such as those proposed in this thesis.

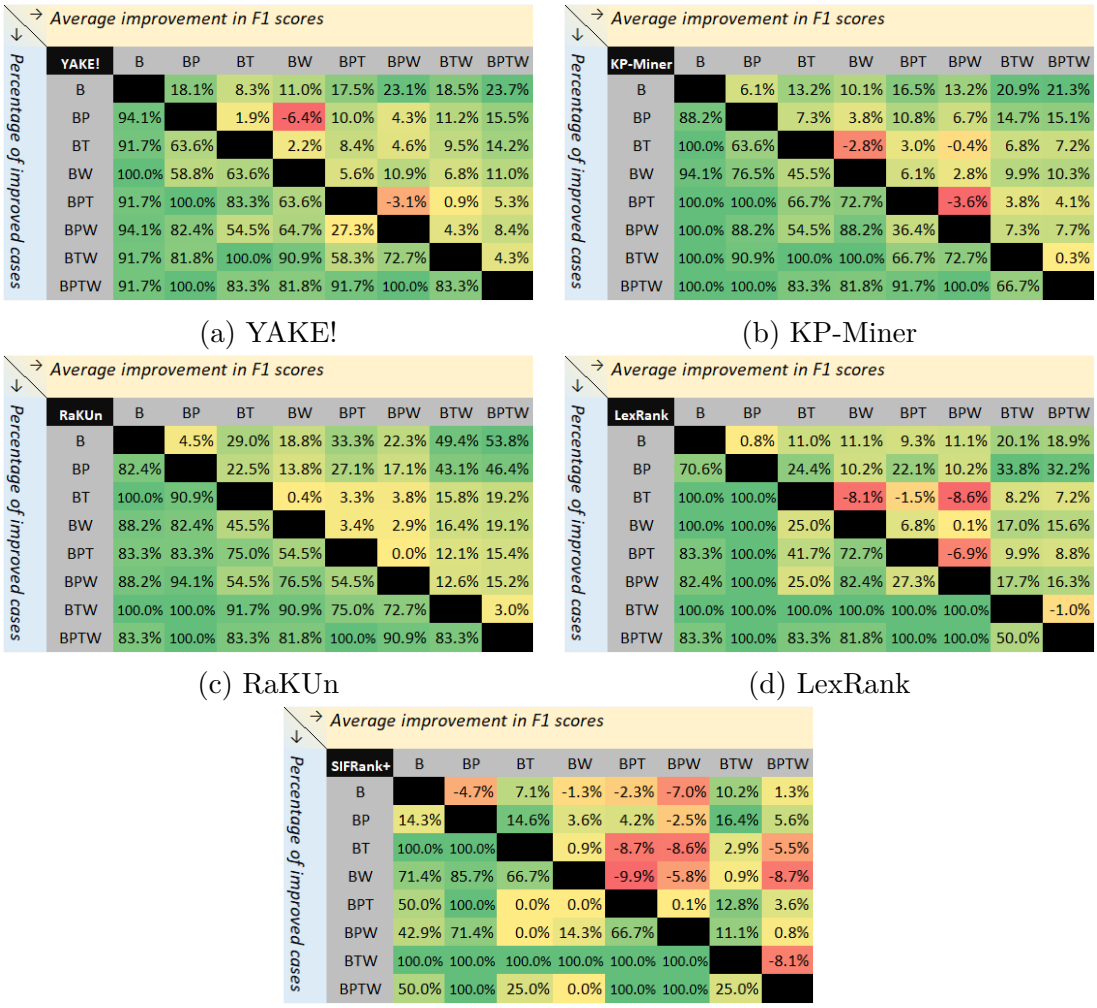
Table 20: Comparison of the precision, recall, and F1 score of the original SIFRank+ and the one utilising Wikipedia named entities, at 10 extracted keywords

Dataset	SIFRank+			SIFRank+ + Wiki		
	P%	**R%**	**F1%**	**P%**	**R%**	**F1%**
KPCrowd	26.08	5.30	8.81	**27.46**	**5.58**	**9.27**
DUC-2001	**28.34**	**35.09**	**31.36**	22.82	28.26	25.25
Inspec	35.68	25.29	29.60	**36.60**	**25.94**	**30.36**
KDD	5.68	13.87	8.06	**6.11**	**14.90**	**8.66**
KPTimes	7.92	15.74	10.54	**9.22**	**18.31**	**12.26**
SemEval2017	**41.66**	**24.08**	**30.52**	41.34	23.89	30.28
WWW	6.59	13.69	8.90	**7.50**	**15.57**	**10.12**
Avg. Score (%)	**21.71**	**19.01**	**18.26**	21.58	18.92	18.03
Improvement (%)				-0.60	-0.47	-1.26

6.3.5 Keyword Extraction and Selection in aedFaCT

Based on the findings mentioned in the previous subsections, aedFaCT leverages AKE with a human-in-the-loop setting to learn the context of the given news article. The choice of a human-in-the-loop setting provides a more controlled environment for users and helps avoid topic drift while the extracted keywords are used in searching. More precisely, the keyword extraction and selection process starts with fetching and parsing the input news article with the help of the Newspaper3k[11] library. Then, AKE is performed by employing SIFRank+ due to its superior performance over the other state-of-the-art AKE methods mentioned in the previous subsections. Based on the experiment results presented in Figure 22, SIFRank+ is enhanced with context-aware thesauri and Wikipedia integration. Other than improving the AKE accuracy, these enhancements are to capture more contextual and domain-specific keywords that are more likely

[11] `https://newspaper.readthedocs.io/en/latest/`



(a) YAKE!

(b) KP-Miner

(c) RaKUn

(d) LexRank

(e) SIFRank+

Figure 22: Average improvements in F1 scores across all the datasets (upper side), and percentages of the improved cases across all the datasets (bottom side), for different AKE methods. (B: Baseline, P: PoS-tagging, T: Thesaurus integration, W: Wikipedia integration)

to limit the search results in the evidence retrieval stage, minimising the risk of aedFaCT returning too general results. After obtaining an initial list of ten keywords, users are asked to select the ones relevant to the article through a pop-up window shown in the Web browser, as depicted in Figure 23, since AKE methods are incapable of providing sufficient accuracy (Papagiannopoulou and Tsoumakas 2020). Users are also allowed to add and select their own keywords through the

user interface.

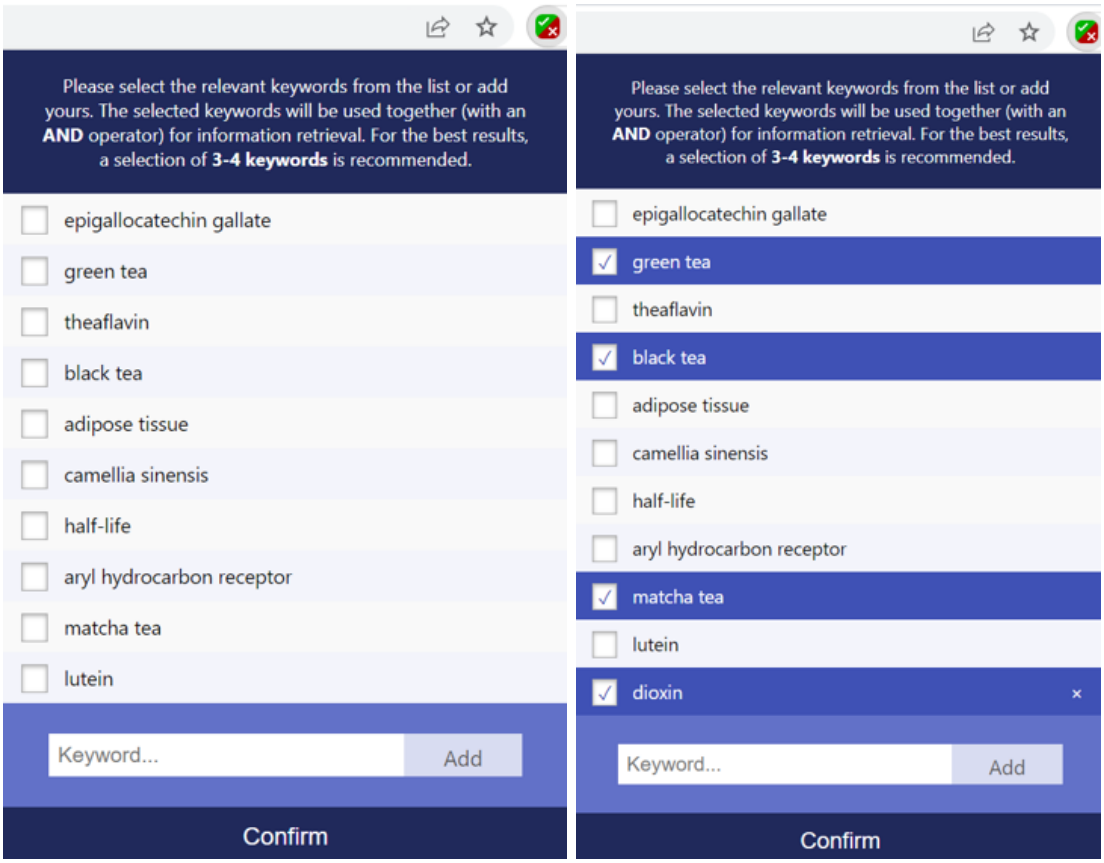


Figure 23: The pop-up window for the keyword extraction and selection step in aedFaCT. The initial list of keywords obtained with AKE is shown on the left while the user's final selection is shown on the right. Note that the last keyword on the right is added by the user.

6.4 Evidence Retrieval

Subsequent to the keyword extraction and selection step, the user's final selection of keywords is used to search for relevant evidence for fact-checking. As mentioned in Section 2.3.2, searching for evidence in aedFaCT includes document retrieval and rationale selection stages. For the former stage, the evidence is searched from two main sources – expert opinions covered in relevant news articles and relevant

scientific publications. Then, the retrieved documents are processed to discover relevant expert opinions and publication abstracts for the rationale selection stage.

6.4.1 Expert Opinion Discovery through News Articles

This stage seeks to investigate the scientific perspectives or comments of domain experts on the identified topic in the news media. The system generates a search query for relevant news items by combining the keywords selected by the user in the previous stage with the *AND* operator. As keywords can also be phrases, each keyword is surrounded by double quotation marks to look for exact matches in the search results. Although a more useful query can be built by combining different logical operators, for the sake of simplicity, the *AND* operator was merely utilised. The searches are carried out using Google's search APIs and take into account the following categories of news sources:

1. *Mainstream News Outlets:* This includes credible news outlets with a high traffic volume and extensive news coverage. A Google site-restricted custom search engine was set up, which allows for the inclusion of ten websites, and ten news outlets with high credibility in English and having no paywall were included. To determine the news outlets with high credibility, the Media Bias/Fact Check (MBFC) credibility ratings[12] was considered because it has been used in various previous research (Krieg et al. 2020; Chen and Freire 2020; Weld, Glenski and Althoff 2021). The left side of Table 21 lists the news outlets that have been included.

2. *Scientific News Outlets:* This category includes credible pro-science news websites that feature scientific perspectives and latest research discoveries. Another Google site-restricted custom search engine with ten selected news websites was set up for this category. With the use of bias, credibility, and

[12] `https://mediabiasfactcheck.com/`

traffic filters, the selection of websites was made based on the MBFC credibility ratings, and websites with a pro-science bias, extensive news coverage, higher traffic, and no paywall were preferred. The right side of Table 21 shows the selection of websites of this sort.

3. *Other Credible News Sources:* Other forms of news sources that may feature expert opinions, in addition to the ones mentioned above, include news issued by institutions and domain-specific news websites (e.g., Medscape, News Medical). To supplement the search results containing the other two categories of news sources, a Google custom search engine with no site restrictions was set up. Because Google Search can return results from non-news websites, the search results were confined with the *NewsArticle*[13] Schema.org type to news articles only. In addition, the Iffy Index of Unreliable Sources[14], which is based on MBFC, was used to filter out untrustworthy news sources from the search results.

Table 21: News outlets covered by the site-restricted custom search engines

Mainstream News Outlets	**Scientific News Outlets**
NPR (`npr.org`)	Science (`science.org`)
NBC News (`nbcnews.com`)	EurekAlert (`eurekalert.org`)
Sky News (`news.sky.com`)	The Scientist (`the-scientist.com`)
ABC News (`abcnews.go.com`)	Science News (`sciencenews.org`)
Euronews (`euronews.com`)	MIT Technology Review (`technologyreview.com`)
Reuters (`reuters.com`)	Popular Science (`popsci.com`)
BBC News (`bbc.com`)	Science Daily (`sciencedaily.com`)
PBS NewsHour (`pbs.com/newshour`)	Science Alert (`sciencealert.com`)
Associated Press (`apnews.com`)	Live Science (`livescience.com`)
CBS News (`cbsnews.com`)	The Conversation (`theconversation.com`)

[13] `https://schema.org/NewsArticle`
[14] `https://iffy.news/index/`

The search results from the three search engines are pooled in the order specified. Although the three search engines can be combined into a single custom search engine, separate site-restricted search engines were preferred for the first two types of news sources based on an observation that site-restricted search engines provide more reliable results, whereas custom search engines configured to search the entire Web are limited to a subset of the Google Web Search corpus[15]. As a result, a standard custom search engine was used as a secondary source to populate the site-restricted search engine results. If there is no search result after the aggregation (especially due to too many keywords selected by the user), the searching process is repeated with the same search query excluding the last keyword. This is because the last keyword is most likely the least useful as it has the lowest AKE score among the selected keywords.

Once the aggregated set of search results is obtained, the system attempts to capture expert opinions from each article, which are mostly in the form of reported speeches, as they contain the most indicative elements of page usefulness for fact-checking (e.g., reported speeches, named entities, and quotes) (Hasanain and Elsayed 2022). It begins by downloading the news article using the Newspaper3k library. The article is then tokenised with two consecutive newline characters to generate paragraphs. Finally, named entities are extracted for each paragraph using the spaCy library's NER feature. Only paragraphs with at least one person name, one academic organisation name (with an indicative word or phrase, such as *university*, *institute*, *academy*, and *research centre*), and a pair of single or double quotation marks (indicating a reported speech) are chosen. As an exception, users are provided with the summary extracted by the Newspaper3k library for the *The Conversation* news articles instead of retrieved expert opinions since they have already been prepared by researchers and academics. The selected paragraphs are incorporated and displayed to users in an individual box that additionally

[15] `https://support.google.com/programmable-search/answer/70392`

includes the source type (icon on the top-left), source name, and publish date, as illustrated in Figure 24. If the displayed expert opinions are insufficient for a decision and further reading is required, users can click on the box to view the complete article. To improve explainability, a green clickable tick linking to the respective MBFC credibility rating webpage is added next to the names of mainstream and science news sources.

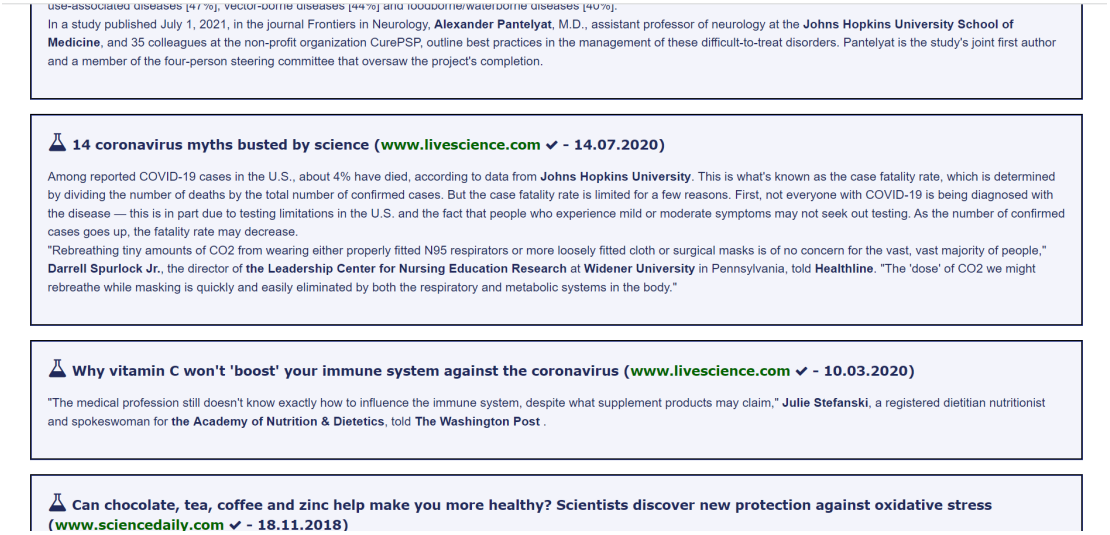


Figure 24: An example output from aedFaCT showing some of the retrieved news articles

6.4.2 Scientific Evidence Retrieval through Research Papers

Being that they are produced by domain experts, scientific publications can also be regarded as a source of expert opinions. As a result, the goal of this stage is to obtain research papers relevant to the topic of the input article.

As aedFaCT aims to include only records with high credibility, this stage requires covering only peer-reviewed publications. In this regard, the Scopus API is used (through the Pybliometrics library (Rose and Kitchin 2019)) to search for relevant peer-reviewed publications. Similar to the previous stage, the searches

are performed by combining the selected keywords with an *AND* operator. Furthermore, each term is surrounded by double quotation marks to allow for the inclusion of loose matches via wildcards and lemmatisation while avoiding phrases to be searched as separate words (Beatty 2022). Although Scopus' search engine also supports exact matching, it was not preferred as it was observed that exact matching tends to return no results at all rather than more relevant results. The obtained search results are presented to the user inside individual boxes including the title, source, publication year, and abstract, in the order of relevance and publication year, as shown in Figure 25.

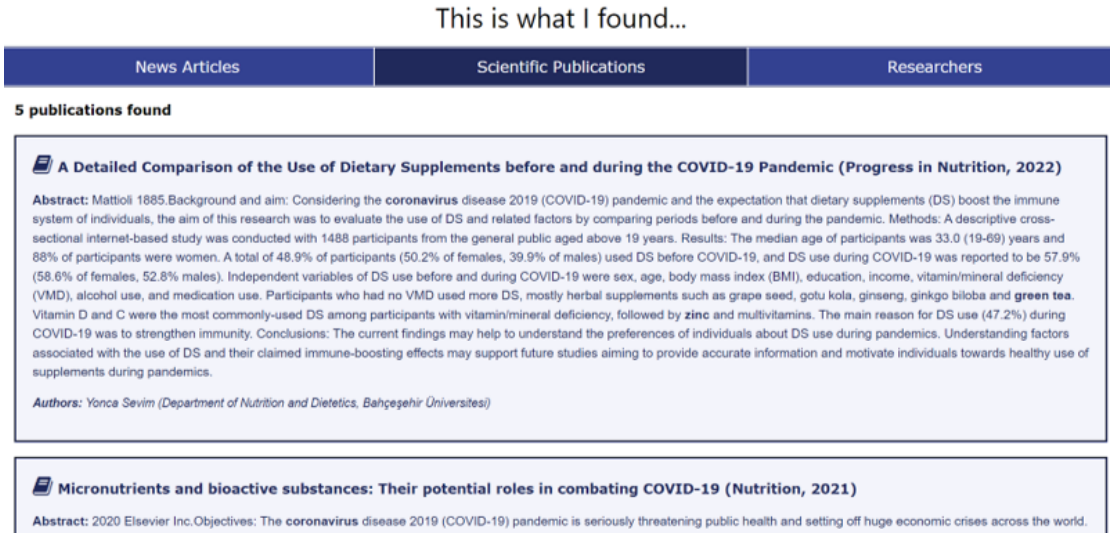


Figure 25: An example output from aedFaCT showing some of the retrieved scientific publications

In addition to the scientific evidence offered by the tool, users, particularly fact-checkers and journalists, may choose to become aware of the experts on the subject in order to follow their research and/or establish contact with them. To do this, aedFaCT profiles the co-authors of the papers collected by retrieving contact information such as profile links from their Scopus and ORCID profiles. The obtained researcher profiles are sorted by their number of publications in the search

results. If this number is equal, they are ranked based on the amount of information in their profile to prioritise more reachable researchers. Figure 26 illustrates an example output from the user interface displaying a list of researchers.

Figure 26: An example output from aedFaCT showing some of the retrieved researcher profiles

6.5 Evaluation of aedFaCT

The evaluation of aedFaCT was performed in two stages. Firstly, a pilot study with three computer science researchers was carried out for the initial assessment of the tool. Then, a larger-scale evaluation study was conducted with 16 participants in the form of a two-part online survey.

Regarding the evaluation metrics, Google's search quality guidelines (Google 2022) were adhered to, which were developed for reviewing Google search engine results using human raters. Despite concerns about the insufficiency of such retrieval effectiveness tests (Lewandowski 2015), similar methodologies are still utilised in the literature (Ciccone and Vickery 2015). The guidelines primarily entail two tasks: determining the extent to which the page fulfils its purpose ("*Page*

Quality") and deciding if search results are helpful ("*Needs Met*"). For the evaluation of aedFaCT, the rating scales of these general search engine evaluation tasks were tailored to focus on *information quality* and *usefulness for fact-checking*, respectively, and the ratings were specialised for the fact-checking process. More precisely, Tables 22 and 23 indicate the rating scales with their descriptions used to evaluate aedFaCT in terms of information quality and usefulness for fact-checking, respectively.

Table 22: Rating scale for information quality

Rating	**Description**
Highest	All the results express high credibility and trustworthiness.
High	Most results express high credibility and trustworthiness although there are a few low-quality results.
Medium	Some of the results are highly credible while there are also low-quality, untrustworthy results. OR, most results are neither from highly credible nor untrustworthy sources.
Low	Many results lack trustworthiness and credibility.
Lowest	Almost all results are untrustworthy, deceptive, harmful to people or society, or have other highly undesirable characteristics.

Table 23: Rating scale for usefulness for fact-checking

Rating	**Description**
Fully Useful	All or almost all users would be immediately and fully satisfied by the results and would not need to view other results to satisfy their need.
Highly Useful	Very helpful for many or most users. Some users may wish to see additional results.
Moderately Useful	Helpful for many users OR very helpful for some users. Some or many users may wish to see additional results.
Slightly Useful	Helpful for fewer users. There is a connection between the query and the result, but not a strong or satisfying connection. Many or most users would wish to see additional results.
Fails to be Useful	Completely fails to meet the needs of the users. All or almost all users would wish to see additional results.

6.5.1 Pilot Study

To verify the validity and functionality of aedFaCT, a preliminary evaluation was carried out as a pilot study with one male and two female computer science researchers. They were provided with 20 health news articles from multiple sources with varying levels of credibility. Since it was beyond the participants' areas of knowledge, the health domain was chosen to provide a better simulation of common readers. The participants were then asked to investigate the veracity of each news article and provide ratings for the displayed output in two rounds: 1) manually by following their daily investigation practices, such as using a Web search engine and/or a research database; 2) by using the proposed tool, aedFaCT.

The participants' responses were collected through a survey prepared on Google Forms. The survey included a rating scale for each processed news article in both rounds, as well as a figure that explained each option on the scale. Furthermore, the survey contained two questions for assessing the perceived success of aedFaCT in terms of which strategy was faster and more beneficial (with the options *manual investigation*, *investigation with aedFaCT*, and *no difference*). Finally, the survey was concluded with an open-ended question asking for feedback about the tool.

Since aedFaCT only considers credible sources, the pilot study aimed to measure only the usefulness of aedFaCT's output for fact-checking, rather than information quality. As a result of the evaluation, the three raters' average rating when manually investigating the provided news articles was 4.35. When they used aedFaCT in their investigations, this average increased to 4.57. Furthermore, with a Fleiss' Kappa of 53.33, the raters were in moderate agreement that aedFaCT offered better or equivalent results to what they were able to acquire with their manual investigations. However, all of the raters agreed that fact-checking with aedFaCT was faster than their own practices. These findings suggest that aedFaCT can assist users do fact-checking faster while maintaining the quality of

retrieved evidence. However, further experimentation with a larger sample is required to assess its performance more properly and obtain statistically significant results.

6.5.2 Evaluation with a Larger Sample

For a more comprehensive evaluation of aedFaCT, a two-part online user study was conducted, aiming to recruit around 20 participants. To be able to obtain statistically significant results and show the effectiveness of aedFaCT for different domains, the initial dataset of 20 health news articles used in the pilot study was expanded to 40 news articles. The newly added articles belong to various domains, including nanotechnology, physics, space, chemistry, and biology, and were randomly selected from the SciNews (Pu et al. 2024) dataset. Since doubling the dataset size would also double the study completion time and to avoid unnecessary time consumption for aedFaCT installation, this time, the evaluation study was conducted as a two-part online survey, focusing on the manual processes in aedFaCT. These two parts are described below:

Part A: This part includes aedFaCT's keyword selection step. The participants were given the news articles with 11 candidate keywords (ten extracted from the articles and one randomly generated non-English word for attention check). Then, they were asked to select up to three candidate keywords that they think are the most useful for fact-checking the claim(s) in the news article. This part also covered some demographic questions, i.e., age, gender, and education level.

Part B: Corresponding to the user's veracity assessment process based on aedFaCT's output, the keywords selected in Part A were used to automatically generate screenshots of aedFaCT's output under different settings. Then, the same participants were asked to rate each given screenshot in terms of

information quality and *usefulness for fact-checking*. The content generated under different settings of aedFaCT and shown to the participants were as follows:

(a) *Summaries of relevant news articles published by credible sources only:* These include the extractive summaries of relevant news articles found instead of expert quotes shown in the original implementation. The articles were summarised using BERT Extractive Summarizer (Miller 2019).

(b) *Expert quotes extracted from relevant news articles published by credible sources only:* These correspond to the original output of aedFaCT shown in the "News Articles" tab on the output page.

(c) *Expert quotes from any relevant news articles regardless of the credibility of the publishing source:* Unlike aedFaCT, these were sourced from any news article returned by the search engine without any exclusion due to source credibility. As mentioned before, Google's custom search engines to search the entire web are limited to a subset of the Google Web Search corpus. Therefore, DuckDuckGo API was utilised via the duckduckgo_search Python library[16] to search for relevant news items.

(d) *Abstracts of relevant peer-reviewed scientific publications:* These correspond to the original output of aedFaCT shown in the "Scientific Publications" tab on the output page.

The user study procedure was approved by the Central Research Ethics Advisory Group at the University of Kent (Application No. CREAG002-09-24). Similar to the survey presented in Chapter 5, the local survey system of iCSS was used to collect the participants' responses, and the participants were recruited anonymously via Prolific. The participants were limited to adults at least 18 years old

[16] `https://github.com/deedy5/duckduckgo_search`

and fluent in English to understand the tasks fully and give useful responses, and no other requirements were applied. After completing Part A with 20 participants, the same participants were invited to attend Part B through the Prolific's user interface. Out of 20 participants of Part A, 16 of them also attended Part B and did not fail the attention checks. They received financial compensation at the rate of £6/hour for their time spent on completing the required tasks. The average completion times of the 16 participants who attended both parts were around 49 minutes for Part A and 1 hour 12 minutes for Part B. As a result, each participant received around £12 for their time. Table 24 shows the demographic information of the 16 participants.

Table 24: The demographics of the 16 user study participants

Characteristics	n	%
Gender		
Male	6	37.5
Female	9	56.3
Prefer Not to Say	1	6.3
Age		
18-29	9	56.3
30-49	5	31.3
50-64	1	6.3
Over 65	1	6.3
Education		
High school or eq.	3	18.8
Bachelor's degree or eq.	10	62.5
Master's degree or eq.	1	6.3
Doctoral degree or eq.	2	12.5

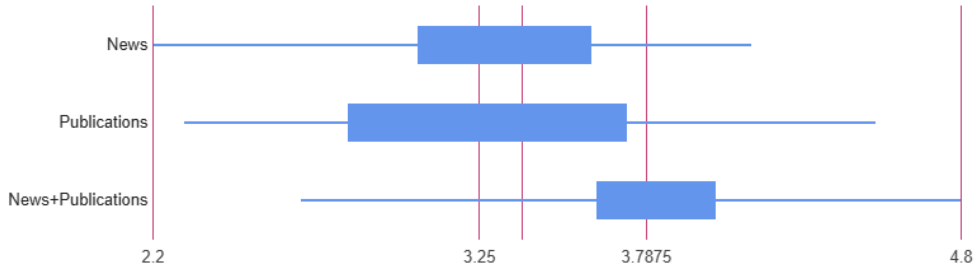


Figure 27: The box plot showing the distribution of the mean ratings when news articles, scientific publications, or both were considered in aedFaCT. The red lines passing through the middle of the boxes indicate the median values.

For the analysis of the responses, IBM SPSS Statistics (version 29) was leveraged to run descriptive statistics, as well as paired samples t-tests to compare the ratings for different settings (a)-(d). The results obtained with statistical analysis are summarised below.

aedFaCT's Effectiveness

The effectiveness of aedFaCT was measured using the usefulness for fact-checking ratings. On average, displaying expert quotes from news articles ($M = 3.32, SD = 0.482$) and abstracts from scientific publications ($M = 3.32, SD = 0.639$) were considered moderately useful. Additionally, assuming that the usefulness of either content type can satisfy the user to reach a verdict, the ratings given for (b) and (d) were unified by considering the highest one as the rating of the entire aedFaCT output. In this case, the mean value for the unified rating appeared as $M = 3.82, SD = 0.517$, higher than the mean ratings for both content types, as shown in Figure 27. This shows that leveraging news articles and scientific publications together in evidence retrieval increased the effectiveness of aedFaCT towards the level of *highly useful*, according to the scale in Table 23.

Expert Quotes vs Summaries

The comparison of the ratings for (a) and (b) showed no statistically significant difference in either information quality ($t(15) = -0.72, p = .242$) or usefulness for fact-checking ($t(15) = -0.37, p = .359$) between displaying the retrieved evidence as expert quotes and summaries in the output. This implies that displaying expert quotes might be an alternative way of showing the output of the evidence retrieval stage, and this approach does not necessarily degrade the usefulness and/or quality of the information presented by aedFaCT.

Only Credible Sources vs Any Sources

When it comes to the trustworthiness of aedFaCT, the information quality ratings given for (b) and (d) were considered for the evaluation. The mean value of the participants' ratings for news articles was $M = 3.58, SD = 0.348$ and for peer-reviewed scientific publications was $M = 3.69, SD = 0.629$. Based on the information quality ratings given for (b) and (c), limiting the news sources to credible ones ($M = 3.58, SD = 0.35$) led to significantly higher ratings, compared to the regular search engine results ($M = 3.15, SD = 0.42$), $t(15) = 7.63, p < .001$. This indicates that aedFaCT can offer more trustworthiness in fact-checking investigations than regular search engine-based manual investigations.

News Articles vs Scientific Publications

The ratings for the two types of evidence supported by aedFaCT – news articles (b) and publication abstracts (d) – did not significantly differ, in terms of either information quality ($t(15) = -0.75, p = .233$) or usefulness for fact-checking ($t(15) = 0.05, p < .519$).

6.6 Summary

Based on the derived conceptual understanding and supported by the empirical evidence discussed in previous chapters, this chapter introduced aedFaCT, a semi-automated fact-checking tool focusing on the evidence retrieval stage of a typical fact-checking process. It accelerates manual fact-checking practices using common searching tools, including Web search engines and a research database. More precisely, it gathers scientific evidence through expert comments from multiple credible sources, such as news articles from mainstream and scientific news outlets, and peer-reviewed research papers. From the perspective of the EER model presented in Chapter 4, aedFaCT's workflow for fact-checking a news article on a topic T can be modelled as shown in Figure 28.

The design of aedFaCT is supported by the survey findings in Chapter 5. Firstly, a widely used search engine (i.e., Google Web Search) is integrated to search for relevant news articles, which appeared as the most common information source for searching. Secondly, science communication is considered by displaying expert comments retrieved from news articles, leveraging the science communication expertise developed in news media. Moreover, the difference between trust in experts and the use of research databases for information searching justifies the need for aedFaCT's expert-oriented design. Finally, aedFaCT joins the forces of human expertise and automation. While human expertise is utilised during the keyword selection and evidence-based veracity assessment steps, the extraction of candidate keywords and retrieving relevant evidence are performed by automation. Additionally, links to the source credibility ratings for each news item are provided in aedFaCT's user interface for better explainability. Apart from these, the finding on utilising audiovisual communication motivates further enhancements on the current aedFaCT implementation to support audiovisual claims and evidence in the future.

The presented tool was evaluated through a two-part online user study with 16

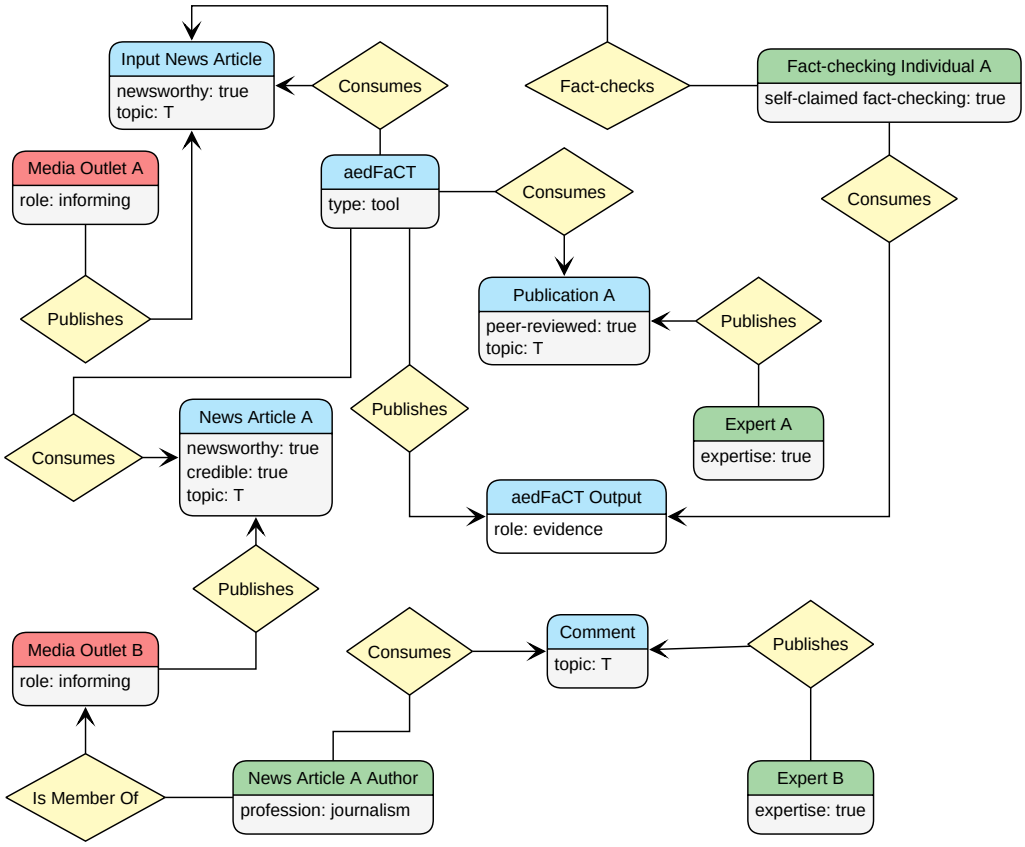


Figure 28: The workflow of aedFaCT for fact-checking a news article on a topic T, modelled with the EER model introduced in Chapter 4.

participants, following a pilot study with three computer science researchers. The user study findings shed light on how an effective fact-checking tool with enhanced trustworthiness can be implemented, addressing RQ4 mentioned in Section 1.2. More precisely, aedFaCT appeared as a useful tool for retrieving evidence from news articles and peer-reviewed publications, increasing its effectiveness. Furthermore, its credibility-focused approach was considered more trustworthy by the participants, compared to regular searching practices. Finally, aedFaCT introduced an alternative way of presenting evidence for fact-checking tools, utilising expert comments.

Despite its several strengths mentioned, the existing implementation of aedFaCT has some limitations. First, it is dependent on the amount of request quotas set by external APIs, i.e., Google and Scopus. While Scopus Search APIs[17] have a weekly quota of 5,000 to 20,000 queries, depending on the utilised API service, Google Custom Search APIs[18] allow 10,000 requests per day. This poses a significant challenge to deploy aedFaCT freely for a wider community. The tool's considerably slow keyword extraction performance is another drawback. Since existing AKE methods already suffer from poor accuracy (Papagiannopoulou and Tsoumakas 2020), accuracy was prioritised over speed for aedFaCT. Therefore, SIFRank+ was chosen as it appeared as the most accurate AKE algorithm despite the existence of various lightweight AKE methods. Additionally, the *AND* operator was used to simply combine the selected keywords. This results in fewer records from the searches, particularly if the user selected an excessive number of keywords. Better search results thus need a smarter strategy that makes use of a mix of logical operators (e.g., utilising the *OR* operator for similar keywords). Finally, aedFaCT has been evaluated with a limited number of participants with similar backgrounds. Therefore, larger-scale experiments including a more representative and diverse set of participants – including professionals such as journalists and fact-checkers, as well as common readers – are needed.

In addition to resolving the limitations, aedFaCT's capabilities are aimed to be enhanced. The current version overlooks evidence retrieval from official sources, e.g., governmental bodies, NGOs, and academic institutions. These can be easily identified from the general Web search results by seeking their corresponding URL extensions. Furthermore, aedFaCT can be improved by incorporating research on claim detection, which would allow extracted keywords to be more precisely targeted at certain claims in the input text. Lastly, as mentioned previously, the current implementation can be enhanced to support audiovisual claims and

[17] `https://dev.elsevier.com/api_key_settings.html`
[18] `https://developers.google.com/custom-search/v1/overview#pricing`

evidence to reach wider audiences.

Chapter 7

Conclusion

7.1 Summary of Contributions

As a problem as old as human history, false information has been harmful to society for centuries. With the proliferation of digital technologies such as the Internet and social media, however, the issue of false information has become much more complicated and challenging, which made it the focus of many research disciplines. This complex nature of the problem revealed the need for developing more comprehensive approaches to fight against it, considering the different aspects of the problem and the findings obtained from different disciplines. With this respect, this thesis aims to contribute to the development of a common understanding of the false information and fact-checking ecosystem, as well as to propose a working prototype based on the developed understanding. While the contributions of each chapter from Chapter 3 to Chapter 6 are explained in the summary sections at the end of each chapter, this section is intended to provide a general summary of the contributions.

Due to the diversity of disciplines researching false information and the lack of unified conceptual frameworks and models covering false information, Chapters 3

and 4 attempted to provide a better understanding of the phenomenon with different conceptual models based on the literature. Initially, Chapter 3 introduced a data-driven topical analysis of the relevant literature through topic modelling, covering publications from a considerably long period, to shed light on the concepts studied by the research community. In consequence, the performed topical analysis revealed several topics that belong to the ecosystem, corresponding to different types, applications, and methods regarding false information. This pointed out that the presented topic models can be leveraged to construct an ontology of false information, which could be a milestone in developing a common understanding of the phenomenon across different disciplines. Moreover, the identified topics by the topic models can guide researchers who study false information and relevant concepts in determining potential research gaps.

Inspired by the close relationship between different topics identified, then, Chapter 4 presented an enhanced entity-relationship (EER) model of the false information and fact-checking ecosystem to explore the stakeholders involved in the ecosystem and the relationships between them. As the first of its kind, the constructed EER model can be used by researchers and practitioners in many different ways. Out of a wide range of its potential applications, two example applications of the proposed model, i.e., modelling real-world scenarios and conducting literature reviews, were provided to demonstrate its usefulness. Other potential applications of the model involve designing empirical studies for a more representative recruitment of participants, modelling and simulating the ecosystem with agent-based models based on the EER model, and studying the dissemination of false information and the overall evolution of the ecosystem. In addition, the EER model can be utilised to construct an ontology useful for different purposes, such as conducting large-scale Internet measurement studies to explore the false information propagation and the roles of different actors, as well as identifying new relevant concepts via automated reasoning.

When it comes to combatting false information, Chapters 5 and 6 focused on identifying needs for better fact-checking solutions and offering a working prototype as an example implementation. Despite highly technical research involved in fact-checking, considering human aspects is crucial for the practical adoption and trust development of proposed solutions. This refers to taking a closer look at the proposed EER model to explore user perspectives on fact-checking. Therefore, Chapter 5 aimed to closely examine these relationships through an online user survey with 302 participants regarding human perceptions and attitudes towards fact-checking tools. To ensure that the survey covered the current landscape comprehensively, it was prepared based on a taxonomy of fact-checking tools constructed based on available tools and proposed approaches. Unlike existing taxonomies mostly focusing on the stages of fact-checking where tools are used, for categorisation, the proposed taxonomy covers different characteristics of tools, such as their level of automation, their contribution to fact-checking, and the platforms they support. As a result, the conducted survey revealed some design cues that can guide researchers in developing more trustworthy and usable fact-checking solutions, which include integrating common searching platforms, using audiovisual communication strategies, providing enhanced science communication, and combining the forces of human expertise and automation speed.

Finally, Chapter 6 combined the conceptual and empirical understanding obtained in the previous chapters to develop a more trustworthy fact-checking solution. As the first of its kind, the presented solution, *aedFaCT* (i.e., **a**utomatic **e**xpert **d**iscovery-based **Fa**ct-**C**hecking **T**ool), is a Web browser extension for semi-automated fact-checking entirely based on expert opinion discovery. In line with the findings presented in Chapter 5, aedFaCT targets better science communication by gathering scientific evidence from multiple sources, i.e., news articles and scientific publications, that might appeal to users with different educational

backgrounds. Furthermore, its semi-automated architecture leaving veracity assessment to users can build trust among users more conveniently, compared to fully automated fact-checking systems. Ultimately, aedFaCT was designed in accordance with the daily fact-checking practices of professional fact-checkers and common readers, which can facilitate its practical adoption.

The findings obtained throughout this thesis address the overarching research question, RQ. More specifically, each research question presented in Section 1.2 is answered below based on the findings:

RQ1: What are the common concepts that have been studied by the research community in the false information literature?

In Chapter 3, a variety of concepts related to false information arose from the application of topic modelling to the relevant literature. These include different types of false information, methods to analyse or detect false information, contexts where it can exist, and different applications of false information.

RQ2: What stakeholders exist in the false information and fact-checking ecosystem, and how are they related?

The presented EER model in Chapter 4 revealed various stakeholders, including different individual and organisational actors, as well as their roles in the ecosystem. Furthermore, the relationships between these actors were explained for the main processes in the ecosystem, i.e., information generation, fact-checking, and lawmaking and regulation.

RQ3: What are the user perspectives on fact-checking, the stakeholders involved, and the fact-checking tools?

The survey results presented in Chapter 5 indicated that search engines, video streaming platforms, mainstream news outlets, and online encyclopedias were the most common sources to search for specific information.

When it comes to the stakeholders, experts and academic institutions were the most trusted actors. Regarding fact-checking tools, the results revealed the superiority of semi-automated approaches, especially those leveraging the human component in veracity assessment (at least for correcting the verdict), over manual and fully automated solutions in terms of trustworthiness. Moreover, explainability appeared as an important factor for the trustworthiness of fact-checking tools.

RQ4: How the findings obtained from the answers to the previous research questions can be used for the development of a new more trustworthy and effective fact-checking tool?

The conceptual understanding obtained from the answers to RQ1 and RQ2 shed light on the scope of existing research on the ecosystem and the workflow of typical fact-checking processes. This supported the development of a new fact-checking tool to address RQ4, aedFaCT, accelerating and facilitating typical fact-checking processes. The design of aedFaCT was further supported by the findings addressing RQ3 for developing an effective tool with enhanced trustworthiness.

7.2 Limitations and Future Work

The research presented in this thesis has some limitations that can be focused on in future to increase its impact. To begin with, the topic models mentioned in Chapter 3 were based on research published before 2020, excluding the COVID-19 pandemic. Thus, reconstructing the models with more recent research, especially including post-COVID publications, could reveal new concepts that emerged during the COVID-19 pandemic and shed light on the implications of the pandemic to false information research. Besides, during the generation of topic models, many publications have been overlooked due to the lack of institutional access to some

publishers and research databases, such as APA PsychInfo. One way of overcoming this limitation could be by augmenting the collected dataset of research items with other accessible publications sharing the same research area as those unable to be reached.

An obvious next step for the conceptual models presented in Chapters 3 and 4 might be developing an ontology, representing the common knowledge regarding the false information and fact-checking ecosystem. Nevertheless, this task is quite challenging due to the diversity of terminologies across disciplines studying false information. Therefore, (semi-)automated data-driven approaches (Al-Aswadi, Chan and Gan 2020; Zhong et al. 2023) can be leveraged to reduce the laborious work during ontology construction which can be overwhelming because of the need to include several disciplines with different points of view in the construction process.

When it comes to the user survey exploring human attitudes and behaviours towards fact-checking tools, the lack of older participants over 65 and the overrepresentation of those younger than 30 is a limitation of the study. As discussed in Section 5.1, recent research has indicated scientific evidence regarding generational differences in fact-checking practices and using fact-checking tools. Therefore, a more balanced population of participants can reveal specific user needs for different age groups, especially those underrepresented in the survey. This can also be useful in enhancing the capabilities of aedFaCT by considering the needs of older people.

Apart from that, an important limitation of aedFaCT, complicating its publicly available deployment, is its dependence on the request quotas set by the external APIs used, i.e., Google and Scopus. Although this can be resolved by requiring users to have their own API keys, the low usability of API keys remains a challenge (Myers and Stylos 2016). In addition, the automatic keyword extraction (AKE) process is the bottleneck of the tool's performance since accuracy

was preferred over speed when choosing the AKE algorithm to be employed. To accelerate aedFaCT, the AKE process can be improved by incorporating state-of-the-art claim detection research and identifying keywords with a less accurate but faster AKE method from a set of claims retrieved from the input text instead of the entire text. Finally, aedFaCT needs more experiments for the evaluation of its accuracy and usability, which should be conducted with an adequate number of both professionals and common readers.

Other than the limitations to be resolved in future work, the capabilities of aedFaCT can be improved in many ways. For example, official sources, such as academic institutions and governmental bodies, can be utilised to retrieve trustworthy evidence as the websites of such sources can be distinguished via their specific URL extensions. In addition, aedFaCT can be upgraded to support non-English languages, especially high-resource languages, as well provided that a multilingual AKE algorithm is employed. In that case, the language of the input text can be detected automatically with a language identifier, which can achieve promising outcomes (Jauhiainen et al. 2019). Last but not least, aedFaCT could benefit from semantic similarity measures to improve its rationale selection during evidence retrieval.

From a higher-level perspective, the advancements in two research fields might be game-changing for the development of fact-checking tools with enhanced accuracy, trustworthiness, and usability, which are discussed in more detail in the following subsections.

7.2.1 Large Language Models (LLMs)

With the advancements in deep learning and NLP, pre-trained language models with a high capability of understanding and generating natural language have emerged (Jones et al. 2022). Recently, large pre-trained language models (LLMs) with massive parameter sizes have been introduced, which enables more natural

conversations with users through their user interfaces. Examples include OpenAI's ChatGPT[1], Google's Bard[2], and many others. From the perspective of this thesis, this can be expected to have two major effects on false information and fact-checking research in future:

LLM-generated false information and its detection. Due to their superior performance in many tasks, LLMs are increasingly used as an information source like a search engine. However, LLMs can unintentionally generate false or misleading content too, known as *hallucinations* (Augenstein et al. 2023). Combined with the use of eloquent language and a confident tone, hallucinations and hallucinatory explanations generated to justify them can be quite persuasive and mislead people. Recent research has indicated that detecting LLM-generated false information is harder for both humans and detectors compared to those generated by humans, meaning that it can pose a serious threat to society (Chen and Shu 2023). Compared to false information provided by typical search engines, preventing users from those generated by LLMs can be more challenging as LLMs lack clear indications of information sources. Despite efforts to address LLM hallucinations during pretraining or with post-processing approaches, more research is needed for hallucination detection and correcting LLM outputs (Wang et al. 2023). Other than misinformation, LLMs can also be instrumentalised to generate disinformation at scale to perform malicious activities, such as personalised attacks, generating infinite variations of the same fake content to bypass monitoring, impersonating the writing style of individuals, and generating convincing fake profiles (Augenstein et al. 2023).

Using LLMs in fact-checking. In addition to generating false information, LLMs can also be leveraged to combat false information. Early findings on the use of existing LLMs for fact-checking indicated that they can serve as a moderately successful (with an accuracy between 60-70%) false information classifier (Hoes,

[1] `https://chat.openai.com/`
[2] `https://bard.google.com/`

Altay and Bermeo 2023; Caramancion 2023). Besides, LLMs can be fine-tuned to support different tasks in fact-checking, including claim matching (Choi and Ferrara 2023), evidence retrieval (Xin, Bowei and Ai Ti 2023), identifying credibility signals (Leite et al. 2023), and veracity assessment (Zhang and Gao 2023; Kareem and Abbas 2023). Such developed models can also be combined in a modular end-to-end solution, each module corresponding to different stages of fact-checking (Li et al. 2024). Nonetheless, the accuracy of existing LLMs in fact-checking implies that they cannot replace human fact-checkers yet, emphasising the need for human-LLM-teamed solutions (Quelle and Bovet 2023).

7.2.2 Explainable AI (xAI)

As briefly discussed in Section 2.6.3, explainability needs to be considered in fact-checking systems with a level of automation to build trust among their users. Furthermore, regulations on digital services, such as GDPR and the Digital Services Act, require digital service providers to provide transparent reporting and explanations (Augenstein 2021). In this manner, research in explainable AI (xAI) can be leveraged in (semi-)automated fact-checking solutions for enhanced transparency and trustworthiness. In addition, explanations offered by such systems can prevent users from overtrusting the system, considering that AI can never guarantee perfect accuracy (Mohseni et al. 2021).

Existing literature on explainable fact-checking targets experts for the provided explanations, rather than end users (A.B., Kumar and Chacko 2023). Hence, in the context of fact-checking, xAI-based solutions need further improvements to become more human-friendly for their public adoption. Initially, more research is needed to develop usable user interfaces to present concise explanations tailored according to user needs (Lim and Perrault 2023). Although longer explanations can improve the overall performance of users, there may be a trade-off in terms

of the time and attention required to understand them (Linder et al. 2021). Regarding the type of explanations provided, example-based or natural language explanations might be a better choice than feature-based explanations revealing the inner workings of the AI models. The reason is that malicious actors may attempt to benefit from feature-based explanations to manipulate the system with adversarial inputs, which could result in questioning the AI model's competence and distrust of the output generated by the system. Finally, in some cases, there may not exist adequate evidence to reach a verdict regarding the veracity of a claim. When this happens, an xAI-based fact-checking system should also be able to explain why the collected evidence is insufficient (Kotonya and Toni 2020). Bearing current research gaps regarding xAI-based fact-checking in mind, the xAI component of fact-checking systems can be designed to work with a human component in collaboration, e.g., human correction of the xAI outputs (Zhang, Rudra and Anand 2021), especially in the short-term.

Bibliography

A.B., A., Kumar, S. M. and Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122, pp. 106087:1–106087:13, https://doi.org/10.1016/j.engappai.2023.106087.

Abdul Wahab, A. W. et al. (2014). Passive video forgery detection techniques: A survey. In *Proceedings of the 10th International Conference on Information Assurance and Security*, IEEE, pp. 29–34, https://doi.org/10.1109/ISIAS.2014.7064616.

Ada Lovelace Institute and The Alan Turing Institute (2023). How do people feel about AI? Tech. rep., https://www.adalovelaceinstitute.org/wp-content/uploads/2023/06/How-do-people-feel-about-AI-Ada-Turing.pdf.

Adams, S. (2010). Revisiting the online health information reliability debate in the wake of "Web 2.0": An inter-disciplinary literature and website review. *International Journal of Medical Informatics*, 79(6), pp. 391–400, https://doi.org/10.1016/j.ijmedinf.2010.01.006.

Ahsan, M., Kumari, M. and Sharma, T. P. (2019). Rumors detection, verification and controlling mechanisms in online social networks: A survey. *Online Social Networks and Media*, 14, pp. 100050:1–100050:12, https://doi.org/10.1016/j.osnem.2019.100050.

Al-Aswadi, F. N., Chan, H. Y. and Gan, K. H. (2020). Automatic ontology construction from text: A review from shallow to deep learning trend. *Artificial Intelligence Review*, 53, pp. 3901–3928, https://doi.org/10.1007/s10462-019-09782-9.

Al-Qershi, O. M. and Khoo, B. E. (2013). Passive detection of copy-move forgery in digital images: State-of-the-art. *Forensic Science International*, 231(1), pp. 284–295, https://doi.org/10.1016/j.forsciint.2013.05.027.

Alemanno, A. (2018). How to counter fake news? A taxonomy of anti-fake news approaches. *European Journal of Risk Regulation*, 9(1), pp. 1–5, https://doi.org/10.1017/err.2018.12.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), pp. 211–236, https://doi.org/10.1257/jep.31.2.211.

Allen, J. et al. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36), pp. eabf4393:1–eabf4393:10, https://doi.org/10.1126/sciadv.abf4393.

Alothali, E. et al. (2018). Detecting social bots on Twitter: A literature review. In *Proceedings of the 2018 International Conference on Innovations in Information Technology*, IEEE, pp. 175–180, https://doi.org/10.1109/INNOVATIONS.2018.8605995.

Altay, S., Berriche, M. and Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, 9(1), https://doi.org/10.1177/20563051221150412.

Altay, S. et al. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4), pp. 1–34, https://misinforeview.hks.harvard.

edu/article/a-survey-of-expert-views-on-misinformation-definitions-determinants-solutions-and-future-of-the-field/.

Altuncu, E. et al. (2022). Improving performance of automatic keyword extraction (AKE) methods using pos-tagging and enhanced semantic-awareness. https://doi.org/10.48550/arXiv.2211.05031.

Altuncu, E. et al. (2023). aedFaCT: Scientific fact-checking made easier via semi-automatic discovery of relevant expert opinions. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, AAAI, pp. 27:1–27:10, https://doi.org/10.36190/2023.27.

Amiram, D. et al. (2018). Financial reporting fraud and other forms of misconduct: A multidisciplinary review of the literature. *Review of Accounting Studies*, 23(2), pp. 732–783, https://doi.org/10.1007/s11142-017-9435-x.

Arora, S., Liang, Y. and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the 2017 International Conference on Learning Representations*, pp. 1–16, https://openreview.net/forum?id=SyK00v5xx.

Arroyo, D. et al. (2023). On the design of a misinformation widget (MsW) against cloaked science. In *Network and System Security*, Springer, pp. 385–396, https://doi.org/10.1007/978-3-031-39828-5_21.

Atanasova, P. (2024). *Accountable and Explainable Methods for Complex Reasoning over Text*. Springer, https://doi.org/10.1007/978-3-031-51518-7.

Atanasova, P. et al. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, 11(3), https://doi.org/10.1145/3297722.

Augenstein, I. (2021). *Towards Explainable Fact Checking*. Ph.D. thesis, Department of Computer Science, Faculty of Science, University of Copenhagen, https://static-curis.ku.dk/portal/files/287065737/Isabelle_Doktordisputats.pdf.

Augenstein, I. et al. (2017). SemEval 2017 Task 10: ScienceIE-extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, ACL, pp. 546–555, https://doi.org/10.18653/v1/S17-2091.

Augenstein, I. et al. (2023). Factuality challenges in the era of large language models. https://doi.org/10.48550/arXiv.2310.05189.

Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76, https://doi.org/10.1016/j.ejpoleco.2022.102256.

Bandhakavi, A., Hoffmann, H. and Lear, P. (2022). Tackling misinformation with HAMLET (human and machine in the loop evaluation and training). Tech. rep., Logically, https://www.logically.ai/download-tackling-misinformation-with-hamlet-an-expert-in-the-loop-ai-framework.

Barni, M., Stamm, M. C. and Tondi, B. (2018). Adversarial multimedia forensics: Overview and challenges ahead. In *Proceedings of the 26th European Signal Processing Conference*, IEEE, pp. 962–966, http://doi.org/10.23919/EUSIPCO.2018.8553305.

Beatty, S. (2022). 6 simple search tips: Lessons learned from the Scopus webinar. https://blog.scopus.com/posts/6-simple-search-tips-lessons-learned-from-the-scopus-webinar.

Bennani-Smires, K. et al. (2018a). Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational*

Natural Language Learning, pp. 221–229, https://doi.org/10.18653/v1/K18-1022.

Bennani-Smires, K. et al. (2018b). Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 221–229, https://doi.org/10.18653/v1/K18-1022.

Bhuiyan, M. M. et al. (2021). Designing transparency cues in online news platforms to promote trust: Journalists' & consumers' perspectives. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), https://doi.org/10.1145/3479539.

Birajdar, G. K. and Mankar, V. H. (2013). Digital image forgery detection using passive techniques: A survey. *Digital Investigation*, 10(3), pp. 226–245, https://doi.org/10.1016/j.diin.2013.04.007.

Blaber, Z. N. et al. (2019). Supply and demand of fake news: Review and implications for business research. *Journal of Applied Business and Economics*, 21(4), pp. 11–28, https://doi.org/10.33423/jabe.v21i4.2127.

Botambu Collins, N. T. N., Dinh Tuyen Hoang and Hwang, D. (2021). Trends in combating fake news on social media – A survey. *Journal of Information and Telecommunication*, 5(2), pp. 247–266, https://doi.org/10.1080/24751839.2020.1847379.

Botnevik, B., Sakariassen, E. and Setty, V. (2020). BRENDA: Browser extension for fake news detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 2117–2120, https://doi.org/10.1145/3397271.3401396.

Boudin, F. (2016). PKE: An open source Python-based keyphrase extraction toolkit. In *Proceedings of the 26th International Conference on Computational*

Linguistics: System Demonstrations, pp. 69–73, http://aclweb.org/anthology/C16-2015.

Bouleimen, A. et al. (2024). Online search is more likely to lead students to validate true news than to refute false ones. In *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media*, AAAI, pp. 24:1–24:12, https://doi.org/10.36190/2024.24.

Bowman, S. and Willis, C. (2003). We Media: How audiences are shaping the future of news and information. Tech. rep., The Media Center at the American Press Institute, http://www.flickertracks.com/blog/images/we_media.pdf.

Bucchi, M. (2017). Credibility, expertise and the challenges of science communication 2.0. *Public Understanding of Science*, 26(8), pp. 890–893, https://doi.org/10.1177/0963662517733368.

Burel, G. and Alani, H. (2023). The fact-checking observatory: Reporting the co-spread of misinformation and fact-checks on social media. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, ACM, https://doi.org/10.1145/3603163.3609042.

Campos, R. et al. (2020). YAKE! keyword extraction from single documents using multiple local features. *Information Sciences*, 509, pp. 257–289, https://doi.org/10.1016/j.ins.2019.09.013.

Cao, J. et al. (2018). Automatic rumor detection on microblogs: A survey. https://doi.org/10.48550/arXiv.1807.03505.

Caracciolo, C. et al. (2013). The AGROVOC linked dataset. *Semantic Web*, 4(3), pp. 341–348, https://doi.org/10.3233/SW-130106.

Caramancion, K. M. (2023). News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking. https://doi.org/10.48550/arXiv.2306.17176.

Casillo, M. et al. (2021). Fake news detection using LDA topic modelling and k-nearest neighbor classifier. In *Computational Data and Social Networks: Proceedings of the 2021 International Conference on Computational Data and Social Networks*, Springer, pp. 330–339, https://doi.org/10.1007/978-3-030-91434-9_29.

Cazalens, S. et al. (2018). A content management perspective on fact-checking. In *Companion Proceedings of the The Web Conference 2018*, ACM, pp. 565–574, https://doi.org/10.1145/3184558.3188727.

Ceron, W., de Lima-Santos, M.-F. and Quiles, M. G. (2021). Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content. *Online Social Networks and Media*, 21, pp. 100116:1–100116:14, https://doi.org/10.1016/j.osnem.2020.100116.

Chen, C. and Shu, K. (2023). Can LLM-generated misinformation be detected? In *Proceedings of the NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 16, pp. 16:1–16:29, https://openreview.net/forum?id=yi8KGilFFk.

Chen, Y., Conroy, N. J. and Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, ACM, pp. 15–19, https://doi.org/10.1145/2823465.2823467.

Chen, Z. and Freire, J. (2020). Proactive discovery of fake news domains from real-time social media feeds. In *Companion Proceedings of the Web Conference 2020*, ACM, pp. 584–592, https://doi.org/10.1145/3366424.3385772.

Choi, E. C. and Ferrara, E. (2023). Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. https://doi.org/10.48550/arXiv.2310.09223.

Chong, M. and Choy, M. (2020). An empirically supported taxonomy of misinformation. In *Navigating Fake News, Alternative Facts, and Misinformation in a Post-Truth World*, IGI Global, pp. 117–138, https://doi.org/10.4018/978-1-7998-2543-2.ch005.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, Le Centre De Hautes Etudes Internationales D'Informatique Documentaire, pp. 609–623, https://doi.org/10.5555/3374430.3374485.

Chuang, J., Manning, C. D. and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the 2012 International Working Conference on Advanced Visual Interfaces*, ACM, pp. 74–77, https://doi.org/10.1145/2254556.2254572.

Ciccone, K. and Vickery, J. (2015). Summon, EBSCO Discovery Service, and Google Scholar: A comparison of search performance using user queries. *Evidence Based Library and Information Practice*, 10(1), pp. 34–49, https://doi.org/10.18438/B86G6Q.

Clark, B. (2023). Proactive institutional repository collection development techniques: Archiving gold open access articles and metadata retrieved with web scraping. *Journal of Library Administration*, 63(6), pp. 743–765, https://doi.org/10.1080/01930826.2023.2240190.

Colepicolo, E. (2015). Information reliability for academic research: Review and

recommendations. *New Library World*, 116(11/12), pp. 646–660, https://doi.org/10.1108/NLW-05-2015-0040.

Conroy, N. J., Rubin, V. L. and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1), pp. 1–4, http://doi.org/10.1002/pra2.2015.145052010082.

Cooke, N. A. (2017). Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The Library Quarterly*, 87(3), pp. 211–221, https://doi.org/10.1086/692298.

Das, A. et al. (2023). The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2), https://doi.org/10.1016/j.ipm.2022.103219.

Das Gollapalli, S. and Caragea, C. (2014). Extracting keyphrases from research papers using citation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), https://doi.org/10.1609/aaai.v28i1.8946.

Deka, P., Jurek-Loughrey, A. and Deepak (2022). Unsupervised keyword combination query generation from online health related content for evidence-based fact checking. In *The 23rd International Conference on Information Integration and Web Intelligence*, ACM, pp. 267–277, https://doi.org/10.1145/3487664.3487701.

Dholakia, N., Ozgun, A. and Atik, D. (2023). The miasma of misinformation: A social analysis of media, markets, and manipulation. *Consumption Markets & Culture*, pp. 1–16, https://doi.org/10.1080/10253866.2022.2149508.

Dorminey, J. et al. (2012). The evolution of fraud theory. *Issues in Accounting Education*, 27(2), pp. 555–579, https://doi.org/10.2308/iace-50131.

Douglas, K., Sutton, R. M. and Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science*, 26, pp. 538–542, http://doi.org/10.1177/0963721417718261.

Douglas, K. M. et al. (2019). Understanding conspiracy theories. *Political Psychology*, 40(S1), pp. 3–35, https://doi.org/10.1111/pops.12568.

Dwivedi, R. et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), pp. 194:1–194:33, https://doi.org/10.1145/3561048.

Ecker, U. K. H. et al. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), pp. 13–29, https://doi.org/10.1038/s44159-021-00006-y.

Edelman (2023). Edelman Trust Barometer 2023: Global Report. Tech. rep., https://www.edelman.com/sites/g/files/aatuss191/files/2023-03/2023%20Edelman%20Trust%20Barometer%20Global%20Report%20FINAL.pdf.

El-Beltagy, S. R. and Rafea, A. (2009). KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information Systems*, 34(1), pp. 132–144, https://doi.org/10.1016/j.is.2008.05.002.

Elizabeth, J. (2014). Who are you calling a fact checker? *Civic Discourse & Democracy, American Press Institute*, https://americanpressinstitute.org/fact-checking-project/fact-checker-definition/.

Erlich, A. and Garner, C. (2023). Is Pro-Kremlin disinformation effective? Evidence from Ukraine. *The International Journal of Press/Politics*, 28(1), pp. 5–28, https://doi.org/10.1177/19401612211045221.

Ermakova, L. et al. (2023). Overview of the CLEF 2023 SimpleText task 3: Simplification of scientific texts. In *Working Notes of the Conference and Labs of*

the Evaluation Forum, vol. 3497, CEUR Workshop Proceedings, pp. 2855–2875, https://ceur-ws.org/Vol-3497/paper-240.pdf.

Fallis, D. (2015). What is disinformation? *Library Trends*, 63(3), pp. 401–426, https://doi.org/10.1353/lib.2015.0014.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, 4(5), pp. e5738:1–e5738:11, https://doi.org/10.1371/journal.pone.0005738.

Fard, A. E. and Cunningham, S. (2019). Assessing the readiness of academia in the topic of false and unverified information. *Journal of Data and Information Quality*, 11(4), https://doi.org/10.1145/3313788.

Farid, H. (2009). Image forgery detection. *IEEE Signal Processing Magazine*, 26(2), pp. 16–25, https://doi.org/10.1109/MSP.2008.931079.

Faul, F. et al. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4), pp. 1149–1160, https://doi.org/10.3758/BRM.41.4.1149.

Ferragina, P. and Scaiella, U. (2010). TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, pp. 1625–1628, https://doi.org/10.1145/1871437.1871689.

Ferrara, E. et al. (2016). The rise of social bots. *Communications of the ACM*, 59(7), pp. 96–104, https://doi.org/10.1145/2818717.

Figueira, Á. and Oliveira, L. (2017). The current state of fake news: Challenges and opportunities. *Procedia Computer Science*, 121, pp. 817–825, https://doi.org/10.1016/j.procs.2017.11.106.

Francis, E. et al. (2023). Transparency, trust, and security needs for the design of digital news authentication tools. In *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, ACM, pp. 101:1–101:44, https://doi.org/10.1145/3579534.

Free, C. (2015). Looking through the fraud triangle: A review and call for new directions. *Meditari Accountancy Research*, 23(2), pp. 175–196, https://doi.org/10.1108/MEDAR-02-2015-0009.

Freiling, I. et al. (2023). Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during covid-19. *New Media & Society*, 25(1), pp. 141–162, https://doi.org/10.1177/14614448211011451.

Froehlich, T. J. (2020). Ten lessons for the age of disinformation. In *Navigating Fake News, Alternative Facts, and Misinformation in a Post-truth World*, IGI Global, pp. 36–88, https://doi.org/10.4018/978-1-7998-2543-2.ch003.

Gallina, Y., Boudin, F. and Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, ACL, pp. 130–135, https://doi.org/10.18653/v1/W19-8617.

García Lozano, M. et al. (2020). Veracity assessment of online data. *Decision Support Systems*, 129, https://doi.org/10.1016/j.dss.2019.113132.

Gay, C. W., Kayaalp, M. and Aronson, A. R. (2005). Semi-automatic indexing of full text biomedical articles. In *Proceedings of the 2005 AMIA Symposium*, American Medical Informatics Association (AMIA), pp. 271–275, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560666/.

Gazendam, L., Wartena, C. and Brussee, R. (2010). Thesaurus based term ranking for keyword extraction. In *Proceedings of the 2010 Workshops on Database and*

Expert Systems Applications, IEEE, pp. 49–53, https://doi.org/10.1109/DEXA.2010.31.

Ghasemaghaei, M. and Hassanein, K. (2016). A macro model of online information quality perceptions: A review and synthesis of the literature. *Computers in Human Behavior*, 55, pp. 972–991, https://doi.org/10.1016/j.chb.2015.09.027.

Ginsca, A. L., Popescu, A. and Lupu, M. (2015). Credibility in information retrieval. *Foundations and Trends in Information Retrieval*, 9(5), pp. 355–475, https://doi.org/10.1561/1500000046.

Google (2022). Search quality rater guidelines: An overview. https://services.google.com/fh/files/misc/hsw-sqrg.pdf.

Graves, L. (2017). Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique*, 10(3), pp. 518–537, https://doi.org/10.1111/cccr.12163.

Graves, L. (2018). Understanding the promise and limits of automated fact-checking. Tech. rep., Reuters Institute for the Study of Journalism, https://doi.org/10.60625/risj-nqnx-bg89.

Graves, L. and Cherubini, F. (2016). The rise of fact-checking sites in Europe. *Digital News Project Report*, https://doi.org/10.60625/risj-tdn4-p140.

Griffiths, M. (2010). Crime and gambling: A brief overview of gambling fraud on the internet. *Internet Journal of Criminology*, pp. 1–7, https://www.internetjournalofcriminology.com/_files/ugd/b93dd4_7504744e957b403ab18d03b176e6772f.pdf.

Gruzd, A. and Mai, P. (2020). Inoculating against an infodemic: A Canada-wide

COVID-19 news, social media, and misinformation survey. Tech. rep., Ryerson University Social Media Lab, https://doi.org/10.5683/SP2/JLULYA.

Guo, B. et al. (2020). The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys*, 53(4), https://doi.org/10.1145/3393880.

Guo, Z., Schlichtkrull, M. and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10, pp. 178–206, https://doi.org/10.1162/tacl_a_00454.

Gupta, V. et al. (2021). Supporting verification of news articles with automated search for semantically similar articles. In *Proceedings of the 2021 Workshop Reducing Online Misinformation through Credible Information Retrieval*, CEUR Workshop Proceedings, pp. 47–58, https://ceur-ws.org/Vol-2838/paper5.pdf.

Hamel, L. et al. (2021a). KFF COVID-19 vaccine monitor: June 2021. Tech. rep., KFF, https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-june-2021/.

Hamel, L. et al. (2021b). KFF COVID-19 vaccine monitor: Media and misinformation. Tech. rep., KFF, https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-media-and-misinformation/.

Hameleers, M. (2023). Disinformation as a context-bound phenomenon: Toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, 33(1), pp. 1–10, https://doi.org/10.1093/ct/qtac021.

Hanselowski, A. (2020). *A Machine-Learning-Based Pipeline Approach to Automated Fact-Checking*. Ph.D. thesis, Technische Universität Darmstadt, https://doi.org/10.25534/tuprints-00014136.

Harjule, P. et al. (2023). Analysing misinformation sharing amongst college students in India during COVID-19. *Procedia Computer Science*, 218, pp. 671–685, https://doi.org/10.1016/j.procs.2023.01.048.

Hasanain, M. and Elsayed, T. (2022). Studying effectiveness of web search for fact checking. *Journal of the Association for Information Science and Technology*, 73(5), pp. 738–751, https://doi.org/10.1002/asi.24577.

Hashem, F. S. and Sulong, G. (2015). Passive approaches for detecting image tampering: A review. *Jurnal Teknologi*, 73(2), pp. 31–36, https://doi.org/10.11113/jt.v73.4189.

Hassan, N., Li, C. and Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ACM, pp. 1835–1838, https://doi.org/10.1145/2806416.2806652.

Hassan, N. et al. (2015). The quest to automate fact-checking. In *Proceedings of the 2015 Computation+Journalism Symposium*, pp. 1–5, http://cj2015.brown.columbia.edu/papers/automate-fact-checking.pdf.

He, L., Hu, S. and Pei, A. (2023). Debunking disinformation: Revolutionizing truth with nlp in fake news detection. https://doi.org/10.48550/arXiv.2308.16328.

Heydari, A. et al. (2015). Detection of review spam: A survey. *Expert Systems with Applications*, 42(7), pp. 3634–3642, https://doi.org/10.1016/j.eswa.2014.12.029.

Hidey, C. et al. (2020). DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational

Linguistics, pp. 8593–8606, https://doi.org/10.18653/v1/2020.acl-main.761.

High-Level Expert Group (HLEG) on Fake News and Online Disinformation (2018). A multi-dimensional approach to disinformation. Tech. rep., Brussels: European Commission, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.

Hoes, E., Altay, S. and Bermeo, J. (2023). Leveraging ChatGPT for efficient fact-checking. https://doi.org/10.31234/osf.io/qnjkf.

Hrckova, A. et al. (2022). Automated, not automatic: Needs and practices in european fact-checking organizations as a basis for designing human-centered ai systems. https://doi.org/10.48550/arXiv.2211.12143.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 216–223, https://aclanthology.org/W03-1028.

Hulth, A. et al. (2001). Automatic keyword extraction using domain knowledge. In *Computational Linguistics and Intelligent Text Processing: Procedings of the Second International Conference on Intelligent Text Processing and Computational Linguistics*, *Lecture Notes in Computer Science*, vol. 2004, Springer, pp. 472–482, https://doi.org/10.1007/3-540-44686-9_47.

Huynh, T. K. et al. (2015). A survey on image forgery detection techniques. In *Proceedings of the 2015 IEEE International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future*, pp. 71–76, http://doi.org/10.1109/RIVF.2015.7049877.

Ibrahim, Y., Safieddine, F. and Pourghomi, P. (2023). Attitudes to fake news verification: Youth orientations to 'right click' authenticate. *Journal of Applied*

Journalism & Media Studies, 12(1), pp. 77–97, https://doi.org/10.1386/ajms_00051_1.

Ipsos MORI (2022). Ipsos MORI veracity index. https://www.ipsos.com/en-uk/ipsos-veracity-index-2022.

Islam, M. S. et al. (2020). COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4), pp. 1621–1629, https://doi.org/10.4269/ajtmh.20-0812.

Jabiyev, B. et al. (2021). FADE: Detecting fake news articles on the web. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, ACM, pp. 15:1–15:10, https://doi.org/10.1145/3465481.3465751.

Jack, C. (2017). Lexicon of lies: Terms for problematic information. Tech. rep., Data & Society Research Institute, https://datasociety.net/wp-content/uploads/2017/08/DataAndSociety_LexiconofLies.pdf.

Jamil, N. B. C. E. et al. (2015). A systematic review on the profiling of digital news portal for big data veracity. *Procedia Computer Science*, 72, pp. 390–397, the Third Information Systems International Conference, http://doi.org/10.1016/j.procs.2015.12.154.

Jandaghi, P. and Pujara, J. (2023). Identifying quantifiably verifiable statements from text. In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data*, ACL, pp. 14–22, https://doi.org/10.18653/v1/2023.matching-1.2.

Jaster, R. and Lanius, D. (2021). Speaking of fake news: Definitions and dimensions. In *The Epistemology of Fake News*, Oxford University Press, pp. 19–45, https://doi.org/10.1093/oso/9780198863977.003.0002.

Jauhiainen, T. et al. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65, pp. 675–782, https://doi.org/10.1613/jair.1.11675.

Johnston, P. and Elyan, E. (2019). A review of digital video tampering: From simple editing to full synthesis. *Digital Investigation*, 29, pp. 67–81, https://doi.org/10.1016/j.diin.2019.03.006.

Jones, H. S. and Moncur, W. (2018). The role of psychology in understanding online trust. *Psychological and Behavioral Examinations in Cyber Security*, pp. 109–132, https://doi.org/10.4018/978-1-5225-4053-3.ch007.

Jones, K. et al. (2022). A comprehensive survey of natural language generation advances from the perspective of digital deception. https://doi.org/10.48550/arXiv.2208.05757.

Joudaki, H. et al. (2015). Using data mining to detect health care fraud and abuse: A review of literature. *Global Journal of Health Science*, 7(1), pp. 194–202, https://doi.org/10.5539/gjhs.v7n1p194.

Juneja, P. and Mitra, T. (2022). Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), https://doi.org/10.1145/3555143.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1), pp. 9–27, https://doi.org/10.1017/S1351324900000048.

Kalsnes, B. (2018). Fake news. *Oxford Research Encyclopedia of Communication, Oxford University Press*, https://doi.org/10.1093/acrefore/9780190228613.013.809.

Kandul, S. et al. (2023). Explainable AI: A review of the empirical literature. https://doi.org/10.2139/ssrn.4325219.

Kapantai, E. et al. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5), pp. 1301–1326, https://doi.org/10.1177/1461444820959296.

Kareem, W. and Abbas, N. (2023). Fighting lies with intelligence: Using large language models and chain of thoughts technique to combat fake news. In *Artificial Intelligence XL: Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Springer, pp. 253–258, https://doi.org/10.1007/978-3-031-47994-6_24.

Karlova, N. A. and Lee, J. H. (2011). Notes from the underground city of disinformation: A conceptual investigation. *Proceedings of the American Society for Information Science and Technology*, 48(1), pp. 1–9, https://doi.org/10.1002/meet.2011.14504801133.

Kartal, Y. S. and Kutlu, M. (2023). Re-think before you share: A comprehensive study on prioritizing check-worthy claims. *IEEE Transactions on Computational Social Systems*, 10(1), pp. 362–375, https://doi.org/10.1109/TCSS.2021.3138642.

Kaur, H. and Jindal, N. (2020). Image and video forensics: A critical survey. *Wireless Personal Communications*, 112, pp. 1281–1302, https://doi.org/10.1007/s11277-020-07102-x.

Kempf, A. O. and Neubert, J. (2016). The role of thesauri in an open web: A case study of the STW thesaurus for economics. *KO Knowledge Organization*, 43(3), pp. 160–173, https://doi.org/10.5771/0943-7444-2016-3-160.

Kevin, V. et al. (2018). Information nutrition labels: A plugin for online news

evaluation. In *Proceedings of the 1st Workshop on Fact Extraction and VERification*, ACL, pp. 28–33, https://doi.org/10.18653/v1/W18-5505.

Khodabakhsh, A., Busch, C. and Ramachandra, R. (2018). A taxonomy of audiovisual fake multimedia content creation technology. In *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval*, IEEE, pp. 372–377, http://doi.org/10.1109/MIPR.2018.00082.

Kim, S. N. et al. (2010). SemEval-2010 Task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, ACL, pp. 21–26, https://aclanthology.org/S10-1004.

Kingra, S., Aggarwal, N. and Singh, R. D. (2016). Video inter-frame forgery detection: A survey. *Indian Journal of Science & Technology*, 9(44), pp. 1–9, https://doi.org/10.17485/ijst/2016/v9i44/105142.

Konstantinovskiy, L. et al. (2021). Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2), https://doi.org/10.1145/3412869.

Korus, P. (2017). Digital image integrity – A survey of protection and verification techniques. *Digital Signal Processing*, 71, pp. 1–26, https://doi.org/10.1016/j.dsp.2017.08.009.

Kosti, N., Levi-Faur, D. and Mor, G. (2019). Legislation and regulation: Three analytical distinctions. *The Theory and Practice of Legislation*, 7(3), pp. 169–178, https://doi.org/10.1080/20508840.2019.1736369.

Kotonya, N. and Toni, F. (2020). Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, pp. 5430–5443, https://doi.org/10.18653/v1/2020.coling-main.474.

Krapivin, M., Autaeu, A. and Marchese, M. (2009). Large dataset for keyphrases extraction. Departmental Technical Report DISI-09-055, University of Trento, Italy, http://eprints.biblio.unitn.it/1671/.

Krieg, S. J. et al. (2020). Pandemic pulse: Unraveling and modeling social signals during the COVID-19 pandemic. *Digital Government: Research and Practice*, 2(2), pp. 19:1–19:9, https://doi.org/10.1145/3431805.

Krithiga, R. and Ilavarasan, E. (2019). A comprehensive survey of spam profile detection methods in online social networks. In *Proceedings of the 2019 International Conference on Physics and Photonics Processes in Nano Sciences*, vol. 1362, IOP Publishing, pp. 012111:1–012111:13, https://doi.org/10.1088/1742-6596/1362/1/012111.

Kucuk, D. and Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys*, 53(1), pp. 12:1–12:37, https://doi.org/10.1145/3369026.

Kumar, S. and Shah, N. (2018). False information on web and social media: A survey. https://doi.org/10.48550/arXiv.1804.08559.

Kyriakidou, M. et al. (2022). Questioning fact-checking in the fight against disinformation: An audience perspective. *Journalism Practice*, pp. 1–17, https://doi.org/10.1080/17512786.2022.2097118.

La Barbera, D., Roitero, K. and Mizzaro, S. (2022). A hybrid human-in-the-loop framework for fact checking. In *Proceedings of the 6th Workshop on Natural Language for Artificial Intelligence*, CEUR Workshop Proceedings, pp. 4:1–4:10, https://ceur-ws.org/Vol-3287/paper4.pdf.

Lampou, E. and Antonopoulos, N. (2023). Ranked by truth metrics: A new communication method approach, on crowd-sourced fact-checking platforms for journalistic and social media content. *Studies in Media and Communication*, 11(6), pp. 231–243, https://doi.org/10.11114/smc.v11i6.6166.

Lawson, M. A., Anand, S. and Kakkar, H. (2023). Tribalism and tribulations: The social costs of not sharing fake news. *Journal of Experimental Psychology: General*, https://doi.org/10.1037/xge0001374.

Lazer, D. M. J. et al. (2018). The science of fake news. *Science*, 359(6380), pp. 1094–1096, https://doi.org/10.1126/science.aao2998.

Leite, J. A. et al. (2023). Detecting misinformation with LLM-predicted credibility signals and weak supervision. https://doi.org/10.48550/arXiv.2309.07601.

Lemieux, V. and Smith, T. D. (2018). Leveraging archival theory to develop a taxonomy of online disinformation. In *2018 IEEE International Conference on Big Data*, pp. 4420–4426, https://doi.org/10.1109/BigData.2018.8622391.

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, 66(9), pp. 1763–1775, https://doi.org/10.1002/asi.23304.

Lewandowsky, S. and van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2), pp. 348–384, https://doi.org/10.1080/10463283.2021.1876983.

Lewandowsky, S. et al. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), pp. 106–131, https://doi.org/10.1177/1529100612451018.

Li, G. and Wang, H. (2014). Improved automatic keyword extraction based on TextRank using domain knowledge. In *Proceedings of the Third CCF International Conference on Natural Language Processing and Chinese Computing*,

Communications in Computer and Information Science, vol. 496, Springer, pp. 403–413, https://doi.org/10.1007/978-3-662-45924-9_36.

Li, M. et al. (2024). Self-Checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Association for Computational Linguistics, pp. 163–181, https://doi.org/10.18653/v1/2024.findings-naacl.12.

Li, Y. et al. (2016). A survey on truth discovery. *SIGKDD Explorations Newsletter*, 17(2), pp. 1–16, http://doi.org/10.1145/2897350.2897352.

Lian, S. and Zhang, Y. (2010). Multimedia forensics for detecting forgeries. In *Handbook of Information and Communication Security*, Springer, pp. 809–828, https://doi.org/10.1007/978-3-642-04117-4_37.

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, 5(3), https://doi.org/10.1177/2053168018786848.

Lim, G. and Perrault, S. T. (2023). XAI in automated fact-checking? the benefits are modest and there's no one-explanation-fits-all. https://doi.org/10.48550/arXiv.2308.03372.

Linder, R. et al. (2021). How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, 2(4), pp. e49:1–e49:19, https://doi.org/10.1002/ail2.49.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3), pp. 265–266, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/.

Liu, H. et al. (2023). Human-centered NLP fact-checking: Co-Designing with fact-checkers using matchmaking for AI. https://doi.org/10.48550/arXiv.2308.07213.

Lovelace, R. et al. (2016). From big noise to big data: Toward the verification of large data sets for understanding regional retail flows. *Geographical Analysis*, 48(1), pp. 59–81, https://doi.org/10.1111/gean.12081.

Lu, Y. and Li, S. (2022). From data flows to privacy-benefit trade-offs: A user-centric semantic model. *Security and Privacy*, 5(4), https://doi.org/10.1002/spy2.225.

Mahdian, B. and Saic, S. (2010). A bibliography on blind methods for identifying image forgery. *Signal Processing: Image Communication*, 25(6), pp. 389–399, https://doi.org/10.1016/j.image.2010.05.003.

Mahmood, T. et al. (2015). A survey on block based copy move image forgery detection techniques. In *Proceedings of the 2015 International Conference on Emerging Technologies*, pp. 1–6, http://doi.org/10.1109/ICET.2015.7389169.

Marietta, M., Barker, D. C. and Bowser, T. (2015). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, 13(4), pp. 577–596, https://doi.org/10.1515/for-2015-0040.

Martel, C. et al. (2023). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, pp. 1–12, https://doi.org/10.1177/17456916231190388.

Martín, A. et al. (2022). FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, 251, pp. 109265:1–109265:19, https://doi.org/10.1016/j.knosys.2022.109265.

Marujo, L. et al. (2013). Supervised topical key phrase extraction of news stories

using crowdsourcing, light filtering and co-reference normalization. https://doi.org/10.48550/arXiv.1306.4886.

McClure Haughey, M., Povolo, M. and Starbird, K. (2022). Bridging contextual and methodological gaps on the "misinformation beat": Insights from journalist-researcher collaborations at speed. In *Proceedings of the 2022 ACM SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 244:1–244:15, https://doi.org/10.1145/3491102.3517503.

Medelyan, O., Frank, E. and Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 1318–1327, https://aclanthology.org/D09-1137.

Medelyan, O. and Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, pp. 296–297, https://doi.org/10.1145/1141753.1141819.

Medelyan, O. and Witten, I. H. (2008). Domain-independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology*, 59(7), pp. 1026–1040, https://doi.org/10.1002/asi.20790.

Medelyan, O., Witten, I. H. and Milne, D. (2008). Topic indexing with wikipedia. In *Proceedings of the 2008 AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, AAAI, pp. 19–24, https://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-004.pdf.

Mena, P. (2019). Principles and boundaries of fact-checking: Journalists' perceptions. *Journalism Practice*, 13(6), pp. 657–672, https://doi.org/10.1080/17512786.2018.1547655.

Merrouni, Z. A., Frikh, B. and Ouhbi, B. (2020). Automatic keyphrase extraction: A survey and trends. *Journal of Intelligent Information Systems*, 54(2), pp. 391–424, https://doi.org/10.1007/s10844-019-00558-9.

Micallef, N. et al. (2022). True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), pp. 127:1–127:44, https://doi.org/10.1145/3512974.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, https://aclanthology.org/W04-3252/.

Milani, S. et al. (2012). An overview on video forensics. *APSIPA Transactions on Signal and Information Processing*, 1, pp. e2:1–e2:18, http://doi.org/10.1017/ATSIP.2012.2.

Miller, D. (2019). Leveraging BERT for extractive text summarization on lectures. https://doi.org/10.48550/arXiv.1906.04165.

Mirza, S. et al. (2023). Tactics, threats and targets: Modeling disinformation and its mitigation. In *Proceedings of the 2023 Network and Distributed System Security Symposium*, pp. 1–18, https://doi.org/10.14722/ndss.2023.23657.

Mizher, M. A. et al. (2017). A review of video falsifying techniques and video forgery detection techniques. *International Journal of Electronic Security and Digital Forensics*, 9(3), pp. 191–208, https://doi.org/10.1504/IJESDF.2017.085196.

Modirrousta-Galian, A. and Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, https://doi.org/10.1037/xge0001395.

Mohammadi, E. et al. (2022). Identifying frames of the COVID-19 infodemic: Thematic analysis of misinformation stories across media. *JMIR Infodemiology*, 2(1), pp. e33827:1–e33827:16, https://doi.org/10.2196/33827.

Mohawesh, R. et al. (2023). Semantic graph based topic modelling framework for multilingual fake news detection. *AI Open*, 4, pp. 33–41, https://doi.org/10.1016/j.aiopen.2023.08.004.

Mohseni, S. et al. (2021). Machine learning explanations to prevent overtrust in fake news detection. AAAI Press, pp. 421–431, https://doi.org/10.1609/icwsm.v15i1.18072.

Molina, M. D. et al. (2021). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, 65(2), pp. 180–212, https://doi.org/10.1177/0002764219878224.

Myers, B. A. and Stylos, J. (2016). Improving API usability. *Communications of the ACM*, 59(6), pp. 62–69, http://doi.org/10.1145/2896587.

Nakov, P. et al. (2021). Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, pp. 4551–4558, survey Track, https://doi.org/10.24963/ijcai.2021/619.

Nane, G. F. et al. (2023). COVID-19 and the scientific publishing system: Growth, open access and scientific fields. *Scientometrics*, 128(1), pp. 345–362, https://doi.org/10.1007/s11192-022-04536-x.

Neumann, T., De-Arteaga, M. and Fazelpour, S. (2022). Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability,*

and Transparency, ACM, pp. 1504–1515, https://doi.org/10.1145/3531146.3533205.

Newman, N. et al. (2023). Digital news report 2023. Tech. rep., Reuters Institute for the Study of Journalism, https://coilink.org/20.500.12592/3sq026.

Ngai, E. W. T. et al. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), pp. 559–569, https://doi.org/10.1016/j.dss.2010.08.006.

Nguyen, A. T. et al. (2018). Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, ACM, pp. 189–199, https://doi.org/10.1145/3242587.3242666.

Nguyen, T. D. and Kan, M.-Y. (2007). Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers: Proceedings of the 10th International Conference on Asian Digital Libraries*, *Lecture Notes in Computer Science*, vol. 4822, Springer, pp. 317–326, https://doi.org/10.1007/978-3-540-77094-7_41.

Nieminen, S. and Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3), pp. 296–309, https://doi.org/10.1177/1478929918786852.

Noy, N. F., McGuinness, D. L. et al. (2001). Ontology development 101: A guide to creating your first ontology. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

Orman, L. (1984). Fighting information pollution with decision support systems. *Journal of Management Information Systems*, 1(2), pp. 64–71, https://doi.org/10.1080/07421222.1984.11517704.

Osborne, F. and Motta, E. (2015). Klink-2: Integrating multiple web sources to generate semantic topic networks. In *The Semantic Web: Proceedings of the 14th International Semantic Web Conference, Part I*, *Lecture Notes in Computer Science*, vol. 9366, Springer, pp. 408–424, https://doi.org/10.1007/978-3-319-25007-6_24.

Palmer, A. (2020). *Scientific Facts in the Space of Public Reason: Moderate Idealization, Public Justification, and Vaccine Policy Under Conditions of Widespread Misinformation and Conspiracism*. Ph.D. thesis, Bowling Green State University, https://scholarworks.bgsu.edu/philosophy_diss/43/.

Panjwani, A. (2022). Videos claiming to show Ukraine invasion sourced to old TikTok videos. https://fullfact.org/online/ukraine-invasion-tiktok-clip/.

Papagiannopoulou, E. and Tsoumakas, G. (2018). Local word vectors guiding keyphrase extraction. *Information Processing & Management*, 54(6), pp. 888–902, https://doi.org/10.1016/j.ipm.2018.06.004.

Papagiannopoulou, E. and Tsoumakas, G. (2020). A review of keyphrase extraction. *WIREs Data Mining and Knowledge Discovery*, 10(2), pp. e1339:1–e1339:45, https://doi.org/10.1002/widm.1339.

Parashar, N. and Tiwari, N. (2015). A survey of digital image tampering techniques. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(10), pp. 91–96, http://doi.org/10.14257/ijsip.2015.8.10.10.

Pasi, G. and Viviani, M. (2020). Information credibility in the social web: Contexts, approaches, and open issues. https://doi.org/10.48550/arXiv.2001.09473.

Pay, T. (2016). Totally automated keyword extraction. In *Proceedings of the 2016 IEEE International Conference on Big Data*, IEEE, pp. 3859–3863, https://doi.org/10.1109/BigData.2016.7841059.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830, https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Petter Bae Brandtzaeg, A. F. and Ángeles Chaparro Domínguez, M. (2018). How journalists and social media users perceive online fact-checking and verification services. *Journalism Practice*, 12(9), pp. 1109–1129, https://doi.org/10.1080/17512786.2017.1363657.

Philipp Schmidt, F. B. and Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4), pp. 260–278, https://doi.org/10.1080/12460125.2020.1819094.

Phua, C. et al. (2012). A comprehensive survey of data mining-based fraud detection research. *Computers in Human Behavior*, 28(3), pp. 1002–1013, https://doi.org/10.1016/j.chb.2012.01.002.

Pierri, F. and Ceri, S. (2019). False news on social media: A data-driven survey. *ACM SIGMOD Record*, 48(2), pp. 18–27, https://doi.org/10.1145/3377330.3377334.

Pirrong, C. (2017). The economics of commodity market manipulation: A survey. *Journal of Commodity Markets*, 5, pp. 1–17, https://doi.org/10.1016/j.jcomm.2017.02.001.

Piva, A. (2013). An Overview on Image Forensics. *International Scholarly Research Notices Signal Processing*, 2013, pp. 496701:1–496701:22, http://doi.org/10.1155/2013/496701.

Poisel, R. and Tjoa, S. (2011). Forensics investigations of multimedia data: A review of the state-of-the-art. In *Proceedings of the Sixth International Conference on IT Security Incident Management and IT Forensics*, IEEE, pp. 48–61, https://doi.org/10.1109/IMF.2011.14.

Posetti, J. and Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. Tech. rep., International Center for Journalists, https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf.

Poynter Institute for Media Studies, MediaWise and YouGov Inc. (2022). A global study on information literacy: Understanding generational behaviors and concerns around false and misleading information online. Tech. rep., https://www.poynter.org/wp-content/uploads/2022/08/A-Global-Study-on-Information-Literacy-1.pdf.

Procter, R. et al. (2023). Some observations on fact-checking work with implications for computational support. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, AAAI, pp. 28:1–28:11, https://doi.org/10.36190/2023.28.

Pu, D. et al. (2024). SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, ELRA and ICCL, pp. 14429–14444, https://aclanthology.org/2024.lrec-main.1258.

Purificato, E., Shahania, S. and De Luca, E. W. (2022). Tell me why it's fake: Developing an explainable user interface for a fake news detection system. In *Proceedings of the 2022 Italian Workshop on Explainable Artificial Intelligence*,

CEUR Workshop Proceedings, pp. 51–63, https://ceur-ws.org/Vol-3277/paper4.pdf.

Putniņš, T. J. (2012). Market manipulation: A survey. *Journal of Economic Surveys*, 26(5), pp. 952–967, https://doi.org/10.1111/j.1467-6419.2011.00692.x.

Qazi, T. et al. (2013). Survey on blind image forgery detection. *IET Image Processing*, 7(7), pp. 660–670, https://doi.org/10.1049/iet-ipr.2012.0388.

Quelle, D. and Bovet, A. (2023). The perils & promises of fact-checking with large language models. https://doi.org/10.48550/arXiv.2310.13549.

Qureshi, M. A. and Deriche, M. (2015). A bibliography of pixel-based blind image forgery detection techniques. *Signal Processing: Image Communication*, 39, pp. 46–74, https://doi.org/10.1016/j.image.2015.08.008.

Ramalingam, D. and Chinnaiah, V. (2018). Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering*, 65, pp. 165–177, https://doi.org/10.1016/j.compeleceng.2017.05.020.

Redi, J. A., Taktak, W. and Dugelay, J.-L. (2011). Digital image forensics: A booklet for beginners. *Multimedia Tools and Applications*, 51(1), pp. 133–162, https://doi.org/10.1007/s11042-010-0620-1.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, pp. 45–50, http://is.muni.cz/publication/884893/en.

Reuter, C., Kaufhold, M.-A. and Steinfort, R. (2017). Rumors, fake news and social bots in conflicts and emergencies: Towards a model for believability in

social media. In *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*, ISCRAM, pp. 583–591, http://tubiblio.ulb.tu-darmstadt.de/108176/.

Reuters (2021). Trump's FCC chairman does not object to Facebook, Twitter blocking president. https://www.reuters.com/article/us-usa-trump-social-media-idUSKBN29C386.

Rocha, A. et al. (2011). Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Computing Surveys*, 43(4), pp. 26:1–26:42, https://doi.org/10.1145/1978802.1978805.

Rodríguez-Pérez, C. et al. (2022). Purposes, principles, and difficulties of fact-checking in Ibero-America: Journalists' perceptions. *Journalism Practice*, pp. 1–19, https://doi.org/10.1080/17512786.2022.2124434.

Romero, L. (2022). How 'War on Fakes' uses fact-checking to spread pro-Russia propaganda. https://www.politifact.com/article/2022/aug/08/how-war-fakes-uses-fact-checking-spread-pro-russia/.

Rose, M. E. and Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, 10, https://doi.org/10.1016/j.softx.2019.100263.

Rose, S. et al. (2010). Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory*, Wiley, chap. 1, pp. 1–20, https://doi.org/10.1002/9780470689646.ch1.

Rosenthol, L. (2022). C2PA: The world's first industry standard for content provenance. In *Applications of Digital Image Processing XLV*, vol. 12226, International Society for Optics and Photonics, SPIE, https://doi.org/10.1117/12.2632021.

Rothkopf, D. J. (2003). When the buzz bites back. *The Washington Post*, https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd.

Roy, P. K. et al. (2023). Securing social platform from misinformation using deep learning. *Computer Standards & Interfaces*, 84, https://doi.org/10.1016/j.csi.2022.103674.

Rubin, V. and Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24(1), https://doi.org/10.7152/acro.v24i1.14671.

Rubin, V. L., Chen, Y. and Conroy, N. J. (2015). Deception detection for news: Three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, American Society for Information Science, pp. 1–4, https://doi.org/10.1002/pra2.2015.145052010083.

Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. *Informatica Economica*, 16(1), pp. 110–122, http://revistaie.ase.ro/content/61/10%20-%20sabau.pdf.

Sakhnini, N. and Chattopadhyay, D. (2022). A review of smartphone fact-checking apps and their (non) use among older adults. In *Adjunct Publication of the 24th International Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, pp. 24:1–24:8, https://doi.org/10.1145/3528575.3551448.

Salatino, A. A. et al. (2018). The Computer Science Ontology: A large-scale taxonomy of research areas. In *The Semantic Web: Proceedings of the 17th International Semantic Web Conference, Part II*, *Lecture Notes in Computer*

Science, vol. 11137, Springer, pp. 187–205, https://doi.org/10.1007/978-3-030-00668-6_12.

Saling, L. L. et al. (2021). No one is immune to misinformation: An investigation of misinformation sharing by subscribers to a fact-checking newsletter. *PLOS ONE*, 16(8), pp. e0255702:1–e0255702:13, https://doi.org/10.1371/journal.pone.0255702.

Samuel, H. and Zaïane, O. (2018). MedFact: Towards improving veracity of medical information in social media using applied machine learning. In *Advances in Artificial Intelligence: Proceedings of the 31st Canadian Conference on Artificial Intelligence*, Springer, pp. 108–120, https://doi.org/10.1007/978-3-319-89656-4_9.

Saquete, E. et al. (2020). Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141, pp. 112943:1–112943:27, https://doi.org/10.1016/j.eswa.2019.112943.

Saunders, L. (2022). Faculty perspectives on mis- and disinformation across disciplines. *College & Research Libraries*, 83(2), https://doi.org/10.5860/crl.83.2.221.

Schutz, A. T. (2008). *Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods*. Master's thesis, National University of Ireland, Galway, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=08b75d31a90f206b36e806a7ec372f6f0d12457e.

Seböck, W., Biron, B. and Lampoltshammer, T. J. (2023). Barriers to the introduction of artificial intelligence to support communication experts in media and the public sector to combat fake news and misinformation. In *Electronic Participation: Proceedings of the 2023 International Conference on Electronic*

Participation, Springer, pp. 67–81, https://doi.org/10.1007/978-3-031-41617-0_5.

Sharma, D. K. and Garg, S. (2021). IFND: A benchmark dataset for fake news detection. *Complex & Intelligent Systems*, pp. 1–21, https://doi.org/10.1007/s40747-021-00552-1.

Sharma, K. et al. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3), pp. 21:1–21:42, https://doi.org/10.1145/3305260.

Sharma, S. and Dhavale, S. V. (2016). A review of passive forensic techniques for detection of copy-move attacks on digital videos. In *Proceedings of the Third International Conference on Advanced Computing and Communication Systems*, vol. 01, pp. 1–6, https://doi.org/10.1109/ICACCS.2016.7586396.

Shelke, S. and Attar, V. (2019). Source detection of rumor in social network – A review. *Online Social Networks and Media*, 9, pp. 30–42, https://doi.org/10.1016/j.osnem.2018.12.001.

Sheoran, A., Jadhav, G. V. and Sarkar, A. (2022). Submodrank: Monotone submodularity for opinionated key-phrase extraction. In *Proceedings of the IEEE 16th International Conference on Semantic Computing*, pp. 159–166, https://doi.org/10.1109/ICSC52841.2022.00032.

Shi, L. et al. (2022). The effects of interactive AI design on user behavior: An eye-tracking study of fact-checking COVID-19 claims. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, ACM, pp. 315–320, https://doi.org/10.1145/3498366.3505786.

Shi, T. et al. (2008). Improving keyphrase extraction using wikipedia semantics. In *Proceedings of the 2nd International Symposium on Intelligent Information*

Technology Application, vol. 2, IEEE, pp. 42–46, https://doi.org/10.1109/IITA.2008.211.

Shih-Yi Chien, C.-J. Y. and Yu, F. (2022). Xflag: Explainable fake news detection model on social media. *International Journal of Human–Computer Interaction*, 38(18–20), pp. 1808–1827, https://doi.org/10.1080/10447318.2022.2062113.

Shin, J. and Chan-Olmsted, S. (2023). User perceptions and trust of explainable machine learning fake news detectors. *International Journal of Communication*, 17, https://ijoc.org/index.php/ijoc/article/view/19534.

Shu, K., Bernard, H. R. and Liu, H. (2019). Studying fake news via network analysis: Detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, Springer, pp. 43–65, https://doi.org/10.1007/978-3-319-94105-9_3.

Shu, K. et al. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter*, 19(1), pp. 22–36, https://doi.org/10.1145/3137597.3137600.

Shu, K. et al. (2020). *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, Springer, chap. Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. pp. 1–19, https://doi.org/10.1007/978-3-030-42699-6_1.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the 2014 Workshop on Interactive Language Learning, Visualization, and Interfaces*, ACL, pp. 63–70, https://doi.org/10.3115/v1/W14-3110.

Singhal, N. and Gandhani, S. (2015). Analysis of copy-move forgery image forensics: A review. *International Journal of Signal Processing, Image Processing*

and Pattern Recognition, 8(7), pp. 265–272, http://doi.org/10.14257/ijsip.2015.8.7.25.

Sitara, K. and Mehtre, B. M. (2016). Digital video tampering detection: An overview of passive techniques. *Digital Investigation*, 18, pp. 8–22, https://doi.org/10.1016/j.diin.2016.06.003.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), pp. 143–178, https://aclanthology.org/J93-1007.

Søe, S. O. (2017). Algorithmic detection of misinformation and disinformation: Gricean perspectives. *Journal of Documentation*, 74(2), pp. 309–332, https://doi.org/10.1108/JD-05-2017-0075.

Søe, S. O. (2021). A unified account of information, misinformation, and disinformation. *Synthese*, 198(6), pp. 5929–5949, https://doi.org/10.1007/s11229-019-02444-x.

Song, X. et al. (2021). Classification aware neural topic model for COVID-19 disinformation categorisation. *PLOS ONE*, 16(2), pp. 1–22, https://doi.org/10.1371/journal.pone.0247086.

Sood, A. (2022). Video does not show 'Russian jets over Ukraine' – it has circulated in old posts about air show rehearsal. https://factcheck.afp.com/doc.afp.com.323T8CJ.

Soprano, M. et al. (2024). Cognitive biases in fact-checking and their countermeasures: A review. *Information Processing & Management*, 61(3), pp. 103672:1–103672:29, https://doi.org/10.1016/j.ipm.2024.103672.

Stamm, M. C., Wu, M. and Liu, K. J. R. (2013). Information forensics: An overview of the first decade. *IEEE Access*, 1, pp. 167–200, https://doi.org/10.1109/ACCESS.2013.2260814.

Sun, Y. et al. (2020). SIFRank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, 8, pp. 10896–10906, https://doi.org/10.1109/ACCESS.2020.2965087.

Sundriyal, M. et al. (2022). Document retrieval and claim verification to mitigate COVID-19 misinformation. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, ACL, pp. 66–74, https://doi.org/10.18653/v1/2022.constraint-1.8.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Tandoc Jr., E. C., Lim, Z. W. and Ling, R. (2018). Defining "fake news". *Digital Journalism*, 6(2), pp. 137–153, https://doi.org/10.1080/21670811.2017.1360143.

Tao, J., Jia, L. and You, Y. (2017). Review of passive-blind detection in digital video forgery based on sensing and imaging techniques. In *Proceedings of the 2017 International Conference on Optoelectronics and Microelectronics Technology and Application*, vol. 10244, International Society for Optics and Photonics, pp. 102441C:1–102441C:6, https://doi.org/10.1117/12.2267503.

Tchechmedjiev, A. et al. (2019). ClaimsKG: A knowledge graph of fact-checked claims. In *The Semantic Web – Proceedings of the 18th International Semantic Web Conference*, Springer, pp. 309–324, https://doi.org/10.1007/978-3-030-30796-7_20.

Temmermans, F. et al. (2023). Towards an international standard to establish trust in media production, distribution and consumption. In *Proceedings of the 24th International Conference on Digital Signal Processing*, https://doi.org/10.1109/DSP58604.2023.10167884.

Teyssou, D. (2019). *Video Verification in the Fake News Era*, Springer, chap.

Applying Design Thinking Methodology: The InVID Verification Plugin. pp. 263–279, https://doi.org/10.1007/978-3-030-26752-0_9.

Thajeel, S. A. and Sulong, G. (2013). State of the art of copy-move forgery detection techniques: A review. *International Journal of Computer Science Issues*, 10(6), pp. 174–183, https://www.ijcsi.org/papers/IJCSI-10-6-2-174-183.pdf.

Thorne, J. and Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, ACL, pp. 3346–3359, https://aclanthology.org/C18-1283.

Time (2021). Incited by the president, pro-Trump rioters violently storm the Capitol. https://time.com/5926883/trump-supporters-storm-capitol/.

Tkaczyk, D. et al. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4), pp. 317–335, https://doi.org/10.1007/s10032-015-0249-8.

Tolosana, R. et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, pp. 131–148, https://doi.org/10.1016/j.inffus.2020.06.014.

Triphati, G., Ahad, M. A. and Haq, Z. A. (2020). Multimedia tampering detection: A comprehensive review of available techniques and solutions. In *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals*, Springer, pp. 223–235, https://doi.org/10.1007/978-981-15-0339-9_18.

Trompeter, G. M. et al. (2012). A synthesis of fraud-related research. *AUDITING: A Journal of Practice & Theory*, 32(Supplement 1), pp. 287–321, http://doi.org/10.2308/ajpt-50360.

Tucker, J. et al. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. Tech. rep., The William and Flora Hewlett Foundation, https://hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.

Turnnidge, S. (2022). BBC breakfast uses old footage of Russian parade rehearsal to show invasion of Ukraine. https://fullfact.org/europe/bbc-footage-russian-flyover-ukraine/.

Ushio, A., Liberatore, F. and Camacho-Collados, J. (2021). Back to the basics: A quantitative analysis of statistical and graph-based term weighting schemes for keyword extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 8089–8103, https://doi.org/10.18653/v1/2021.emnlp-main.638.

van der Meer, T. G. L. A., Hameleers, M. and Ohme, J. (2023). Can fighting misinformation have a negative spillover effect? How warnings for the threat of misinformation can decrease general news credibility. *Journalism Studies*, pp. 1–21, https://doi.org/10.1080/1461670X.2023.2187652.

Vartapetiance, A. and Gillam, L. (2014). Deception detection: Dependable or defective? *Social Network Analysis and Mining*, 4(1), pp. 166:1–166:14, https://doi.org/10.1007/s13278-014-0166-8.

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), pp. 910–932, 10.1109/JSTSP.2020.3002101.

Vivian, A. C. and Hussain, S. M. (2014). First party fraud: A review of the forms and motives of fraudulent consumer behaviours in e-tailing. *International*

Journal of Retail & Distribution Management, 42(9), pp. 805–817, https://doi.org/10.1108/IJRDM-05-2013-0112.

Viviani, M. and Pasi, G. (2017). Credibility in social media: Opinions, news, and health information – A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5), pp. e1209:1–e1209:25, https://doi.org/10.1002/widm.1209.

Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, ACL, pp. 18–22, https://doi.org/10.3115/v1/W14-2508.

Vo, N. and Lee, K. (2018). The rise of guardians: Fact-checking URL recommendation to combat fake news. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, pp. 275–284, https://doi.org/10.1145/3209978.3210037.

Walter, N. et al. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3), pp. 350–375, https://doi.org/10.1080/10584609.2019.1668894.

Wan, X. and Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI, pp. 855–860, https://www.aaai.org/Papers/AAAI/2008/AAAI08-136.pdf.

Wang, J. and Yu, B. (2021). News2PubMed: A browser extension for linking health news to medical literature. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 2605–2609, https://doi.org/10.1145/3404835.3462788.

Wang, S. (2010). A comprehensive survey of data mining-based accounting-fraud detection research. In *Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation*, vol. 1, pp. 50–53, https://doi.org/10.1109/ICICTA.2010.831.

Wang, W., Dong, J. and Tan, T. (2009). A survey of passive image tampering detection. In *Digital Watermarking: 8th International Workshop, IWDW 2009, Guildford, UK, August 24-26, 2009, Proceedings*, Springer, pp. 308–322, http://doi.org/10.1007/978-3-642-03688-0_27.

Wang, X., Luo, X. and Liu, H. (2015). Measuring the veracity of web event via uncertainty. *Journal of Systems and Software*, 102, pp. 226–236, https://doi.org/10.1016/j.jss.2014.07.023.

Wang, Y. et al. (2023). Factcheck-GPT: End-to-end fine-grained document-level fact-checking and correction of LLM output. https://doi.org/10.48550/arXiv.2311.09000.

Warbhe, A. D., Dharaskar, R. V. and Thakare, V. M. (2016). A Survey on Keypoint Based Copy-paste Forgery Detection Techniques. *Procedia Computer Science*, 78, pp. 61–67, the First International Conference on Information Security & Privacy, http://doi.org/10.1016/j.procs.2016.02.011.

Wardle, C. (2017). Fake news. It's complicated. https://firstdraftnews.org/articles/fake-news-complicated/.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Tech. rep., Council of Europe Strasbourg, https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf.

Webb, H. et al. (2016). Digital wildfires: Propagation, verification, regulation,

and responsible innovation. *ACM Transactions on Information Systems*, 34(3), pp. 15:1–15:23, https://doi.org/10.1145/2893478.

Weld, G., Glenski, M. and Althoff, T. (2021). Political bias and factualness in news sharing across more than 100,000 online communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1), pp. 796–807, https://doi.org/10.1609/icwsm.v15i1.18104.

Westlund, O. et al. (2022). *Disinformations Studies: Perspectives from An Emerging Field*, LABCOM Communication & Arts, chap. Technologies and Fact-Checking: A Sociotechnical Mapping. pp. 193–237, https://labcomca.ubi.pt/wp-content/uploads/2022/11/2022_DisinformationStudies_JCorreiaPJeronimoIAmaral.pdf.

World Economic Forum, in partnership with Marsh & McLennan Companies and Zurich Insurance Group (2024). The global risks report 2024 19th edition. Tech. rep., World Economic Forum, https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf.

Xin, T., Bowei, Z. and Ai Ti, A. (2023). Evidence-based interpretable open-domain fact-checking with large language models. https://doi.org/10.48550/arXiv.2312.05834.

Yang, K.-C., Ferrara, E. and Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science*, 5(2), pp. 1511–1528, https://doi.org/10.1007/s42001-022-00177-5.

Yang, K.-C. and Menczer, F. (2023). Large language models can rate news outlet credibility. https://doi.org/10.48550/arXiv.2304.00228.

Yin, J. et al. (2023). Emulating reader behaviors for fake news detection. https://doi.org/10.48550/arXiv.2306.15231.

Yousuf, M. (2023). Mediating the truth: Influences of routines on legacy news media fact-checking. *Journalism Practice*, in press, https://doi.org/10.1080/17512786.2023.2169187.

Yu, Y. and Ng, V. (2018). WikiRank: Improving unsupervised keyphrase extraction using background knowledge. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), pp. 3723–3727, https://aclanthology.org/L18-1588.

Yuan, H. et al. (2022). Graphical models of false information and fact checking ecosystems. https://doi.org/10.48550/arXiv.2208.11582.

Zakariah, M., Khan, M. K. and Malik, H. (2018). Digital multimedia audio forensics: Past, present and future. *Multimedia Tools and Applications*, 77(1), pp. 1009–1040, https://doi.org/10.1007/s11042-016-4277-2.

Zannettou, S. et al. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, 11(3), https://doi.org/10.1145/3309699.

Zeng, X., Abumansour, A. S. and Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10), pp. e12438:1–e12438:21, https://doi.org/10.1111/lnc3.12438.

Zervanou, K. (2010). UvT: The UvT term extraction system in the keyphrase extraction task. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, ACM, pp. 194–197, https://dl.acm.org/doi/abs/10.5555/1859664.1859706.

Zhang, L. et al. (2022). MDERank: A masked document embedding rank approach for unsupervised keyphrase extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, ACL, pp. 396–409, https://doi.org/10.18653/v1/2022.findings-acl.34.

Zhang, W. et al. (2023). NewsQuote: A dataset built on quote extraction and attribution for expert recommendation in fact-checking. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, AAAI, pp. 22:1–22:11, https://doi.org/10.36190/2023.22.

Zhang, X. and Gao, W. (2023). Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, ACL, pp. 996–1011, https://doi.org/10.18653/v1/2023.ijcnlp-main.64.

Zhang, X. and Ghurbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management*, 57(2), pp. 102025:1–102025:26, http://doi.org/10.1016/j.ipm.2019.03.004.

Zhang, Z., Rudra, K. and Anand, A. (2021). FaxPlainAC: A fact-checking tool based on EXPLAINable models with HumAn correction in the loop. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ACM, pp. 4823–4827, https://doi.org/10.1145/3459637.3481985.

Zhao, A. and Hu, L. (2023). Platform, jurors, and reporters: How different actors facilitate community-driven content moderation on weibo. https://doi.org/10.31235/osf.io/da8fz.

Zheng, L., Zhang, Y. and Thing, V. L. L. (2019). A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, 58, pp. 380–399, http://doi.org/10.1016/j.jvcir.2018.12.022.

Zhong, L. et al. (2023). A comprehensive survey on automatic knowledge graph construction. https://doi.org/10.48550/arXiv.2302.05019.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), https://doi.org/10.1145/3395046.

Zubiaga, A. et al. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2), pp. 32:1–32:36, https://doi.org/10.1145/3161603.

Škrlj, B., Repar, A. and Pollak, S. (2019). RaKUn: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. In *Proceedings of the 7th International Conference on Statistical Language and Speech Processing*, *Lecture Notes in Computer Science*, vol. 11816, Springer, pp. 311–323, https://doi.org/10.1007/978-3-030-31372-2_26.

Appendix A

Survey Questionnaire

A.1 Participant Information Sheet

This survey is part of an ongoing PhD project in the School of Computing at the University of Kent in the UK. The main aim of this survey is to learn about people's fact-checking attitudes and behaviours.

The researchers participating in this study include the following:

`Enes Altuncu, Shujun Li, Jason Nurse, Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, UK`

Data from this survey will be accessible to the participating researchers.

In this survey, we will ask questions about the following aspects:

- Information and information providers that require fact-checking
- Reliable information providers for fact-checking and fact-checkers
- Familiarity with and usage patterns of fact-checking tools and services
- Trust in fact-checking tools with a level of automation
- Explainability in fact-checking tools

Your participation in this survey is completely voluntary; choosing not to take part will not disadvantage you in any way. If you do decide to take part, you are still free to withdraw at any time and without giving a reason. You will receive full payment only upon completing the survey fully and after we have verified that your response is valid. The survey is anonymous and we will not collect any personal data about you. Instructions are provided during the survey. Please read the instructions carefully and answer the questions accordingly. Remember that there are no wrong or right answers. We are interested in your honest opinions. We will consider all reasonable answers as valid, however, if some of your answers contain too many obvious contradictions or meaningless responses, or if you fail the attention checks (questions in the survey which test that you are paying attention), they may be considered invalid and you will not receive any payment.

What happens to the information you provide?

All data will remain anonymous. The data collected will be stored on a secure university server (which hosts this survey itself) and no one other than the researchers mentioned above will be given access to the data stored on the server.

If you require any further information, wish to withdraw your data, or have any queries about this study please contact the lead researcher Enes Altuncu via his email address ea483@kent.ac.uk.

If you would like any further information, or in case of a complaint, please contact the School of Computing's Ethics Officer via his email or postal address.

A.2 Consent Form

- I confirm that I have read and understood the information sheet for the survey. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.

- I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason.
- I understand that the survey will not collect any personal data that will allow identification of me.
- My anonymised responses may be shared with other researchers or made available in online data repositories.
- I will provide my answers honestly and responsibly.
- I understand that I will receive payment only if my responses are considered valid by the research team.
- I understand that my responses will be used in research publications by the research team and that the data may be published to support reproducible research.
- I am over the age of 18.

A.3 General Questions

1. Please enter your Prolific ID to start this survey. (Short free text)
2. Please select your age. [Under 18; 18-29; 30-49; 50-64; Over 65; Prefer not to say]
3. Please select your gender. [Male; Female; Other; Prefer not to say]
4. Please select the highest level of education you have completed. [Less than high school or equivalent; High school or equivalent; Bachelor's degree or equivalent; Master's degree or equivalent; Doctoral degree or equivalent; Other (Please specify)]

5. Please select your native language from the list. (Drop-down list of languages)

6. How would you rate your computer knowledge? [Basic; Intermediate; Advanced; Professional; None]

7. About how much time do you spend online every day? [Less than 1 hour; 1-3 hours; 3-5 hours; More than 5 hours]

A.4 Information and Information Providers That Require Fact-Checking

1. Which of the following platform(s) do you use to get information? (Select all that apply) [TV/Radio; Newspapers; Books/Magazines; Websites; Smartphones; Computers or laptops; Tablets; Smartwatches (e.g., Apple Watch, Samsung Galaxy); Smart speakers (e.g., Amazon Alexa, Google Home); Other (Please specify)]

2. How frequently do you use a digital device (such as a smartphone, computer, laptop, or tablet) to get information? [More than once a day; Once a day; Once or twice a week; Less than once a week; Never; Don't know]

3. Which type of device do you prefer to get information? (Select all that apply) [A desktop or laptop computer; A mobile device, such as a smartphone or tablet; Other (Please specify)]

4. Where do you get your information online in your daily life? (Select all that apply) [Mainstream news outlets (e.g., BBC, Guardian, Reuters); News Aggregators (e.g., Apple News, Flipboard, Pocket, Google News); Local news

outlets; Independent journalists who do not work for a news outlet; Informative social media accounts (e.g., Hourly Facts, Word of the Day); Scientific publications for experts (e.g., academic articles, technical reports); Science/Technology magazines (e.g., Wired, Popular Science); Online encyclopaedias (e.g., Wikipedia, Encyclopaedia Britannica, Scholarpedia, RationalWiki); Online forums (e.g., Reddit, 4Chan, StackOverflow, Quora); Blogging websites (e.g., Medium, Substack); Video streaming platforms (e.g., YouTube); Audio/Podcast streaming platforms (e.g., Spotify, Apple Podcasts); Periodic email newsletters; Other (Please specify)]

5. When you would like to learn about something, what do you use to search for information? (Select all that apply) [Search engines (e.g., Google, Bing); Social media platforms (e.g., Facebook, Twitter, Instagram, TikTok); Video streaming platforms (e.g., YouTube); Audio/Podcast streaming platforms (e.g., Spotify, Apple Podcasts); Research databases (e.g., Google Scholar, Scopus); Government websites; Academic institution websites; Online encyclopaedias (e.g., Wikipedia, Scholarpedia, RationalWiki); Online forums (e.g., Reddit, 4Chan, StackOverflow, Quora); Other (Please specify)]

6. How frequently do you encounter false or inaccurate information online? [More than once a day; Once a day; Once or twice a week; Less than once a week; Never; Don't Know]

7. Do you agree that false or inaccurate information is a serious threat to society? [Strongly agree; Agree; Neither agree nor disagree; Disagree; Strongly disagree]

8. How often do you feel the need for fact-checking suspicious claims? [More than once a day; Once a day; Once or twice a week; Less than once a week; Never; Don't know]

9. For each of the following organisations, please select the option most accurately reflects your agreement/disagreement on that you feel them credible when it comes to providing true and accurate information. (5-point Likert from Strongly Disagree to Strongly Agree) [Mainstream news outlets; Local news outlets; Non-governmental organisations (NGOs); Academic institutions; Governmental organisations; Social media platforms]

10. For each of the following individuals, please select the option most accurately reflects your agreement/disagreement on that you feel them credible when they give information about something you haven't heard of. (5-point Likert from Strongly Disagree to Strongly Agree) [Journalists who work for a mainstream news outlet; Journalists who work for a local news outlet; Independent journalists who don't work for any news outlet; Celebrities/Influencers; Activists; Family, friends, and acquaintances; Experts (e.g., scientists, engineers, doctors); Politicians; Social media users that you don't know personally but come across their posts]

A.5 Reliable Information Providers for Fact-Checking and Fact-Checkers

1. How do you know if what you're reading online is accurate and reliable? (Select all that apply) [I check the URL to see if it looks legitimate; I check if the source of what I'm reading is credible; I check for signs showing the website is not a reputable source (e.g., bad web design, text in all caps, excessive punctuation usage); I ask someone whom I trust for their opinion about the accuracy and reliability of what I'm reading; I search for relevant information online via search engines (e.g., Google, Bing) and/or research databases (e.g. Google Scholar, Scopus); I check the websites of

fact-checking organisations to see if they investigated the claim previously; I use fact-checking software to check the accuracy and reliability of what I'm reading; I cross-check from other sources; Other (Please specify)]

2. For each of the following organisations, please select the option most accurately reflects your agreement/disagreement on that you feel them credible when it comes to providing true and accurate information for fact-checking purposes. (5-point Likert from Strongly Disagree to Strongly Agree) [Mainstream news outlets; Local news outlets; News associations; Regulators; Anonymous social media accounts which share posts about their recent fact-checks; Fact-checking organisations (e.g., Full Fact, Snopes); Non-governmental organisations (NGOs); Academic institutions; Governmental organisations; Tech companies, including social media platforms]

3. For each of the following individuals, please select the option most accurately reflects your agreement/disagreement on that you feel them credible when they give information for fact-checking purposes. (5-point Likert from Strongly Disagree to Strongly Agree) [Journalists who work for a mainstream news outlet; Journalists who work for a local news outlet; Independent journalists who don't work for any news outlet; Fact-checkers who work for a fact-checking organisation (e.g., Full Fact, Snopes); Activists; Celebrities/Influencers; Family, friends, and acquaintances; Experts (e.g., scientists, engineers, doctors); Politicians; Crowdsourcing participants; Social media users that you don't know personally but come across their posts]

A.6 Familiarity with and Usage Patterns of Fact-Checking Organisations and Tools

1. "Fact-checking organisations are organisations dedicated to investigating suspicious claims to verify or debunk." Have you ever heard of any such organisation that you think might answer the given definition? [Yes, No, Not sure][1]

2. "Fact-checking tools involve any software, artefact, or resource that can be used to investigate the accuracy of suspicious claims or to seek findings of previous investigations of suspicious claims." Have you ever heard of any such tool that you think might answer the given definition? [Yes, No, Not sure][2]

3. For each of the following types of tools, please select the option most accurately describes how often you use them when investigating suspicious claims or following recent fact-checks? (5-point Likert from Never to Every Day) [Websites of fact-checking organisations (e.g., BBC Reality Check, Full Fact, Snopes); Social media accounts of fact-checking organisations (e.g., BBC Reality Check, Full Fact, Snopes); Mobile applications (e.g., Logically App, FactStream); Messaging apps (e.g., Whatsapp chatbots); Web applications (e.g., ClaimBuster, Google Fact Check Explorer); Web browser extensions (e.g., InVID/WeVerify, The Factual, NewsGuard, MediaBias/Fact Check); Content moderation tools (e.g., Twitter Community Notes/Birdwatch, Meta CrowdTangle); Content analysis tools (e.g., TinEye, Google Reverse Image Search, Deepware, Google Street View); Monitoring Tools (e.g., TweetDeck, Trendolizer, Botometer); Search engines (e.g., Google Search, Bing, Yandex); Scientific databases (e.g., Google Scholar, Scopus); Online written

[1] If "Yes" is selected, the participant is asked to name three of them
[2] If "Yes" is selected, the participant is asked to name three of them

materials (e.g., newsletters, handbooks, guides); Online multimedia materials (e.g., tutorials, video series, visual aids)]

A.7 Trust in Fact-Checking Tools with a Level of Automation

1. Please rank the following fact-checking approaches in terms of your trust (most at the top). (Drag-and-drop ranking) [AI automatically does fact-checking; AI and humans work together to do fact-checking; Humans do fact-checking without any automation]

2. For each of the following fact-checking tool designs, please select the option most accurately reflects your agreement/disagreement on that you feel them credible. (5-point Likert from Strongly Disagree to Strongly Agree) [AI automatically predicts if it's true or false; AI automatically predicts if it's true or false by considering human input in different stages of the fact-checking process; AI automatically finds evidence. Then, based on the evidence, a human fact-checker decides if it's true or false; AI automatically finds evidence. Then, based on the evidence, other users of the tool decide if it's true or false by voting; AI automatically finds evidence. Then, based on the evidence, you will decide on your own if it's true or false; If many users of the tool mark the claim as false, it is considered to be false; A human fact-checker investigates if the claim is true with open and closed sources]

3. For each of the following fact-checking approaches, please select the option most accurately reflects your agreement/disagreement on that you feel them credible when humans and AI work together to do fact-checking. (5-point Likert from Strongly Disagree to Strongly Agree) [Humans use AI as an

assistant to accomplish certain fact-checking tasks (e.g., detecting check-worthy claims, and finding evidence) faster; AI uses humans' opinions (e.g., comments under a social media post) to predict the veracity of claims; AI and humans work in hybrid to assess the veracity of claims – AI requires human input at different stages of fact-checking; AI predicts the veracity of claims, and humans check the AI output for refinement and to avoid any potential mistakes]

4. For each of the following tasks, please select the option most accurately reflects your agreement/disagreement on that you think it is important to be automated. (5-point Likert from Strongly Disagree to Strongly Agree) [Identifying suspicious claims that need to be fact-checked; Prioritising the most important claims that need fact-checking; Monitoring potentially false or misleading information online; Identifying sources of suspicious claims; Verifying sources of suspicious claims; Verifying embedded multimedia (e.g., image, video, audio) to check if they are manipulated; Retrieving relevant and trustworthy evidence; Determining if a claim is true or false; Providing a convincing explanation for the determined veracity of a claim]

A.8 Explainability in Fact-Checking Tools

1. For each of the following statements about fact-checking tools (FCT), please select the option which most accurately reflects your agreement/disagreement. (5-point Likert from Strongly Disagree to Strongly Agree) [I can trust a fully-automated FCT only if it provides explanations on how it predicts the veracity of claims; I never trust a fully-automated FCT regardless of how it is designed; It is important for me that a fully-automated FCT shows which sentences/words in the input text have more contribution to the veracity prediction; Please select "Disagree" for this statement **(Attention check)**;

Even if humans make the final decision about the veracity of claims, an FCT needs to clearly explain how its automated parts work; I never trust the output of an FCT which contains AI; It is important for me that an FCT only considers credible sources when finding evidence to verify/debunk a claim; An FCT must clearly show users what sources it collects the evidence from to verify/debunk a claim; It is sufficient for an FCT to show only a veracity label (e.g., True, False) to users without providing any explanation]