University of Sheffield

Detecting and Tracking the Spread of Debunked Narratives Across Languages

Iknoor Singh

Supervisor: Carolina Scarton and Kalina Bontcheva

A thesis submitted for the degree of Doctor of Philosophy in Computer Science

in the

Department of Computer Science

July 22, 2024

Declaration

I, Iknoor Singh, hereby declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgement, the work presented is entirely my own.

Name: Iknoor Singh

Date: July 22, 2024

Acknowledgements

Embarking on the PhD journey has been a challenging yet enriching experience, and I am indebted to several people who have played pivotal roles in all these doctoral research years.

First and foremost, my deepest gratitude goes to my supervisors, Dr. Carolina Scarton, and Professor Kalina Bontcheva. Their unwavering guidance and unconditional support have been crucial, especially in navigating the complexities of initiating my doctoral research amid the unprecedented challenges posed by the COVID-19 pandemic. The depth of my appreciation for their mentorship goes beyond words, and I am truly fortunate to have had them as pillars of support from day one.

I also extend my gratitude to my PhD panel committee members, Professor Heidi Christensen, Professor Nikolaos Aletras, and Dr. Loic Barrault for their constructive feedback and guidance during my research. I also thank my PhD examiners, Dr. Diana Maynard and Dr. Harith Alani, for examining my thesis and providing valuable insights.

Furthermore, I extend heartfelt thanks to my research collaborators and friends — Kokil Jaidka, Xingyi Song, Muneerah Patel, Diana Maynard, Ahmad Zareie, Ibrahim Abu Farha, Ian Roberts, Mark Greenwood, and the numerous other individuals in the GATENLP group. Their generosity in sharing knowledge has been instrumental in honing the skills required to address the diverse challenges encountered throughout my research journey. The collaborative and nurturing environment fostered by the GATENLP group has significantly contributed to my academic and professional growth.

My sincere appreciation also goes to my colleagues in the department — Amit Meghanani, Dr. Danae Sánchez, Mugdha Pandya, Jason Clarke, Dr. Yida Mu, Miles Williams, Rob Flynn, Yue Li, Jasivan Sivakumar, and others whose friendship and shared experiences have added a layer of warmth and support to this academic endeavour. I am grateful for the laughter, encouragement, and mutual support that we have shared throughout my Ph.D. journey.

In addition, I also want to express gratitude to my family, with a special mention of my mother, Amarpeet Kaur, father, Jaswinder Pal Singh, and brother, Divij Singh. I extend heartfelt thanks for their unwavering love and support, which has been instrumental throughout my PhD journey.

Finally, I would like to express gratitude to my Ph.D. funders – European Union Horizon 2020 Program, grant no. 825297 (WeVerify), the European Union Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", UKRI grant EP/W011212/1, a European Union Horizon 2020 grant (agreement no. 871042) ("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)) and a European Union grant INEA/CEF/ICT/A2020/2381686 (European Digital Media Observatory (EDMO) Ireland, https://edmohub.ie/) – for the PhD scholarship and the generous financial support, which has played a crucial role in facilitating my research.

Abstract

Misinformation and disinformation during critical events, like the COVID-19 pandemic and geopolitical conflicts such as the Ukraine war, poses threats to public perception, social cohesion, and political stability. While fact-checkers strive to counter their spread, a multifaceted problem emerges: the enduring and widespread propagation of similar or nearly duplicate false narratives across multiple languages, modalities, and social media platforms, often persisting long after the initial debunking by a professional fact-checker.

First, this thesis utilises the CoronaVirusFacts Alliance database to identify and uncover repeatedly debunked false narratives related to COVID-19. The spatiotemporal analysis indicates the global prevalence of false narratives related to general medical advice, consistently shared by Facebook users despite the existence of fact-checks that have already debunked similar narratives across different languages. Additionally, the thesis analyses debunks related to the Ukraine conflict, revealing the wider spread of disinformation compared to its debunks and demonstrating the delayed but positive impact of debunks on reducing Ukraine-related disinformation. The thesis ultimately advocates for the implementation of a cross-lingual debunked narrative search tool in the fact-checking pipeline to efficiently identify previously debunked narratives in different languages.

Motivated by the challenges posed by the persistent spread of debunked narratives, this thesis delves into cross-lingual debunked narrative retrieval, aiming to enhance the performance and robustness of retrieval models across various languages. Firstly, it introduces the Multistage BiCross encoder for multilingual access to COVID-19 information, presenting experimental results and search query optimisation techniques. Subsequently, the thesis introduces novel benchmark datasets and computational methods to aid fact-checkers in detecting debunked narratives across multiple languages. It also emphasises the need for social media platforms to adopt similar technologies at scale to optimise fact-checker resources. Finally, the thesis proposes unsupervised methods for training debunked narrative retrieval models, offering effective real-time adaptation without relying on time-consuming and labour-intensive human annotations.

In summary, the research contributes to a comprehensive understanding of the spread of debunked narratives. It offers practical solutions and insights that can inform policy decisions and contribute to the ongoing global efforts against misinformation and disinformation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

False information poses significant threats to public perception, social cohesion, and political stability, especially during critical events such as the COVID-19 pandemic (Posetti et al., 2020; IFCN, 2024) and geopolitical conflicts like the Russia-Ukraine war (Babacan and Tam, 2022; Mejias and Vokuev, 2017; EUvsDisinfo, 2024) and the Israel-Hamas war (Stănescu, 2024; FullFact, 2024a). For instance, people around the world encountered many false narratives about the coronavirus's origin, spread, medical treatments, and vaccines (Posetti et al., 2020; Brennen et al., 2020). Beyond mere inaccuracy, the spread of false information can have profound real-world consequences and has the potential to cause considerable harm (Lewandowsky et al., 2012; Roozenbeek et al., 2020; Nakov and Da San Martino, 2021; Arora et al., 2023). For example, around 800 people lost their lives because of false information related to coronavirus in just the first three months of 2020 (BBC, 2024). On the other hand, false information triggered a massive exodus of migrants during the lockdown, causing a national disturbance in India (TOI, 2024).

Another issue with false information is that it often originates on various social media platforms which makes its authenticity questionable as there is no method in place to quickly check the credibility of the content as well as the source (Limaye et al., 2020; Tasnim et al., 2020; Del Vicario et al., 2016; Singh et al., 2022). Additionally, easy access to social media often provides the playground for bad actors to execute their nefarious motives (Vosoughi et al., 2018; Guess et al., 2020; Di Domenico et al., 2021; Lazer et al., 2018). For example, false narratives were deliberately spread on social media platforms during the 58th US presidential elections in order to influence the election outcome (Watts, 2017; Allcott and Gentzkow, 2017). Hence, social media websites have become key conduits as they not only provide a medium for publishing factually inaccurate information but also offer services like promoting information to target specific people or community[1] (Lazer et al., 2018; Spenkuch and Toniatti, 2016; Flaxman et al., 2016). Moreover, a report by Pew Research (Shearer, 2021) shows that 52% of American adults get news from digital platforms, out of which more than half of the respondents (53%) said that they consume news from social media platforms. Therefore, detecting the spread of false information on social media platforms has become both important and urgent.

Despite the increasing efforts by fact-checking initiatives to combat false information (Graves and Cherubini, 2016; Haque et al., 2018), there exists a critical research gap in detecting and analysing the cross-border propagation of similar debunked narratives and the duplicated efforts of fact-checkers in debunking them. This thesis highlights the persistence of debunked narratives in various regions and languages, emphasising the need for empirical investigations into their spatio-temporal aspects to develop targeted strategies for countering false information on a global scale. Furthermore, it introduces novel methods and datasets for cross-lingual debunked narrative retrieval, aiming to enhance the performance and robustness of retrieval models across various languages.

The next section (Section 1.1) clarifies the terminology associated with the spread of false information, providing essential context for understanding this research. Section 1.2 discusses the problem statement and research motivation in detail. Additionally, Section 1.3 outlines the research questions addressed in this investigation. Finally, Section 1.4 gives a brief overview of each chapter and the corresponding contributions. It also enumerates the publications generated during the research for this study.

1.1 Definition and Terminology

In the context of this thesis, it is essential to clarify certain definitions and terminology associated with the spread of false information. The term "fake news" has frequently been employed in the literature to denote false information (Pennycook and Rand, 2021; Lazer et al., 2018; Gelfert, 2018). However, this study refrains from using the term "fake news" due to its

[1] https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

ambiguity and varying definitions across researchers and organisations.

For instance, Klein and Wueller (2018) defines fake news as an "online publication of intentionally or knowingly false statements of fact". The Ethical Journalism Network (EJN)[2] defines it as "information deliberately fabricated and published with the intention to deceive and mislead others into believing falsehoods or doubting verifiable facts." Notably, renowned dictionaries such as Cambridge[3], Collins[4], and Oxford[5], among others, also provide varying definitions. Given these disparate definitions, this thesis opts for clarity by not employing the term "fake news". Instead, it employs a broader concept of false information which may include false claims or narratives, defined as any content containing false information that is not true. The distinction between a "claim"' and a "narrative" is as follows:

- **Claim**: A claim typically refers to a specific statement or assertion, often presented without accompanying evidence or proof[6]. Claims can be either true or false and frequently serve as building blocks within a broader narrative.
- **Narrative**: A narrative aims to convey a particular message, viewpoint, or perspective, often by structuring various claims or pieces of information into a cohesive storyline[7]. For example, different false narratives related to 5G and coronavirus emerged and spread widely [8].

Additionally, false claims and narratives require further nuance due to the varying purposes behind their spread. As explored by Wardle and Derakhshan (2017), misinformation, disinformation, and malinformation each represent distinct intents behind the spread of false information. Despite these distinctions, this study addresses general false information regardless of specific subtype. The details of each category are as follows:

- **Misinformation**: Information that is false but the person sharing the

[2] https://ethicaljournalismnetwork.org/fake-news-bad-journalism-digital-age

[3] https://dictionary.cambridge.org/dictionary/english/fake-news

[4] https://www.collinsdictionary.com/dictionary/english/fake-news

[5] https://www.oxfordlearnersdictionaries.com/definition/english/fake-news?q=fake+news

[6] https://dictionary.cambridge.org/dictionary/english/claim

[7] https://www.merriam-webster.com/dictionary/narrative

[8] https://fullfact.org/online/5g-and-coronavirus-conspiracy-theories-came/

information has no intent to harm anybody. For example, people shared home remedies to cure COVID-19 in a genuine effort to help others[9].

- **Disinformation**: Information that is false and is shared intentionally by a person or groups of people with an intent to deceive or cause harm to a particular person, community, organisation or country. For example, anti-vaxxer disinformation spread amidst the COVID-19 pandemic[10] amidst the COVID-19 pandemic.
- **Malinformation**: Malinformation also has an intent to deceive or harm but the information itself is true and based on reality. For instance, decontextualised images of empty grocery stores during the start of the COVID-19 pandemic were used to instil panic buying[11] (Brennen et al., 2020).

Furthermore, this thesis introduces the concept of "debunked narrative retrieval" as a central theme. This concept involves the process of identifying and retrieving narratives or claims that have been previously debunked by fact-checking organisations. This retrieval aims to locate instances where false information reappears after it has already been debunked. Finally, the term "debunked narratives" specifically refers to false information that has already been refuted and proven inaccurate. Throughout the thesis, the terms "debunks" and "fact-checks" are used interchangeably.

1.2 Research Motivation

To counter the spread of false information, there has been a significant surge in fact-checking initiatives, dedicated to monitoring and debunking (IFCN, 2024; Graves and Cherubini, 2016; Pavleska et al., 2018; Haque et al., 2018). However, despite these efforts, the immediate and enduring damages caused by false information persist, highlighting the ongoing challenge (Burel et al., 2020; Schuetz et al., 2021; Barrera et al., 2020; Anderson and Rainie, 2017).

The amplification of false information is further exacerbated by social me-

[9] https://www.bbc.co.uk/news/world-51735367

[10] https://www.bloomberg.com/news/newsletters/2021-07-02/anti-vaccine-disinformation-spreads-in-asia

[11] https://www.politifact.com/factchecks/2020/mar/06/facebook-posts/romanian-conspiracy-theory-migrates-us-amid-corona/

dia, resulting in the dissemination of similar false narratives across different countries at varying times (FullFact, 2024b). These narratives are often debunked by multiple fact-checking organisations in multiple languages. This proliferation has led to duplicated debunking efforts, leading to a significant waste of fact-checker resources. For instance, the claims stating that consuming alkaline-rich foods can eliminate coronavirus was initially debunked in Europe (Maldita, 2020) in Spanish. However, this claim persisted, as it faced debunking once more by fact-checking organisations in Asian (Boomlive, 2020; Teyit, 2020), South American (Fatos, 2020; Cocuyo, 2020), and North American (Leadstories, 2020; AP, 2024) countries in multiple languages. Notably, there is an absence of extensive empirical investigations into the cross-border propagation of similar or nearly duplicate debunked narratives across various modalities and languages. Furthermore, even if multiple fact-checkers consistently debunk information online, the current effectiveness of debunking in curbing the overall spread of false narratives on social media platforms remains an unanswered question, especially in the context of the Russia-Ukraine war. Addressing this problem is essential for informing policy decisions, guiding resource allocation, and providing valuable insights into the dynamics of false information on social media platforms.

Simultaneously, in the pursuit of mitigating the spread of false information, researchers have proposed automated fact-checking systems as a complementary approach (Panchendrarajan and Zubiaga, 2024; Shu and Liu, 2022; Nielsen and McConville, 2022). These systems serve the dual purpose of countering false narratives on digital media and easing the workload on fact-checkers (Shang et al., 2023; Wu et al., 2022; Guo et al., 2022; Zhang et al., 2022a; Zeng et al., 2021). A key task of these systems is the detection of previously fact-checked similar claims which aims to detect claims that spread even after they have already been debunked by at least one professional fact-checker (Nakov et al., 2022c, 2021b; Shaar et al., 2020b). This is essentially a retrieval problem, referred to as "debunked narrative retrieval" in this thesis, where a claim serves as the query to extract relevant debunked narratives from a database of already published publicly available fact-checking articles. Previous work has mostly focused on training retrieval models, primarily focusing on monolingual retrieval, where the language of the query claim matches the language of the debunk (Shaar et al., 2020a,b; Nakov et al., 2021b, 2022b; Barrón-Cedeño et al., 2023). However, it is imperative to emphasise that these monolingual retrieval models operate under the assumption that debunked information is exclusively contained within a single language. This assumption is challenged by compelling evidence presented in this thesis, demonstrating the persistent spread of similar false claims across multiple languages, despite the availability of debunks in another language for several months. Thus, the automated detection of debunked narratives in multiple languages is crucial to make the best use of scarce fact-checkers' resources.

Building upon this need, the thesis delves into the realm of "cross-lingual debunked narrative retrieval," which aims to find and retrieve debunked narratives in different languages. To address this, the primary issue at hand revolves around the need to enhance the performance of cross-lingual and multilingual retrieval models using advanced neural techniques while ensuring their robustness. Moreover, significant challenges are associated with cross-lingual retrieval of debunked narratives, including the potential for retrieval models trained on high-resource languages to benefit low-resource languages, as well as the cross-dataset and cross-domain generalisation of models in a zero-shot setting. By addressing these challenges, the thesis seeks to contribute to the global fight against misinformation and disinformation.

As the narrative unfolds, it becomes evident that prompt detection of false narratives is inherently challenging due to its topical nature. This highlights the necessity for models to adapt and learn from new data in real-time to effectively counter ever-evolving false narratives. While creating human-annotated datasets is a solution to regularly update retrieval models on emerging topics, it introduces challenges such as high costs and time constraints. To overcome this, the thesis explores alternative approaches, specifically unsupervised methods, to mitigate the resource-intensive nature of current practices. It also aims to determine various factors that influence the effectiveness of unsupervised methods. In conclusion, the research paves the way for more efficient and cost-effective methods for the development of debunked narrative retrieval models.

1.3 Research Questions

This doctoral research seeks to address two primary research questions (1, 2). It breaks them down into sub-research questions, as outlined below.

1. ***To what extent do false narratives propagate even after being debunked by professional fact-checkers, and how does their spread com-***

pare to the dissemination of corresponding debunks?

(a) *How do false narratives propagate across languages, modalities, and social media platforms, even after they have been debunked by at least one professional fact-checking organisation?*

To answer this, this thesis investigates the spatiotemporal characteristics of false narratives related to COVID-19 that have multiple debunks and see how these differ in terms of country, social media platforms and modality of content. Specifically, the investigation explores how long false narratives persist and continue to circulate after their initial debunking by a fact-checker. The findings are presented in Chapter 2. Additionally, this inquiry is also extended to Ukraine-related false narratives in Chapter 3. Finally, this research question sets the stage for the second primary research question (2) by highlighting the importance of cross-lingual debunked narrative retrieval in the claim verification workflow to detect the spread of debunked narratives in multiple languages.

(b) *How does the spread of Ukraine-related disinformation compare to the dissemination of its corresponding debunks, and is there a causal relationship between them?*

This thesis delves into this by conducting a comparative analysis of engagement, themes, and causality using Ukraine-related debunks and disinformation. Specifically, the Granger causality test is employed to assess the effectiveness of debunking in countering Ukraine-related disinformation on social media platforms. Answering this question contributes to evaluating the current efficacy of debunking for mitigating disinformation on social media platforms. Furthermore, this analysis also provides insights into the potential effectiveness of automated strategies for detecting debunked narratives on social media platforms, as addressed in the second primary research question (2). The problem is studied in detail, and the findings addressing this research question are presented in Chapter 3.

2. ***What are the effective ways by which we can detect and alleviate the repeated dissemination of multilingual debunked narratives?***

(a) *How can advanced novel neural approaches enhance cross-lingual and multilingual retrieval, and what is their impact on improving cross-*

lingual debunked narrative retrieval?

Building on the foundation laid in the first primary research question (1), this research question explores novel neural approaches to enhance cross-lingual and multilingual retrieval. To address this research gap, the thesis reports the results of the University of Sheffield's participation in the Multilingual Information Access (MLIA) shared task on the COVID-19 multilingual semantic search. As part of this effort, the thesis demonstrates the performance of the proposed multistage BiCross encoder in comparison to state-of-the-art methods in the MLIA shared task in both monolingual and cross-lingual retrieval settings. Further details of this work are provided in Chapter 4.

Furthermore, this thesis leverages the knowledge gained through MLIA participation to enhance cross-lingual debunked narrative retrieval. First, a challenging benchmark dataset is developed that stands out as a comprehensive resource compared to its counterparts. Subsequent experiments were conducted to evaluate the performance of state-of-the-art cross-lingual retrieval models in identifying debunked narratives. Drawing inspiration from our proposed multistage BiCross encoder, this research also proposes two multistage retrieval methods that effectively address the cross-lingual nature of the task. The details of this work and its outcomes are presented in Chapter 5.

(b) *What are the key challenges and opportunities in cross-lingual debunked narrative retrieval? Can models trained on high-resource languages help low-resource languages in a zero-shot setting?*

This research question delves into the domain of cross-lingual retrieval of debunked narratives, aiming to identify both challenges and opportunities. In the fact-checking realm, where detecting the recurrence of debunked narratives is crucial, the question seeks to determine whether the models can transcend language barriers and adapt to various datasets and languages. By addressing this issue, the question addresses the overarching challenge of enabling the retrieval of debunked narratives with limited resources and advocates for solutions to promote cross-lingual and cross-dataset evaluations in fact-checking practices. In Chapter 5, the thesis conducts an in-depth study and addresses this research question.

(c) *How can unsupervised methods enhance real-time adaptation in debunked narrative retrieval without relying on human annotations?*

This research addresses the challenge of swiftly detecting false narratives tied to their topical nature. To overcome this, models must dynamically adapt and learn in real-time. Although using human-annotated datasets is a method to regularly update retrieval models on emerging topics, it is time-consuming, labour-intensive, and often limited in scale, which can impede the performance of the retrieval models. The thesis explores the potential of unsupervised methods for training debunked narrative retrieval models to match or surpass the performance of state-of-the-art methods without relying on human-annotated pairs. It aims to overcome the resource-intensive nature of retrieval model development in the fact-checking domain. This problem is thoroughly examined and addressed in Chapter 6.

To address the above-mentioned research questions, this thesis leverages diverse datasets, and cutting-edge deep learning techniques to analyse and mitigate the spread of debunked narratives. The chapters of this thesis are structured to provide a comprehensive examination of the problem, starting with the spatiotemporal characteristics of COVID-19-related debunked narratives, followed by an in-depth analysis of Ukraine-related disinformation, the development of multilingual semantic search methods, and the exploration of cross-lingual retrieval and unsupervised training approaches.

1.4 Thesis Overview: Publications and Contributions

This section outlines the contributions of this thesis. In particular, it adopts the thesis by publication format and comprises five distinct papers in the following order.

Chapter 2

Publication I: The False COVID-19 Narratives That Keep Being Debunked: A Spatiotemporal Analysis

This publication addresses the first primary research question (1a). It examines the spatiotemporal characteristics of similar or nearly duplicate false COVID-19 narratives that have been spreading in multiple languages, modalities and on various social media platforms in different countries, sometimes as much as several months after the first debunk of that narrative has been published by a fact-checker. The contributions of this publication are as follows:

- It utilises the CoronaVirusFacts Alliance database of COVID-19-related debunks and uncovers 10.3% of instances where similar false narratives related to COVID-19 are independently debunked multiple times, highlighting the widespread dissemination of misinformation during the pandemic.
- The spatiotemporal analysis reveals that misinformation on general medical advice is widespread globally and has been repeatedly debunked by multiple fact-checkers. Additionally, it finds that Facebook users consistently share false narratives without being aware that fact-checking organisations have already debunked these narratives in the past, sometimes in different languages.
- It provides compelling evidence for the need for a cross-lingual debunked narrative search tool in the fact-checking pipeline to efficiently determine if a narrative has been previously debunked in another language. Moreover, this approach aims to optimise resources and prevent the repetitive debunking of the same claims, particularly crucial given the labour-intensive nature of manual fact-checking.

This work has been published on ArXiv preprint server.

Singh, I., Bontcheva, K., & Scarton, C. (2021). The False COVID-19 Narratives That Keep Being Debunked: A Spatiotemporal Analysis. arXiv preprint arXiv:2107.12303.

The author contributed to the work by conceptualising, collecting data, developing methodology, validating, and writing.

Chapter 3

Publication II: Comparative Analysis of Engagement, Themes, and Causality of Ukraine-Related Debunks and Disinformation

This publication addresses the first primary research question (1b), where it examines the database of debunks related to the Ukraine conflict by different fact-checking organisations. The contributions of this publication are as follows:

- It offers a comprehensive comparative analysis of the dissemination of Ukraine-related disinformation and corresponding debunks on Twitter, revealing that, despite platform efforts, Ukraine-related disinformation spreads more widely than its debunks. The dataset used is made publicly accessible for further research.
- A bidirectional post-hoc analysis, employing Granger causality tests, impulse response analysis, and forecast error variance decomposition, reveals that debunks eventually exert a positive impact on reducing Ukraine-related disinformation, albeit not immediately.
- It uncovers approximately 18% of debunks that are associated with false claims already debunked by another fact-checking organisation in a different country or language. It suggests practical strategies to mitigate the impact of disinformation, such as utilising cross-lingual search and machine translation to expedite the debunking.

This work has been published in 13th International Conference on Social Informatics (SocInfo) 2023.

Singh, I., Bontcheva, K., Song, X., & Scarton, C. (2022, October). Comparative Analysis of Engagement, Themes, and Causality of Ukraine-Related Debunks and Disinformation. In International Conference on Social Informatics (pp. 128-143). Cham: Springer International Publishing.

The author contributed to the work by conceptualising, collecting data, developing methodology, validating, and writing.

Chapter 4

Publication III: Multistage BiCross encoder for multilingual access to COVID19 health information

This publication addresses the second primary research question (2a), presenting the experimental results from the participation in the Multilingual Information Access (MLIA) shared task on COVID-19 multilingual semantic search. The contributions of this publication are as follows:

- Multistage BiCross encoder method, which is a three-stage ranking pipeline that uses the Okapi BM25 retrieval algorithm and state-of-the-art multilingual transformer-based bi-encoder and cross-encoder by aggregating sentence-level relevance scores for the task of COVID-19 multilingual semantic search.

- Experiments with different types of search queries in order to establish the best performing ones for retrieving COVID-19 health information across millions of documents, in multiple languages. It also presents ways to combine scores from different stages using various rank fusion algorithms.
- An extensive comparison of our runs with other participant runs demonstrates the effectiveness of our methods in achieving high precision for top-ranked documents, as well as high recall for all retrieved documents in both monolingual and cross-lingual search settings.

This work has been published in the Public Library of Science Journal.

Singh, I., Scarton, C., & Bontcheva, K. (2021). Multistage BiCross encoder for multilingual access to COVID-19 health information. PLOS ONE 16(9): e0256874.

The author contributed to the work by conceptualising, developing methodology, validating, and writing.

Chapter 5

Publication IV: Breaking Language Barriers with MMTweets: Advancing Cross-Lingual Debunked Narrative Retrieval for Fact-Checking

This publication addresses the second primary research question (2a, 2b). This study addresses the understudied problem of cross-lingual debunked narrative retrieval, introducing a novel dataset and conducting experiments to benchmark retrieval models, revealing challenges and insights for optimising models to enhance fact-checking efforts. In particular, the novel contributions of this publication are:

- The Multilingual Misinformation Tweets (MMTweets): a novel benchmark that stands out, featuring cross-lingual pairs, images, and fine-grained human annotations, making it a comprehensive resource compared to its counterparts. In total, it comprises 1,600 query tweet claims (in Hindi, English, Portuguese & Spanish) and 30,452 debunk corpus (in 11 different languages) for retrieval. The dataset used is made publicly accessible for further research.

- An extensive evaluation of state-of-the-art (SOTA) cross-lingual retrieval models on the MMTweets dataset. It also introduces two multistage retrieval methods (*BE+CE* and *BE+GPT3.5*) adapting earlier approaches to effectively address the cross-lingual nature of the X-DNR task. Nevertheless, the results suggest that dealing with multiple languages in the MMTweets dataset poses a challenge, and there is still room for improvement in models.
- A comprehensive evaluation aims to investigate: 1) cross-lingual transfer and generalisation across languages within MMTweets; 2) how challenging it is for models trained on existing datasets to transfer knowledge to the MMTweets test set; 3) the impact of the type and count of negative pairs on the model's performance; and 4) insights into the retrieval latency of different models.

This work is currently under review.

Singh, I., Scarton, C., Song, X., & Bontcheva, K. (2023). Finding Already Debunked Narratives via Multistage Retrieval: Enabling Cross-Lingual, Cross-Dataset and Zero-Shot Learning. arXiv preprint arXiv:2308.05680.

The author contributed to the work by conceptualising, collecting data, developing methodology, validating, and writing.

Chapter 6

Publication V: UTDRM: Unsupervised Method for Training Debunked Narrative Retrieval Models

This publication answers the second primary research question (2c). This work proposes a novel Unsupervised Method for Training Debunked Narrative Retrieval Models (UTDRM) in a zero-shot setting, eliminating the need for human-annotated pairs. It leverages fact-checking articles for the generation of synthetic topical claims and employs a neural retrieval model for training. The main contributions of this publication are:

- `UTDRM`, a two-step method for training debunked narrative retrieval models that achieves comparable or superior retrieval scores to supervised models, all without relying on annotations. A comprehensive evaluation of `UTDRM` across seven public datasets establishes its efficacy and generalisability in retrieving accurate debunks for misinformation in tweets, political debates, and speeches.

- A large-scale dataset of synthetic topical claims created using two topical claim generation techniques based on text-to-text transformer-based models and large language models (LLMs).
- Extensive ablation experiments that assess the impact of different factors on UTDRM's performance. This includes: 1) the volume of fact-checking articles utilised, 2) the number of synthetically generated claims used for training, 3) the proposed *entity inoculation* method, and 4) the usage of LLMs, such as LLaMA and ChatGPT, for retrieval.

This work has been published in the EPJ Data Science Journal.

Singh, I., Scarton, C., & Bontcheva, K. (2023). UTDRM: unsupervised method for training debunked-narrative retrieval models. EPJ Data Science, 12(1), 59.

The author contributed to the work by conceptualising, collecting data, developing methodology, validating, and writing.

Chapter 7

Conclusions and Future Work

This thesis concludes with **Chapter** 7, which summarises the thesis, revisits the proposed research questions, and briefly discusses directions for future work.

Other Research Contributions

The author is the main contributor to all the chapter publications mentioned in this thesis. In addition, the author has also contributed to other publications that helped strengthen his PhD work. However, either the research doesn't align with the thesis topic or the author isn't the primary contributor, and hence this thesis excludes the following research.

***Singh, I.**, Li, Y., Thong, M., & Scarton, C. (2022, July). GateNLP-UShef at SemEval-2022 Task 8: Entity-Enriched Siamese Transformer for Multilingual News Article Similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 1121-1128).*

*Jiang, Y., Song, X., Scarton, C., **Singh, I.**, Aker, A., & Bontcheva, K. (2023, September). Categorising Fine-to-Coarse Grained Misinformation: An Empirical Study of the COVID-19 Infodemic. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (pp. 556-567).*

*Mu, Y., Jiang, Y., Heppell, F., **Singh, I.**, Scarton, C., Bontcheva, K., & Song, X. (2023). A Large-Scale Comparative Study of Accurate COVID-19 Information versus Misinformation. ICWSM TrueHealth Workshop 2023.*

*Song, X., Petrak, J., Jiang, Y., **Singh, I.**, Maynard, D., & Bontcheva, K. (2021). Classification aware neural topic model for COVID-19 disinformation categorisation. PloS one, 16(2), e0247086.*

*Jaidka, K., Ceolin, A., **Singh, I.**, Chhaya, N., & Ungar, L. (2021, June). WikiTalkEdit: A Dataset for modelling Editors' behaviours on Wikipedia. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021) (pp. 2191-2200).*

Chapter 2

The False COVID-19 Narratives That Keep Being Debunked: A Spatiotemporal Analysis

Iknoor Singh, Carolina Scarton and Kalina Bontcheva
Department of Computer Science, The University of Sheffield, UK

Abstract

The onset of the COVID-19 pandemic led to a global infodemic that has brought unprecedented challenges for citizens, media, and fact-checkers worldwide. To address this challenge, over a hundred fact-checking initiatives worldwide have been monitoring the information space in their countries and publishing regular debunks of viral false COVID-19 narratives. This study examines the database of the CoronaVirusFacts Alliance, which contains 10,381 debunks related to COVID-19 published in multiple languages by different fact-checking organisations. Our spatiotemporal analysis reveals that similar or nearly duplicate false COVID-19 narratives have been spreading in multiple modalities and on various social media platforms in different countries, sometimes as much as several months after the first debunk of that narrative has been published by an International Fact-checking Network (IFCN) fact-checker. We also find that misinformation involving general medical advice has spread across multiple countries and hence has the highest proportion of false COVID-19 narratives that keep being debunked. Furthermore, as manual fact-checking is an onerous task in itself, therefore the need to repeatedly

> debunk the same narrative in different countries is leading, over time, to a significant waste of fact-checker resources. To this end, we propose the idea of including a multilingual debunk search tool in the fact-checking pipeline, in addition to recommending strongly that social media platforms need to adopt the same technology at scale, so as to make the best use of scarce fact-checker resources

2.1 Introduction

The COVID-19 pandemic has not only triggered a global health emergency but has also led to the emergence of a worldwide infodemic, commonly referred to as a disinfodemic (Posetti and Bontcheva, 2020). In 2020, virtually everyone encountered or was exposed to various false claims concerning the origin, transmission, and medical treatments of the coronavirus[1]. Numerous studies (Limaye et al., 2020; Tasnim et al., 2020) have indicated that a majority of these claims originate on various social media platforms, raising concerns about their authenticity due to the lack of a reliable method for swiftly assessing the credibility of the online content. These unverified claims often fall into the category of misinformation, where the person spreading the claim is unaware of its falsity. Additionally, there is disinformation, involving the intentional spread of false information to deceive (Bontcheva et al., 2020). Both misinformation and disinformation have the potential to inflict significant harm[2]. On the other hand, despite the substantial growth in the number of fact-checking initiatives, these efforts are still unable to effectively mitigate the impact of dis/misinformation in the early stages of its spread due to limited resources (Nakov, 2020; McGlynn et al., 2020; Burel et al., 2020).

Furthermore, a report (FullFact, 2024b) by the UK's independent fact-checking organisation FullFact shows that there have been cases where similar narratives disseminated in different countries at different times have been debunked by multiple fact-checking organisations, given that the debunk (or fact-check) for that narrative already existed before. However, the previous study (FullFact, 2024b) was small-scale and lacked in-depth analysis, a gap we aim to address in this paper. In particular, it is unclear how frequently the same false narratives are spread and debunked across different languages or countries. In this paper, we utilise the International Fact-checking Network (IFCN) CoronaVirusFacts Alliance fact-checks database to find all duplicate debunks of the same false narratives concerning COVID-19. While it is possible that these duplicate debunks were generally published on days that lie in proximity to the publication date of the first debunk, our analysis finds that such duplicates differ by weeks and perhaps even by

[1] https://www.poynter.org/ifcn-covid-19-misinformation/

[2] https://www.bbc.com/news/world-53755067

months from their first appearance. These duplicate debunks usually arise when the same narratives are shared recurrently on various social media platforms in different countries at different times[3]. Although there could be multiple reasons why people persistently repeat debunked narratives (Lewandowsky et al., 2012; Ecker et al., 2010), one notable factor is that well-known figures, such as politicians, are known to reiterate false statements consistently (Nyhan and Reifler, 2010; Pillai and Fazio, 2021). Another possible reason, which we extensively explore in this paper, is that the debunk published in one language might not be available in another language, preventing the spreader from being aware of its debunk. In particular, we address the following research questions in this paper,

RQ1 Does the database of COVID-19-related debunks contain duplicate debunks of the same false narrative? In the case of duplicate debunks, what is the temporal gap between them, i.e. can the same false narrative resurface again significantly later and spread unhindered by the platforms' moderation algorithms in a different language or country?

RQ2 What are the spatiotemporal characteristics of recurrent debunked narratives, and how do these characteristics differ in terms of country, social media platform, and modality of content?

RQ3 What types of misinformation is most prevalent and has been debunked by multiple fact-checkers across different countries?

RQ4 Why integrate a multilingual debunked narrative search tool into the fact-checking pipeline to detect previously debunked narratives in multiple languages?

In this paper, we uncover numerous cases where similar debunked narratives spread at different times, varying in terms of country, social media platform, and modality of content. These narratives usually stem from an original factually inaccurate claim. Additionally, the recurrent spread of narratives of the same false claim gives rise to debunks from multiple fact-checking organisations in different languages. In this paper, we refer to these as "duplicate claim debunks" since they all debunk narratives of the same claim. The term "debunked narratives" or "debunked claims" refers to false narratives or claims that have undergone prior debunking or have been proven inaccurate by professional fact-checkers. Finally, we identify all such duplicate claim debunks in the IFCN database (Section 2.2).

We further investigate the spatiotemporal characteristics of the spread of debunked narratives. The analysis reveals that narratives related to general medical

[3] https://www.aljazeera.com/news/2020/12/2/trump-releases-video-repeating-debunked-election-fraud-claims

advice are particularly prevalent, having disseminated across multiple countries and been debunked multiple times. For instance, narratives regarding the purported benefits of consuming alkaline-rich food to eliminate coronavirus were initially debunked in Europe. Nevertheless, these narratives persisted, as they were again debunked by fact-checking organisations in Asian, South American, and North American countries. Furthermore, the findings also reveal that Facebook users contribute to most of the misinformation, as the same false narratives keep appearing on the platform, oblivious to the fact that the fact-check articles for those narratives have already been published in the past, either in the same language or in a language different from what the user posts in.

Lastly, there is a growing interest in developing automated fact-checking systems (Zhou and Zafarani, 2020; Singh et al., 2020; Thorne and Vlachos, 2018). In this context, before fact-checking a new claim, it is crucial to prevent the spread of narratives that have already been debunked. For instance, a prior study (Reis et al., 2020) on WhatsApp public groups in India and Brazil identified a significant amount of misinformation in the form of images shared within the groups, even after undergoing fact-checking. This recurrent spread of debunked narratives has led to the urgent need for retrieval systems to find fact-checked claims. Recent efforts have been made to address this internal gap (Barrón-Cedeno et al., 2020; Shaar et al., 2020a), with researchers focusing on detecting previously debunked narratives in a monolingual setting. However, this paper underscores the importance of including multilingual debunked narratives in the fact-checking pipeline to determine whether a narrative spreading in one language has already been debunked in the same or a different language (cross-lingual setting). Despite the significance of searching for previously debunked narratives in a multilingual setting, it has largely been overlooked by the research community. Furthermore, given the labour-intensive nature of current fact-checking processes, the ability to search for debunked narratives in a cross-lingual setting can prevent the unnecessary duplication of efforts in debunking the same narratives repeatedly. This approach would allow resources to be allocated more efficiently, enabling the timely fact-checking of other unsubstantiated claims.

In the next section (Section 2.2), we discuss the method used to perform the analysis. Section 2.3 mentions the main finding of this paper and in Section 2.4, we conclude this paper.

2.2 Method

To address the research questions outlined in Section 2.1, we utilise the CoronaVirusFacts Alliance database led by the IFCN Poynter. The IFCN Poynter database comprises debunks from over 100 organisations in 70 countries, covering around 40 languages. All IFCN fact-checkers adhere to specific principles regarding good practices in debunking. We use the IFCN Poynter[4] website to collect all claims that underwent fact-checking in 2020.

We crawl a total of 10,381 claims related to COVID-19 along with their corresponding debunk article page. In addition to the fields provided by the Poynter website[5], we extract the following information fields for each debunked claim on the IFCN Poynter website:

- 'Claim': Original debunked claim statement from the IFCN Poynter website.
- 'Country': List of countries where the claim has spread.
- 'Fact-checking Organisation': Name of the fact-checking organisation that has debunked the claim.
- 'Debunk Link': Link to the fact-checking article about the claim.
- 'Debunk Language': Language used in the fact-checking article detected using *langdetect* Python library[6].
- 'Debunk Date': Date of publication of the fact-checking article detected using *htmldate* Python library[7].
- 'Social media website': List of websites where claims appeared extracted from fact-checking articles using the JAPE rule (Song et al., 2021).
- 'Modality of content': Modality of claims extracted from fact-checking articles using the JAPE rule (Song et al., 2021).

To identify similar debunked narratives, we employ the claim field from the debunks collected earlier to identify semantically similar claims that were debunked by multiple fact-checkers. We formulate this as a retrieval problem, where for each claim field, we conduct a semantic search across all other claims in the dataset. Each claim used as a query is denoted as a "query claim debunk," and their retrieved semantically similar claims are referred to as "duplicate claim debunks".

For retrieval, we initially standardise all references to COVID-19 in the claims (e.g., SARS-CoV-2, COVID-19, 2019-nCoV, COVID) with a unified representation, namely "coronavirus." Following this, we employ a multistage approach (Nogueira and Cho, 2019; Singh et al., 2021b) (see Chapter 4) involving BM25 Okapi algorithm for initial lexical retrieval and a subsequent neural retrieval stage utilising a state-of-the-art text similarity model based on RoBERTa cross-encoder model(Liu et al., 2019) to identify semantically similar claims. We ensure robust and reliable data by setting a strict 0.8 similarity score threshold and manually

[4] https://www.poynter.org/ifcn-covid-19-misinformation/

[5] https://www.poynter.org/wp-content/uploads/2020/05/CORONAVIRUS-FACTS-RFP-Data-Description.pdf

[6] https://pypi.org/project/langdetect/

[7] https://pypi.org/project/htmldate/

Query Claim Debunk			Duplicate Claim Debunk		
Claim	**Debunk Org**	**Date**	**Claim**	**Debunk Org**	**Date**
Vitamin C can cure coronavirus.	Détecteur de rumeurs	2020/04/24	Vitamin C can cure COVID-19.	JTBC news	2020/03/04
			Vitamin C is a miracle cure for the novel coronavirus.	Källkritikbyrån	2020/03/05
			Vitamin C prevents coronavirus.	TjekDet.dk	2020/03/04
			Vitamin C will protect you from the coronavirus.	AFP	2020/03/13
			Consuming large doses of Vitamin C can stop the spread of coronavirus.	Vishvas News	2020/03/04
			Vitamin C can "stop" the new coronavirus.	FactCheck.org	2020/02/12
			The coronavirus can be slowed or stopped with the "immediate widespread use of high doses of vitamin C."	PolitiFact	2020/01/27
Aborted fetal cells are in the COVID-19 vaccine	Science Feedback	2020/11/16	Vaccines, including the one for COVID-19, include aborted fetal tissues.	VoxCheck	2020/04/28
			Aborted babies used to develop COVID-19 vaccine	AAP FactCheck	2020/10/22
			CoronaVac uses cells from aborted fetuses.	Aos Fatos	2020/07/28

Table 2.1: *Some examples of query claim debunks and their corresponding duplicate claim debunks. Note 1) Fact-checking organisation of the query claim debunk and duplicate claim debunks is different. 2) Date of publication of the duplicate claim debunk is before the date of publication of the query claim.*

verifying the quality to include only relevant duplicate claim debunks. In addition to this, there are two retrieval constraints: 1) The fact-checking organisation of the query claim debunk is different from the fact-checking organisation of the retrieved duplicate claim debunk 2) The date of publication of the duplicate claim debunk is before the date of publication of the query claim debunk. These constraints ensure that we do not get duplicate cases and only the ones which have the debunks from different fact-checking organisations published in the past. Moreover, the IFCN Poynter[8] states that the countries mentioned on the debunked claim webpage are where the falsehood was spreading. Therefore, we infer that the claims which have been debunked at different times are the claims that have been spreading in distinct countries at different times.

Finally, for each query claim debunk, we retrieved N ($\geq$1) duplicate claim debunks. For certain analyses (see Section 2.3), we transformed this from a one-to-many relationship into a one-to-one relation between query claim and duplicate claim debunks. Table 2.1 shows examples of query claim debunks and their corresponding duplicate claim debunks.

2.3 Findings

We divide this section into four parts, where each of the below-mentioned findings addresses the four research questions mentioned in section 2.1 in order.

Finding 1. COVID-19 debunks in the IFCN database contain a considerable number of fact-checking articles debunking similar narratives that originate in different countries at different times.

Out of a total of 10,381 debunks in the IFCN database, we identify 1,070 debunks that already have a debunk about a similar claim from a different fact-checking organisation published in the past. This accounts for 10.3% of all the debunks in the IFCN database. Throughout this paper, we refer to these 1,070 debunks as “query claim debunks” and their duplicate counterparts as “duplicate claim debunks” (see Section 2.2). In other words, for each query claim debunk, we have N ($\geq$1) duplicate claim debunks from different fact-checking organisations published in the past. Please refer to Appendix 2.6.1 for the cluster plot visualisation for duplicate claim debunks.

Figure 2.1 (left) is the pie chart distribution of the top 10 countries of query claim debunks, i.e., the top countries where claims already debunked are spreading. India and the United States have the largest number of recurring false narratives,

[8] https://www.poynter.org/wp-content/uploads/2020/05/CORONAVIRUS-FACTS-RFP-Data-Description.pdf

and these get debunked multiple times, leading to a waste of fact-checkers' efforts. It indicates that these countries, particularly India with a total proportion of 19%, are most vulnerable to the spread of narratives that have already been debunked in the past. In general, this also suggests a lack of awareness among the people about prior fact-checked information.

Figure 2.1 (right) illustrates the pie chart distribution of the top 10 fact-checking organisations of query claim debunks, i.e. the top fact-checking organisations that are debunking narratives for which debunks already existed in the past. The results align with Figure 2.1 (left), where Vishvas News, an Indian fact-checking website, publishes a large number of debunks about previously fact-checked claims.

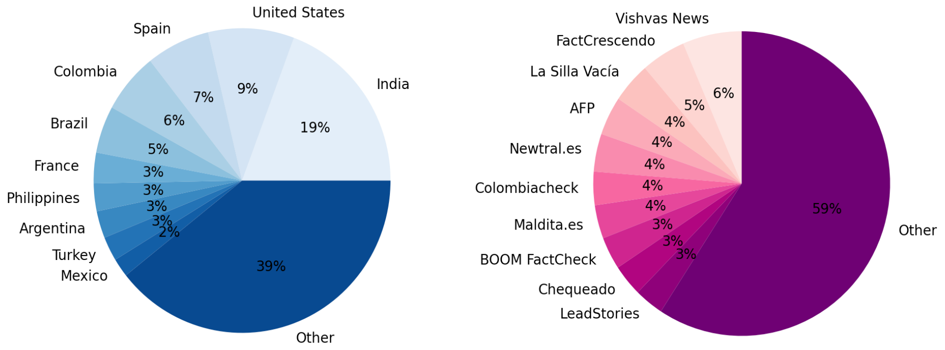


Figure 2.1: *Left: Pie chart distribution for top 10 countries where the claims already debunked were spreading. Right: Pie chart distribution for top 10 fact-checking organisations that published fact-checking articles about the claims that were debunked in the past.*

The difference in days between the publication date of query claim debunks and the duplicate claim debunks is depicted in Figure 2.2. The histogram plot shows the weekly count with the bin interval set at 7 days. For instance, the first bar indicates that there are 884 cases where the publication date difference between query claim debunk and duplicate claim debunk is one week or less. Similarly, the second bar shows nearly 300 cases with a fortnight difference, and so forth. This reveals that misinformation persists and gets debunked multiple times even after relevant debunks are already available. This is worrisome and the subsequent findings help us understand the reasons for the existence of such duplicate claim debunks.

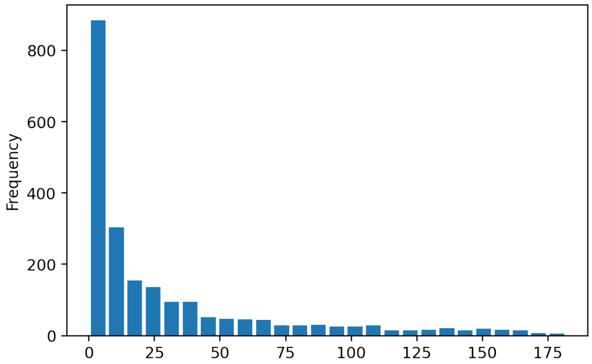


Figure 2.2: *Histogram plot for days difference between query claim debunks and duplicate claim debunks. (Bin set at an interval of 1 week).*

Finding 2. Spatiotemporal characteristics of similar false narratives and their transition between countries, social media platforms and modalities of content.

The spatiotemporal characteristics of both query claim debunks and duplicate claim debunks can help reveal how information flows or changes between different debunks. In Figure 2.3, pie charts illustrate the movement of similar false claims between different countries. For simplicity, we only consider the top 10 country pairs, where Figure 2.3 (left) shows the count of cases where both countries are the same, and Figure 2.3 (right) shows cases where both countries are different.

Since the date of publication of the duplicate claim debunk is before the publication date of the query claim debunk (see Section 2.2), the symbol "←" between the countries can be treated as the flow of false claims between different country pairs. For example, "India ← United States" indicates that there are around 40 cases where the flow of false claims is from the United States to India. We find that the movement of similar false claims is highest between India and the United States, followed by movement from Spain to Columbia. The conceivable reason for this could be the common language of English and Spanish, respectively, for each of the cases.

Figure 2.4 (left) illustrates the change in social media platforms of the claims fact-checked in both query claim debunk and duplicate claim debunk. In other words, it provides insights into the movement of similar false claims from one social media website to another. It suggests that for similar claims, the spread within Facebook itself is the highest, with around 800 cases, followed by occurrences from WhatsApp to Facebook, which has just over 200 instances. This is particularly

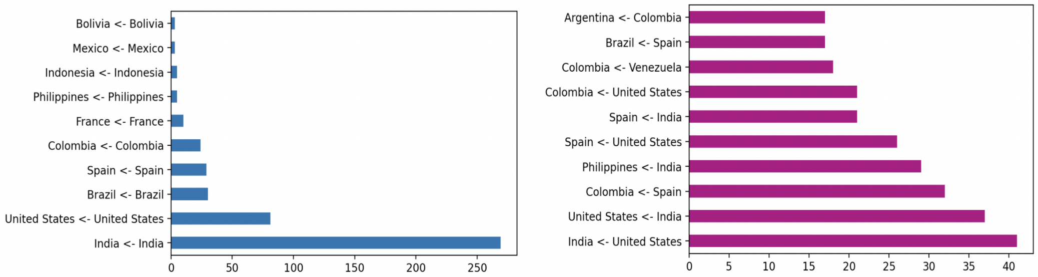


Figure 2.3: *The movement of similar false claims between different country pairs. The bar chart on the left shows the top 10 counts of cases where both the countries are same and the bar chart on the right depicts the top 10 cases where both countries are different.*

concerning given that Facebook, increasingly used as a primary source of news (Bridgman et al., 2020), allows the wide dissemination of content whose falsity has already been fact-checked in the past.

According to a Pew Research report of 2020[9], 52% of American adults get news from digital platforms, out of which more than half of the people (53%) said that they consume news from social media platforms. This is worrisome, especially during the time of COVID-19 pandemic[10] as most false claims regarding government rules, virus cures, vaccines, and more originate on various social media platforms, making users vulnerable to believing misinformation. Although these social media platforms have made efforts[11] to mitigate the spread of false narratives, it remains prevalent, as shown in this study and supported by previous research (Burel et al., 2020).

Furthermore, people use different modalities of content such as text, images, videos, etc., to spread factually inaccurate claims. Figure 2.4 (right) displays the transition in the modality of claims fact-checked in both query claim debunk and duplicate claim debunk. While the modality for text, video, and image remains consistent, there are also considerable cases where there is a transition between the modalities of content that state the same things.

Figure 2.5 shows the difference in the language used in the fact-checking articles

[9] https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/

[10] https://www.pewresearch.org/fact-tank/2021/08/24/about-four-in-ten-americans-say-social-media-is-an-important-way-of-following-covid-19-vaccine-news/

[11] https://www.washingtonpost.com/technology/2020/11/09/facebook-twitter-election-misinformation-labels/

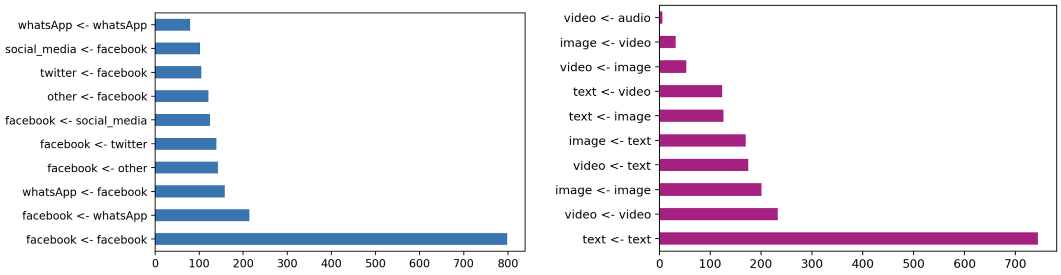


Figure 2.4: *Left: Transition in social media platforms. Right: Transition between modality of content.*

for both the query claim debunk and the duplicate claim debunk for the top 10 language pairs. Here, the first symbol represents the ISO-39 language code of the query claim debunk, and the second one is the language used in duplicate claim debunk articles. It's noteworthy that for monolingual pairs, it's unusual to observe a significant number of duplicate claim debunks for which debunks already exist in the same language. Additionally, there are a considerable number of bilingual pairs, indicating the necessity for cross-lingual search before debunking a new claim, as discussed later in Finding 4.

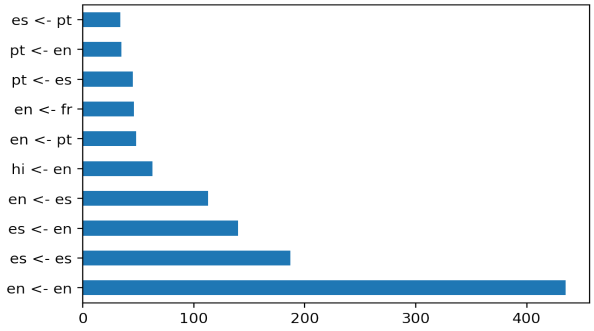


Figure 2.5: *Top 10 count of cases showing the difference in the language used in the fact-checking articles for both the query claim debunk and the duplicate claim debunk. ISO-39 language code is used to denote the language.*

Finding 3. COVID-19 misinformation involving general medical advice got spread across multiple countries and hence has the highest proportion of duplicate claim debunks in our dataset.

To assist fact-checkers in quick debunking, prior work (Brennen et al., 2020) categorised COVID-19 misinformation into various types, such as medical advice, virus origin, etc. We label the claims using CANTM model (Song et al., 2021) to understand which kinds of claims spread the most and have the highest number of duplicate claim debunks.

Figure 2.6 (top) depicts a pie plot of the categories of claims for which multiple debunks exist. The COVID-19 misinformation categories include PubAuthAction (public authority), CommSpread (community spread and impact), GenMedAdv (medical advice, self-treatments, and virus effects), PromActs (prominent actors), Consp (conspiracies), VirTrans (virus transmission), VirOrgn (virus origin and properties), PubPrep (public reaction), Vacc (vaccines, medical treatments, and tests), and None (other). Misleading medical advice appears to be the most consistent topic of misinformation, accounting for the highest proportion at 33%, followed by conspiracy theories, public authority actions, and community spread-based false claims, each making up 13% of all cases. Overall, these recurring topics underscore the necessity for more efficient resource allocation to mitigate redundant debunking efforts.

Furthermore, Figure 2.6 (bottom) is a scatter plot demonstrating the difference in days between query claim and duplicate claim debunks for different categories of claims. We observe that claims on general medical advice are most densely spread, indicating many cases where the publication date of duplicate claim debunks differs by several days. Claims about vaccines and conspiracy theories also exhibit a dense spread compared to others, which are denser on the lower end, depicting that the difference in days between the publication date of query claim and duplicate claim debunk is not much.

In Table 2.2, we examine the top six words (after removing all non-useful words) in various categories of claims that have multiple debunks. Words such as “Water”, “lemon” etc are most dominant in misinforming medical advice, while “Honjo Tasuku” and “Gates” can be observed in repeated claims involving conspiracy.

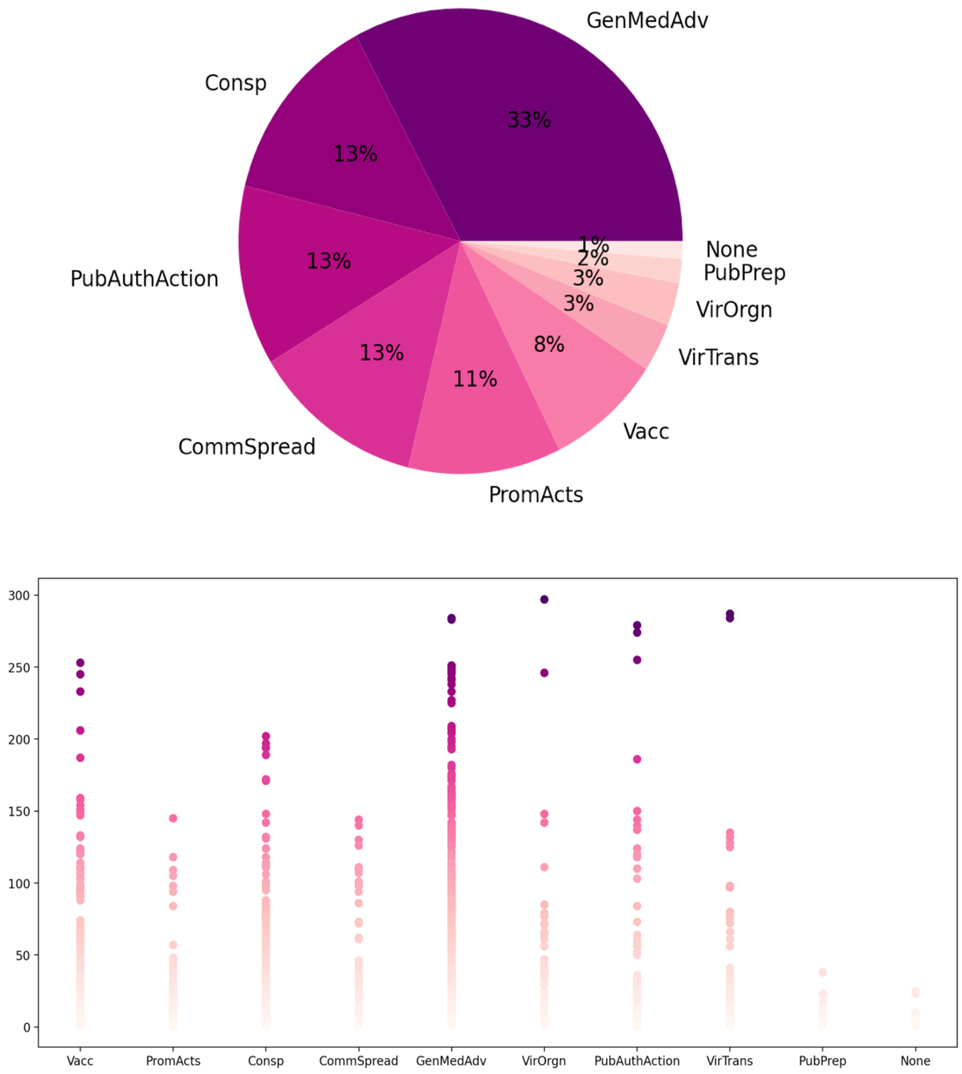


Figure 2.6: *Top: Pieplot for categories of claims. Bottom: Difference in days between query claim debunks and duplicate claim debunks for different categories of claims.*

We further examine claims that are widely spread and have debunks published at different times of the year. Figure 2.7 presents a sample of 10 false claims about fallacious medical advice, including cures, remedies, and prevention methods specific to COVID-19. We find that the duplicate claim debunks for these claims are spread across the entire year and are published in different languages.

Class	Words							
GenMedAdv	water	salt	lemon	cures	breath	vitamin	vinegar	tea
Consp	nobel	honjo	tasuku	lab	wuhan	gates	outbreak	china
PubAuthAction	people	china	patients	government	police	india	court	video
CommSpread	people	photo	italy	video	patients	china	coffins	victims
PromActs	president	ronaldo	cristiano	minister	hospitals	bill	charles	hotels
Vacc	vaccine	people	cure	bill	gates	dna	russia	pfizer
VirTrans	hypoxia	masks	use	mask	chicken	flu	creator	pcr
VirOrgn	video	wuhan	virus	china	market	bats	chicken	hubei
PubPrep	people	lions	streets	russia	homes	masks	berlin	pandemic

Table 2.2: *Top six words in different categories of claims that have multiple debunks; darker blue means higher volume.*

Subsequently, Figure 2.8 illustrates the timeline of debunks for claims about the consumption of an alkaline-rich diet to eliminate the coronavirus. From our dataset, it appears that the claim was first debunked in Spain in March 2020 and after a month a similar claim was debunked in Indonesia and the United States but it was still here to stay. It is surprising and yet worrisome that the same claim was again debunked in Turkey and Brazil in September and December respectively. One thing that might have led to this unknowing spread of previously debunked claims is the language of the fact-checking article, as they all differ (shown in Figure 2.8 with ISO-39 language codes enclosed in brackets after the name of the fact-check organisation).

We also investigate the language and modality of the claims and find that claims written in one language are sometimes transformed into other languages and varied modalities (eg. text to image) before being propagated to other countries. The social media platforms used to spread the claim in different countries also change over time. Figure 2.8 shows that the same claim was shared on Facebook, WhatsApp, and Twitter.

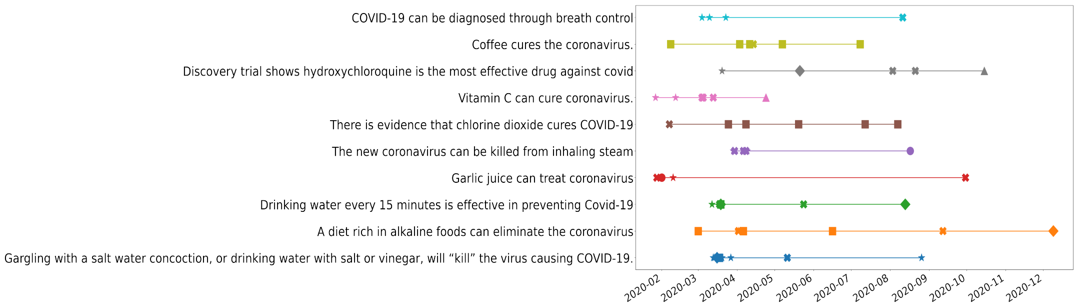
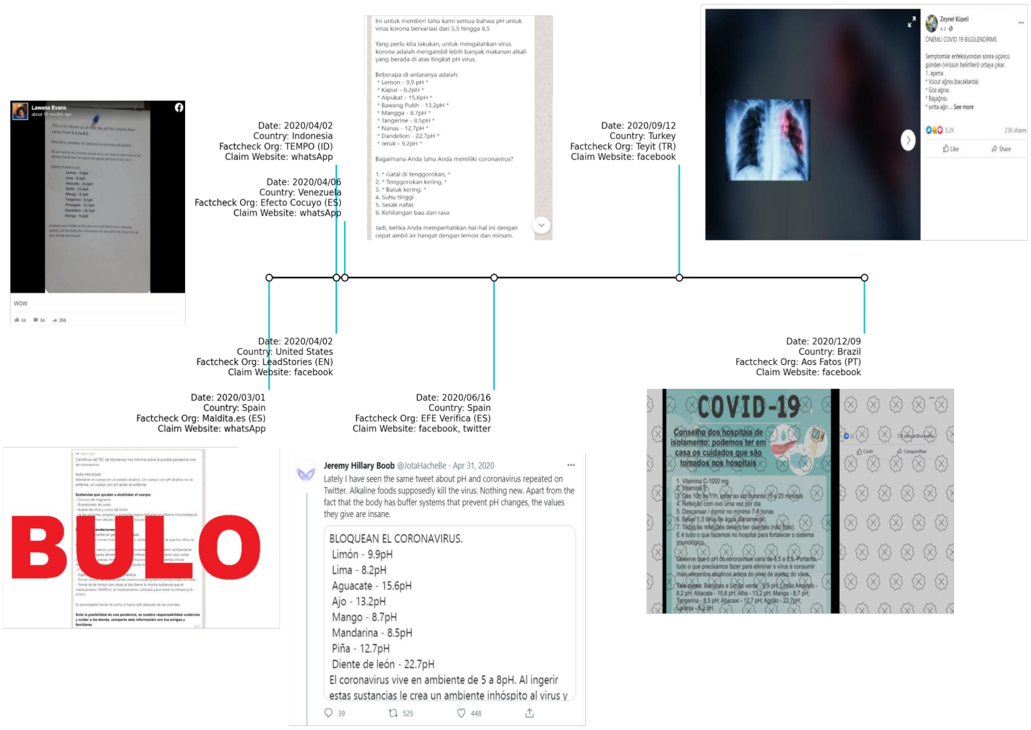


Figure 2.7: *Timeline for a sample of 10 claims about fallacious medical advice. Here the language of debunk article is denoted by different symbols like English: ★; Spanish: ■; Hindi: ●; Portuguese: ♦; French: ▲; Other:* ***X****;*



Ini untuk memberi tahu kami semua bahwa pH untuk virus korona bervariasi dari 5,5 hingga 8,5

Yang perlu kita lakukan, untuk mengalahkan virus korona adalah mengambil lebih banyak makanan alkali yang berada di atas tingkat pH virus.

Beberapa di antaranya adalah:
* Lemon - 9,9 pH *
* Kapur - 8.2pH *
* Alpukat - 15,6pH *
* Bawang Putih - 13,2pH *
* Mangga - 8.7pH *
* Tangerine - 8.5pH *
* Nanas - 12,7pH *
* Dandelion - 22,7pH *
* Jeruk - 9.2pH *

Bagaimana Anda tahu Anda memiliki coronavirus?

1. * Gatal di tenggorokan, *
2. * Tenggorokan kering, *
3. * Batuk kering. *
4. Suhu tinggi
5. Sesak nafas
6. Kehilangan bau dan rasa

Jadi, ketika Anda memperhatikan hal-hal ini dengan cepat ambil air hangat dengan lemon dan minum.

Jeremy Hillary Boob @JotaHacheBe · Apr 31, 2020

Lately I have seen the same tweet about pH and coronavirus repeated on Twitter. Alkaline foods supposedly kill the virus. Nothing new. Apart from the fact that the body has buffer systems that prevent pH changes, the values they give are insane.

BLOQUEAN EL CORONAVIRUS.
Limón - 9.9pH
Lima - 8.2pH
Aguacate - 15.6pH
Ajo - 13.2pH
Mango - 8.7pH
Mandarina - 8.5pH
Piña - 12.7pH
Diente de león - 22.7pH
El coronavirus vive en ambiente de 5 a 8pH. Al ingerir estas sustancias le crea un ambiente inhóspito al virus y

39 525 448

Figure 2.8: *A detailed timeline of claim: "A diet rich in alkaline foods can eliminate the coronavirus". All the images show the same claims being spread on different social media websites in different languages and varied modalities (top left and bottom right are the images shared on Facebook; top right and bottom centre are the images accompanied by some text shared on Facebook and Twitter respectively; top centre and bottom left show text shared on WhatsApp)*

Figure 2.9 illustrates conspiracy theories that have been debunked multiple times. The belief that COVID-19 is linked to 5G technology was common across many countries, despite having been debunked before. Additionally, there are numerous falsely attributed claims and conspiracies involving Bill Gates. For instance, Figure 2.10 displays the timeline of debunks about claims alleging a statement from Bill Gates that the COVID-19 vaccine can change human DNA. All the debunks appear in multiple languages at different times over the time span of five months from June to October 2020.

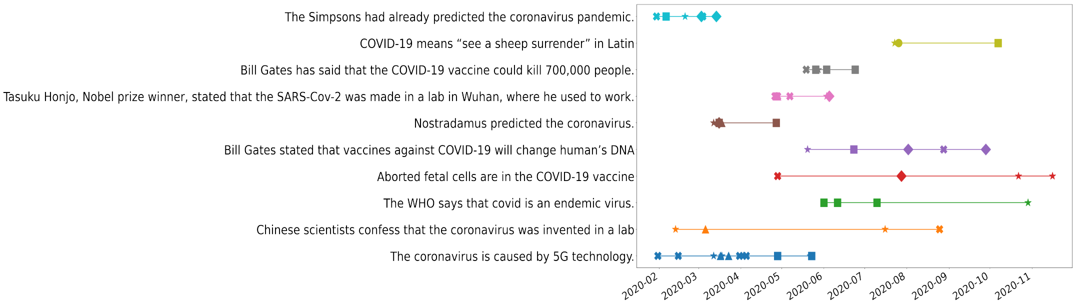


Figure 2.9: *Timeline for a sample of 10 conspiracy theories concerning COVID-19. Here the language of the debunk article is denoted by different symbols like English: ★; Spanish: ■; Hindi: ●; Portuguese: ⧫; French: ▲; Other: X*

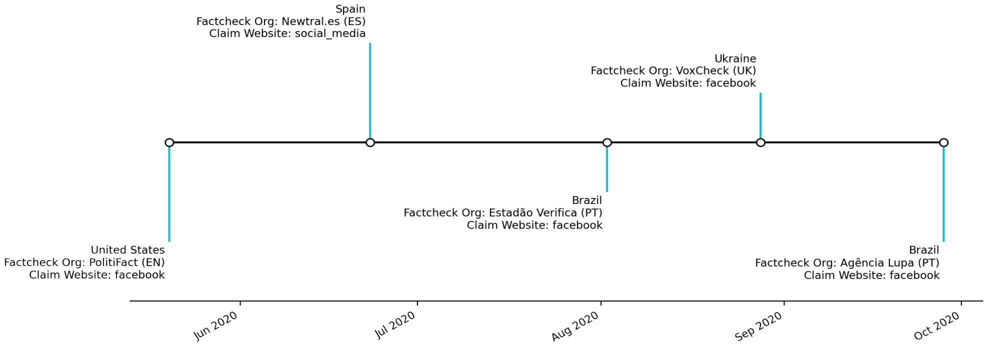


Figure 2.10: *Detailed timeline of claim: "Bill Gates stated that vaccines against COVID-19 will change human's DNA"*

Finding 4. The IFCN database mainly consists of cases where there isn't even a single duplicate claim debunk in the same language as that of the query claim debunk, highlighting the necessity of including multilingual debunk search in the fact-checking pipeline.

As mentioned earlier, we identified 1,070 debunks about claims that already have debunks published by an IFCN fact-checker. Among these 1,070 cases, there are a total of 627 (59%) instances for which we don't have a single duplicate claim debunk in the same language as that of the query claim debunk. Alternatively, this shows that if a person from some country is willing to search for fact-check articles about a claim that has already been debunked in a language different from what the person understands, then he/she might not be able to do so due to the language barrier. Although one can make efforts to search through the content in multiple languages, it's usually not done because it's inefficient and it's probably the reason claims spread, incognizant of the fact that they have already been debunked in the past. Therefore, the need for multilingual and cross-lingual debunk search in the initial stages of the fact-checking pipeline becomes imperative.

Before delving into fact-checking a claim, it is crucial to check whether the claim or its equivalent has already been debunked by a fact-checking organisation in a different language. While there are commercially available debunk database search tools by Google[12] and WeVerify[13], to the best of our knowledge, these tools are limited to monolingual search. Our analysis highlights the need for a cross-lingual/multilingual retrieval search, where a comprehensive pool of debunked narratives from around the world is considered, irrespective of the language used in the fact-checking article. Given the time-consuming nature of manual fact-checking, avoiding duplicated efforts in debunking narratives that have already been debunked in the past is paramount. Therefore, the ability to search for previously debunked narratives in multiple languages is beneficial for fact-checkers.

On the other hand, while it may be impossible to fact-check every claim, social media platforms can take the initiative to warn users before they share content containing previously debunked narratives. Over the years, numerous fact-checking organisations have emerged, accumulating a vast corpus of fact-checking articles (Augenstein et al., 2019; Shahi and Nandini, 2020; Gupta and Srikumar, 2021a) debunking various claims in different languages. This data can be effectively utilised to quickly debunk repeated false narratives appearing on various social media platforms, thereby limiting their spread and potential harm.

2.4 Conclusion

The onset of the COVID-19 pandemic led to a global infodemic that has brought unprecedented challenges for citizens, media, and fact-checkers worldwide. To address this challenge, over a hundred fact-checking initiatives worldwide have been monitoring the information space in their countries and publishing regular debunks of viral false COVID-19 narratives. In this paper, we examine the database of the CoronaVirusFacts Alliance, which contains 10,381 debunks related to COVID-19 published in multiple languages by different fact-checking organisations.

Our spatiotemporal analysis addressed the research questions outlined in the introduction. First, we confirmed the existence of duplicate debunks of the same false narratives across different countries and languages (RQ1), revealing signif-

[12] https://toolbox.google.com/factcheck/explorer

[13] https://weverify.eu/

icant temporal gaps between the initial and subsequent debunks. This demonstrates that false narratives can resurface months later, often without moderation by social media platforms. Second, we identified key spatiotemporal characteristics of these recurrent narratives, highlighting differences across countries, social media platforms, and content modalities (RQ2). Notably, misinformation involving general medical advice has spread across multiple countries and hence has the highest proportion of false COVID-19 narratives that keep being debunked (RQ3).

Lastly, we underscored the necessity of integrating a multilingual debunked narrative search tool into the fact-checking pipeline (RQ4). This tool could significantly enhance efficiency by preventing the redundant effort of debunking previously addressed claims, thereby optimising the use of fact-checker resources. Additionally, we strongly recommend that social media platforms adopt this technology at scale, so as to make the best use of scarce fact-checker resources.

2.5 Limitations and Future Work

Our work should be seen in light of the following limitations: i) For all the fact-checking articles debunking similar narratives, we did not consider any changes in rulings made by fact-checkers over time. In other words, we assume that if a claim is initially declared false by some fact-checking organisation, then it remains false irrespective of the time or place of debunking of a similar claim. This is something we plan to investigate in detail in our future work. ii) While the dataset utilised in our analysis may be considered weakly labelled, we mitigate this limitation by leveraging state-of-the-art semantic similarity models with a high threshold. Additionally, we conduct manual checks to ensure that only relevant duplicate claim debunks are included in our study. iii) The assumption that overlapping debunks indicate redundant efforts is valid but not entirely foolproof. Fact-checkers may have valid reasons to publish their own versions, renew existing checks, or add further evidence. Finally, we presume that the spread of debunked narratives is due to the spreader being unaware of previously debunked articles about similar narratives; however, there can be multiple possible reasons for this (Lewandowsky et al., 2012). Future work should delve deeper into these aspects, exploring the reasons behind duplicated debunks and examining the effectiveness of user engagement with debunks. The main aim of this study is to draw attention to the general public and fact-checkers regarding the presence of duplicate claim debunks, suggesting ways to mitigate the spread of debunked narratives and better deal with potential infodemics in the future.

2.6 Appendix

2.6.1 Gephi Plot

Out of 10,381 debunks in the IFCN database, we find 1070 debunked claims that already had a debunk about the same false narrative from a different fact-checking organisation in the past. We clustered together all such duplicate claim debunks which have more than three debunks that fact-check similar claims and produced a GRAPHML-file to visualise the clusters using java-based network analysis applications such as Gephi (Figure 2.11). The Fructhterman-Reingold force-directed graph drawing algorithm is used to visualise the network in a compact circle with coloured cluster separation based on the modularity class. Here, a node represents a debunk from the fact-checking organisation and the colour represents the cluster of all duplicate claim debunks. The claim statement for each cluster is mentioned as shown in Figure 2.11.

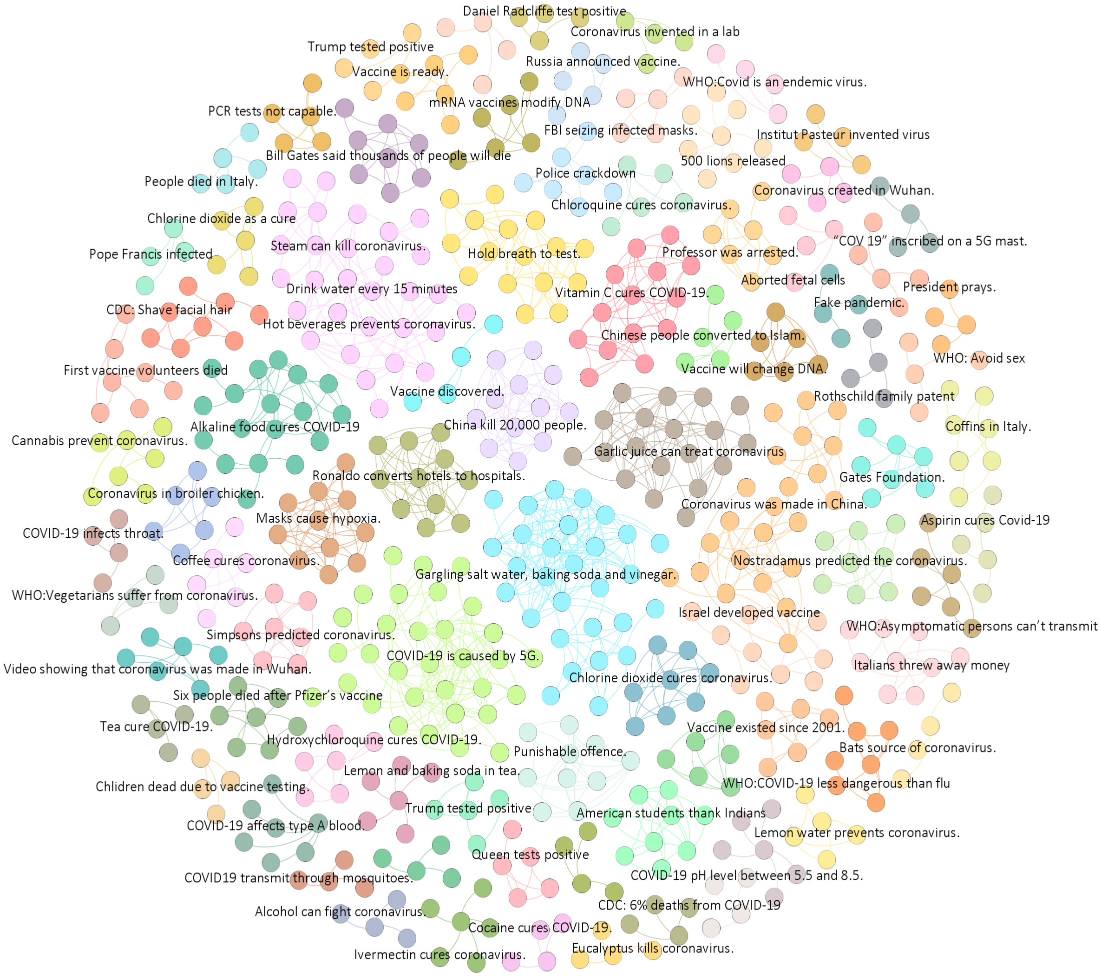

Daniel Radcliffe test positive
Coronavirus invented in a lab
Trump tested positive
Russia announced vaccine.
Vaccine is ready.
WHO:Covid is an endemic virus.
mRNA vaccines modify DNA
PCR tests not capable.
FBI seizing infected masks.
Institut Pasteur invented virus
Bill Gates said thousands of people will die
500 lions released
People died in Italy.
Police crackdown
Coronavirus created in Wuhan.
Chloroquine cures coronavirus.
Chlorine dioxide as a cure
"COV 19" inscribed on a 5G mast.
Steam can kill coronavirus.
Professor was arrested.
Pope Francis infected
Hold breath to test.
Aborted fetal cells
President prays.
Drink water every 15 minutes
Vitamin C cures COVID-19.
Fake pandemic.
CDC: Shave facial hair
Hot beverages prevents coronavirus.
Chinese people converted to Islam.
WHO: Avoid sex
First vaccine volunteers died
Vaccine will change DNA.
Vaccine discovered.
Rothschild family patent
Alkaline food cures COVID-19
China kill 20,000 people.
Coffins in Italy.
Cannabis prevent coronavirus.
Garlic juice can treat coronavirus
Ronaldo converts hotels to hospitals.
Gates Foundation.
Coronavirus in broiler chicken.
Coronavirus was made in China.
Masks cause hypoxia.
Aspirin cures Covid-19
COVID-19 infects throat.
Coffee cures coronavirus.
Nostradamus predicted the coronavirus.
Gargling salt water, baking soda and vinegar.
WHO:Vegetarians suffer from coronavirus.
Israel developed vaccine
WHO:Asymptomatic persons can't transmit
Simpsons predicted coronavirus.
COVID-19 is caused by 5G.
Italians threw away money
Video showing that coronavirus was made in Wuhan.
Chlorine dioxide cures coronavirus.
Six people died after Pfizer's vaccine
Tea cure COVID-19.
Vaccine existed since 2001.
Hydroxychloroquine cures COVID-19.
Punishable offence.
Bats source of coronavirus.
Lemon and baking soda in tea.
Chlidren dead due to vaccine testing.
WHO:COVID-19 less dangerous than flu
Trump tested positive
American students thank Indians
COVID-19 affects type A blood.
Lemon water prevents coronavirus.
Queen tests positive
COVID-19 pH level between 5.5 and 8.5.
COVID19 transmit through mosquitoes.
CDC: 6% deaths from COVID-19
Alcohol can fight coronavirus.
Cocaine cures COVID-19.
Eucalyptus kills coronavirus.
Ivermectin cures coronavirus.

Figure 2.11: *Cluster visualisation for duplicate claim debunks.*

Chapter 3

Comparative Analysis of Engagement, Themes, and Causality of Ukraine-Related Debunks and Disinformation

Iknoor Singh, Kalina Bontcheva, Xingyi Song and Carolina Scarton
Department of Computer Science, The University of Sheffield, UK

Abstract

This paper compares quantitatively the spread of Ukraine-related disinformation and its corresponding debunks, first by considering retweets, replies, and favourites, which demonstrate that despite platform efforts Ukraine-related disinformation is still spreading wider than its debunks. Next, bidirectional post-hoc analysis is carried out using Granger causality tests, impulse response analysis and forecast error variance decomposition, which demonstrate that the spread of debunks has a positive impact on reducing Ukraine-related disinformation eventually, albeit not instantly. Lastly, the paper investigates the dominant themes in Ukraine-related disinformation and their spatiotemporal distribution. With respect to debunks, we also establish that around 18% of fact-checks are debunking claims which have already been fact-checked in another language. The latter finding highlights an opportunity for better collaboration between fact-checkers, so they can benefit from and amplify each other's debunks through translation, citation, and early publication online.

3.1 Introduction

Following on from and interleaved with the COVID-19 infodemic, the war in Ukraine has unleashed a new large stream of mis- and disinformation (Aguerri et al., 2022), as evidenced, amongst others, by fact-checkers from the European Digital Media Observatory (EDMO) who found a record-high Ukraine-related disinformation in March 2022[1]. Examples include viral decontextualised videos from past[2] and a popular pro-Kremlin false narrative about the existence of a biolab in Ukraine funded by Joe Biden's son[3]. To counter this fast-flowing disinformation, the International Fact-checking Network (IFCN) fact-checkers are working together to maintain and publish a unified database of debunks of Ukraine-related disinformation[4]. In order to measure the effectiveness of these efforts, we carry out a comparative analysis of engagement, themes, and predictive causality of Ukraine-related debunks and disinformation.

The novel contributions of this paper are in answering the following three key research questions through a quantitative analysis of Ukraine-related disinformation and debunks on Twitter:

RQ1 What is the overall engagement of Ukraine-related disinformation and debunks on Twitter (Section 3.4)?

RQ2 Does the spread of debunks have a positive impact in reducing Ukraine-related disinformation (Section 3.5)?

RQ3 What are the underlying themes in Ukraine-related disinformation and their spatiotemporal characteristics on Twitter (Section 3.6)?

In the following sections, we will discuss first related work (Section 3.2) and then detail the data acquisition methodology for this study (Section 3.3).

3.2 Related Work

Ukraine-related pro-Kremlin disinformation (Yablokov, 2022) is not new (Mejias and Vokuev, 2017; Aguerri et al., 2022; Lange-Ionatamishvili et al., 2015). For instance, Lange-Ionatamishvili et al. (2015) and Mejias and Vokuev (2017) studied the spread of disinformation on social media after the 2014 annexation of Crimea by the Russian Federation, while Erlich and Garner (2021) investigated

[1] https://edmo.eu/fact-checking-briefs/

[2] https://www.politifact.com/factchecks/2022/may/10/facebook-posts/no-not-footage-ukraine-shooting-down-russian-plane/

[3] https://www.politifact.com/article/2022/apr/01/facts-behind-russian-right-wing-narratives-claimin/

[4] https://ukrainefacts.org/

Table 3.1: *Top domains of disinformation and debunk links.*

Disinformation Domains	Debunk Domains
facebook.com (30%)	dpa-factchecking.com(25%)
tiktok.com (3%)	euvsdisinfo.eu (25%)
twitter.com (3%)	rumorscanner.com (9%)
oroszhirek.hu (2%)	politifact.com (8%)
sputniknews.com (2%)	factly.in (8%)
nabd.com (2%)	verify-sy.com (6%)
arabic.rt.com (1%)	factcrescendo.com (4%)
fb.watch (1%)	verafiles.org (3%)
de.news-front.info (1%)	factcheck.org (2%)
Other (55%)	Other (10%)

if Ukrainian citizens are able to discern between factual information and pro-Kremlin disinformation. Another study (Gerber and Zavisca, 2016) investigated the effectiveness of Russian propaganda in swaying the views of its readers. Recently, Park et al. (2022) released a Ukraine-related dataset of tweets and carried out an analysis of public reactions to tweets by state-affiliated and independent media. Miller et al. (2022) studied the spread of tweets related to hashtags that were trending in February 2022. Nonetheless, these studies do not focus specifically on comparing Ukraine-related disinformation and debunks in terms of engagement, inter-relationship, and topics.

Prior literature on the spread of true and false information is extensive (Vosoughi et al., 2018; Grinberg et al., 2019; Shao et al., 2018). Nevertheless, this paper is related to prior work that studied the spread and dynamics of false information and debunks on Twitter (Burel et al., 2020, 2021; Park et al., 2021; Allcott and Gentzkow, 2017; Singh et al., 2021a; Jiang et al., 2021; Swire et al., 2017; Nyhan and Reifler, 2015; Barrera et al., 2020; Zhang et al., 2022b; Recuero et al., 2022). In particular, Burel et al. (2020) compared COVID-related misinformation and fact-checks using impulse response modelling, causal analysis, and spread variance analysis, while Chen et al. (2021) investigated the reasons why people share fact-checks and ways to encourage this further. Also, Siwakoti et al. (2021) showed that user engagement with fact-checks increased significantly as a result of the COVID-19 pandemic. However, to the best of our knowledge, no study has examined the predictive causality between Ukraine-related disinformation and debunks, or their spatiotemporal characteristics and top disinformation themes.

3.3 Data

The data underpinning our analyses spans disinformation and debunks posted between 1 February and 30 April 2022. Specifically, we focus on Ukraine-related debunks and accompanying links to the corresponding disinformation encompassing: *(i)* 110 debunks and 311 links to disinformation published by EUvsDsinfo[5], which primarily fact-checks pro-Kremlin disinformation; *(ii)* 344 debunks indexed by Google in the ClaimReview format [6], which refer to 439 disinformation links. See Appendix 3.9.1 for details on how we collect the disinformation links from debunks. Similar to Burel et al. (2020), in addition to the above date restrictions, we also applied keyword-based filtering[7] in order to select only Ukraine-related debunks and disinformation.

In total, this study analyses 454 debunk URLs and 750 links to Ukraine-related disinformation. The latter are provided by the fact-checking organisations themselves within the published debunks (see Appendix 3.9.1), therefore we consider them as accurate. Table 3.1 shows the top domains that occur within the disinformation and debunk links. The former point either to content on social media platforms or to Kremlin-backed outlets. For debunks, the main domains are EUvsDisinfo (25%) and Dpa-factchecking (25%).

Next, we use academic research access to the Twitter API[8] to obtain 16,549 unique tweets containing one of the above debunk URLs and another 62,882 unique tweets sharing one of the disinformation links[9]. Retweets are also collected, since we aim to investigate the overall spread of information on Twitter. Hereafter, the tweets containing debunk links are referred to as "debunk tweets" and those containing disinformation links as "disinformation tweets". Figure 3.1 shows the stacked plot of a rolling 7-day average curve for the spread of disinformation and debunk tweets. It shows that Ukraine-related disinformation spiked in the first half of March 2022, which consequently led to an increase in published debunks as it is also reported in EDMO's Fact-checking Briefs [10].

3.4 Comparative Analysis of Engagement

In order to measure the spread of disinformation and debunks through tweets, we first compare the differences in engagement metrics in terms of mean and standard deviation. Table 3.2 shows the statistics for author's followers, author's

[5] https://euvsdisinfo.eu/

[6] https://www.datacommons.org/factcheck/download

[7] Where debunked claims were in languages other than English, these were translated automatically with Google Translate first, prior to filtering with the keywords listed here: https://gist.github.com/greenwoodma/430d9443920a589b6802070f2ca54134

[8] https://developer.twitter.com/en/docs/twitter-api

[9] The dataset used for analysis received ethical approval from the University of Sheffield Ethics Board. This paper only discusses analysis and results in aggregate data, without providing examples or information about individual users.

[10] https://edmo.eu/fact-checking-briefs

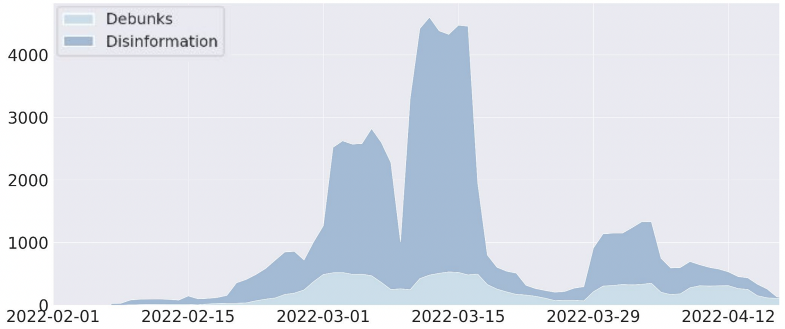


Figure 3.1: *Stacked plot of a rolling 7-day average of the number of disinformation and debunk tweets between 1 February 2022 and 12 April 2022.*

Table 3.2: *Mean and standard deviation (STD) values of metrics of engagement with disinformation and debunk tweets. * represents a statistical significant difference (p≤0.01)*

	Followers	Tweets	Retweets	Replies	Likes	Quote count
Mean - Disinformation	6,814	61,331	15.0*	0.5	4.5	0.2
Mean - Debunks	21,790*	89,098*	1.8	0.4	2.0	0.1
STD - Disinformation	2,04,422	1,22,527	312.3	34.5	631.9	13.9
STD - Debunks	4,55,884	1,55,183	15.0	7.6	28.0	1.3

tweets, number of retweets, replies, likes and the quote count. We find that the number of retweets, replies, likes and the quote count are comparatively higher for disinformation. However, a *t*-test reveals statistically significant difference ($p \leq 0.01$) only for the number of retweets, i.e. significantly more Twitter users are retweeting posts containing disinformation URLs than debunk ones. There is also a statistically significant difference ($p \leq 0.01$) in the number of followers and tweet counts for users sharing debunks as opposed to disinformation. This is as expected since the former are primarily Twitter accounts of fact-checking organisations which naturally have more followers and post more frequently.

Figure 3.2 shows the histogram and kernel density estimate depicting the average number of days between the date of publication of disinformation tweets and their corresponding debunk articles. In this, for each debunk we compute $\sum_{i=1}^{|N|}(DoP_i - DoP_{debunk})/|N|$, where DoP_{debunk} is the date of publication of a debunk by the fact-checking organisation, DoP_i is the date of publication of a disinformation tweet i and $|N|$ is the total count of disinformation tweets for each debunk. The data is positively skewed, with a Fisher-Pearson coefficient of skewness of 3.37, suggesting some spread of disinformation even after the publication of the corresponding debunk article (see Section 3.5).

Since EUvsDsinfo debunks explicitly list countries where the disinformation is spreading, these can be compared to the country of the authors of those tweets. The latter is derived from the self-declared user location field obtained via the Twitter API[11] (when available). We obtain the location information for authors of 51% of the disinformation-sharing tweets. Unsurprisingly, the biggest proportion (9%) comes from cases where EUvsDisinfo has found the disinformation spreading in Ukraine, while the author's self-declared locations are in Russia (Table 3.3). Another key observation is the global nature of the disinformation, with spread extending significantly beyond Europe.

Figure 3.3 shows the most frequent 100 hashtags in disinformation-sharing vs debunk-sharing tweets. Unsurprisingly Ukraine dominates both, while `#FoxNews` is prevalent in tweets sharing disinformation links. This is due to the spread by right-wing American media of a wide-reaching false narrative regarding the presence of U.S.-backed bioweapon labs in Ukraine[12].

We also investigate the presence of identical or highly similar false claims in our dataset that have been debunked multiple times by different fact-checkers. Sim-

[11] https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/user. Where needed, Geopy Python library (Ref. https://pypi.org/project/geopy/) is used to extract the country name from the information provided by the API.

[12] https://www.politifact.com/article/2022/mar/11/russia-china-and-tucker-carlson-lack-evidence-ukra/

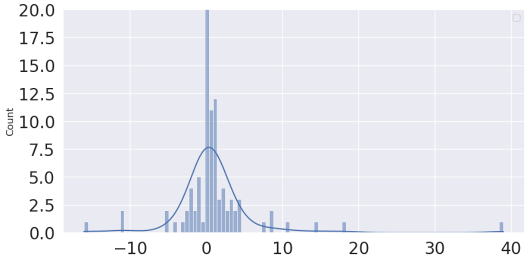


Figure 3.2: *Average difference in days between the date of publication of disinformation tweets and their corresponding debunk article (Fisher-Pearson coefficient of skewness = 3.37).*

Table 3.3: *Top ten cases with country affected by the disinformation and country of the authors of disinformation tweets.*

Affected country	Authors' country	Percentage
Ukraine	Russia	9.0
Ukraine	Germany	7.0
Russia	Russia	6.0
Russia	Germany	5.0
Ukraine	United States	4.0
Ukraine	Venezuela	4.0
Ukraine	Mexico	3.0
United States	Mexico	3.0
Other		59.0

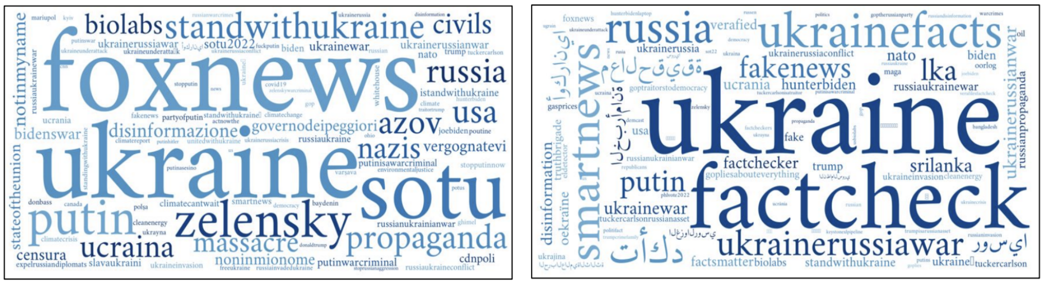


Figure 3.3: *Wordcloud of the most frequent 100 hashtags in disinformation- (left) and debunk-sharing (right) tweets respectively.*

ilar to Singh et al. (2021a), a state-of-the-art semantic search model [13] is used for this task. Out of the 456 debunks in our dataset (see Section 3.3), 84 of them

[13] The multilingual model available at https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2, since it performs best according to the leaderboard (Ref. https://www.sbert.net/docs/pretrained_models.html).

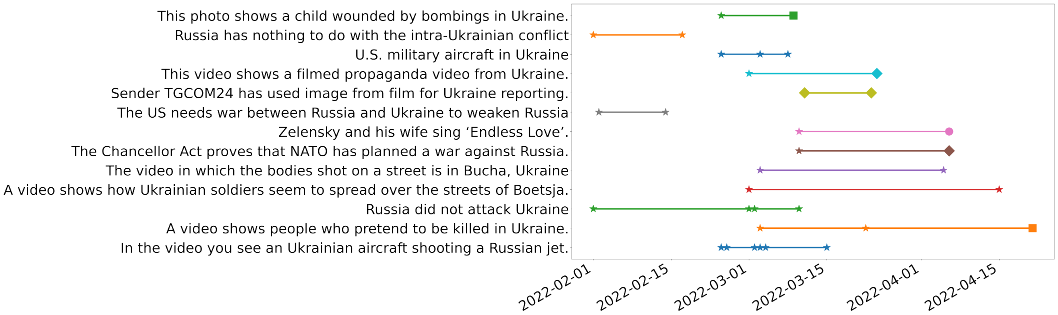


Figure 3.4: *Timeline for a sample of Ukraine-related false narratives that have been debunked multiple times. The y-axis states the false narrative and the x-axis represents the date of publication of its debunks. The language of debunk articles is denoted by different symbols – English: ★; French: ■; Dutch: ●; German: ♦*

(18%) were found to be highly similar to false narratives that have already been debunked by another fact-checking organisation. Figure 3.4 shows some examples of Ukraine-related false narratives that have been debunked multiple times in different languages. This finding demonstrates significant overlap in effort spent by fact-checking organisations in multiple countries, as well as cost- and time-saving opportunities that could be exploited with the help of translation and cross-publishing of debunks.

3.5 Post-hoc Causality Analysis

We test the bi-directional Granger causality (Granger, 1969) between the disinformation-sharing and debunk-sharing tweets. In other words, we want to investigate whether the spread of debunks has a positive impact on reducing the sharing of Ukraine-related disinformation on Twitter. Although identifying causation relationships between different information types is not trivial, a Granger causality test can be used to evaluate the predictive causality i.e. if the spread of one information type can be used to predict the spread of another. In this, we treat the occurrence of disinformation and debunk tweets as two time series variables and then try to find if one variable can be predicted from the other variable's past values and its own past values. [14] First we build a Vector Autoregression (VAR) model (Sims, 1980), where a period of three is applied, based on the Akaike's Information Criterion. The Augmented-Dicky Fuller test identifies the data as stationary ($p \leq 0.01$). The general equation of VAR model is

[14] The Statsmodel Python library is used to perform the Granger causality test. Ref. https://www.statsmodels.org/

$$disinfo(t) = \sum_{i=1}^{k} \alpha_{1,i} disinfo(t-i) + \sum_{i=1}^{k} \beta_{1,i} debunk(t-i) + \epsilon_1 \tag{3.1}$$

$$debunk(t) = \sum_{i=1}^{k} \alpha_{2,i} debunk(t-i) + \sum_{i=1}^{k} \beta_{2,i} disinfo(t-i) + \epsilon_2 \tag{3.2}$$

where $debunk(t)$ and $disinfo(t)$ refers to count of tweets at time t, k is the maximum lag order, $\alpha_{1,i}$ and $\alpha_{2,i}$ are autoregressive coefficients, $\beta_{1,i}$ and $\beta_{2,i}$ are regression coefficients, and ϵ_1 and ϵ_2 are error terms.

The experiments find a Granger causality relation, which shows that debunk spread has predictive causality over disinformation spread ($p \leq 0.01$). In addition, we also observe this weak causation in the opposite direction, i.e. from disinformation to debunks ($p \leq 0.01$). The significant results in both directions imply that changes in the spread of disinformation may induce changes in the spread of debunks and that the spread of debunks may likewise cause changes in the disinformation spread. This is similar to the findings of previous work for COVID-19 misinformation (Burel et al., 2020, 2021). In order to further understand the weak causation between Ukraine-related disinformation and debunks, we use the VAR model to perform an impulse response analysis and forecast the error variance decomposition for 14 days periods.

Impulse response analysis is used to find the effect of shock in one variable to itself and the other variables in the VAR model. The prime reason to investigate this is to check if an increase in the spread of debunks triggers a reduction in disinformation on Twitter. Figure 3.5 shows that an orthogonal shock[15] from debunks leads to an initial spike in disinformation but there is a downward trend for disinformation afterwards. This suggests that debunks will trigger a reduction in overall disinformation eventually, if not instantly. Similarly, an orthogonal shock from disinformation also triggers an initial spike in debunks (Figure 3.5) and an eventual decrease with time, although not instantly. This suggests swift response in debunk publication (mostly from fact-checkers) following a sudden rise in disinformation on social media. Interestingly, we also notice that the shock in disinformation quickly dies as the impact returns to zero with a sharp decrease on the second day, followed by a small post-shock peak during the 4–6 day period which finally converges back to zero between 8–10 day period.

Forecast Error Variance Decomposition (FEVD) helps uncover the proportion of information each variable contributes in predicting a particular variable in the VAR model. FEVD analysis (Figure 3.6) reveals substantial predictive dependencies between Ukraine-related disinformation and debunks. Similar to what we

[15] Cholesky decomposition is used for orthogonalisation

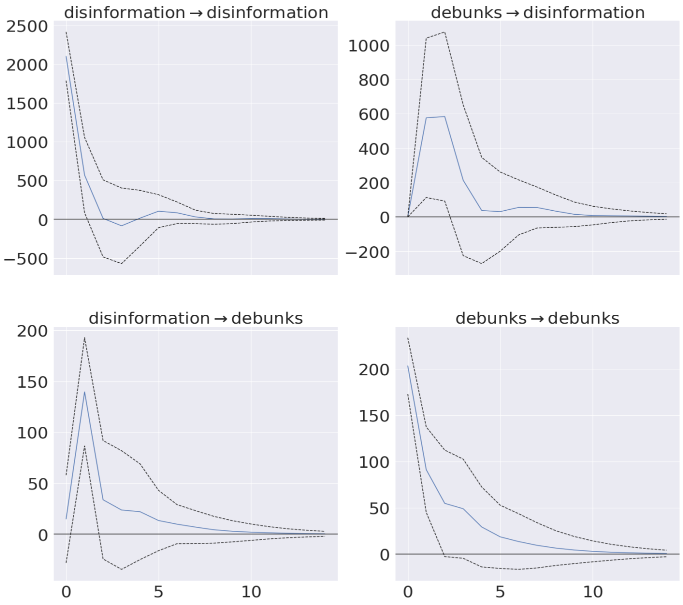


Figure 3.5: *Impulse Response Analysis (x-axis represents 14 days period and y-axis represents effect of shock). Top left and bottom left shows the effect of disinformation shock on disinformation and debunks respectively. Top right and bottom right shows the effect of debunk shock on disinformation and debunks respectively. By default, asymptotic standard errors are presented at the 95% confidence level.*

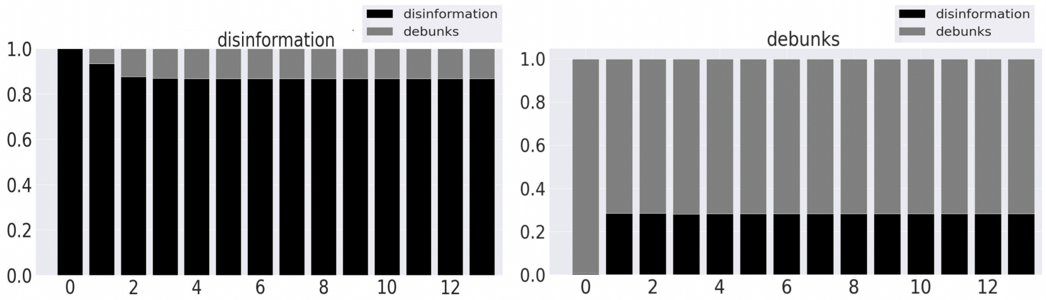


Figure 3.6: *Forecast Error Variance Decomposition (FEVD) plot (x-axis represents a 14-day period and y-axis represents proportion of affect). Left and right represents FEVD for disinformation and debunks respectively.*

find in impulse response analysis, FEVD results show that debunks directly affect disinformation by around 15% by the end of the 14-day period. We also observe that debunks affect the spread of disinformation after an initial delay by a day,

Table 3.4: *Top ten words and count of disinformation tweets in each topic cluster. Order of words depicts its importance from left to right.*

Topic clusters	**Count**
0_ukraine_ukrainian_russia_kyiv_neo_coup_nazis_war_crimea_weapons	39,761
1_poland_nato_polish_alliance_west_security_countries_western_europe_russian	10,225
2_putin_biden_know_think_vladimir_lee_answer_prices_says_oil	4,194
3_video_shows_ukraine_ukrainian_proof_jet_marcos_shot_soldiers_fighter	3,793
4_biolabs_ukraine_financed_state_biological_labs_military_biden_victoria_vaccinated	3,132
5_trump_russia_bucha_massacre_billions_evidence_west_100_planted_united	1,777

after which it rises and becomes constant following the third day. On the other hand, we also find that the spread of debunks is also affected by disinformation by around 30%. The results also show that the impact of disinformation on debunks is delayed initially for a day. In other words, this implies that the spread of debunks is not dependent on how Ukraine-related disinformation spreads initially. In summary, our experimental analysis confirms that the spread of debunking tweets does have a positive impact on reducing Ukraine-related disinformation on Twitter.

3.6 Topical Analysis of Ukraine-related Disinformation

This section investigates the main topics in disinformation and study the engagement around them over time. The debunked claim statements are clustered by applying K-means to embeddings from the semantic search model[16]. The number of clusters is kept at six using the Elbow method and silhouette coefficient score. The model is run for a maximum of 300 iterations with K-means++ used as a method of initialisation. The clustering is applied on debunked claim statements and not on tweets itself. See Appendix 3.9.1 for details on how we collect the debunked claim statements.

The class-based TF-IDF (Grootendorst, 2022) is used to find top words in debunked claim statements in each of the clusters (Table 3.4). Each cluster has distinct words that separate it from the other five. In order to verify the separation between the clusters, we also plot a heatmap of topic similarity (see Appendix 3.9.2). The results show that except clusters zero and one, most of the clusters are distinct in terms of the topics they cover. For instance, cluster four encompasses

[16] We use the BERTTopic (Grootendorst, 2022) Python library for clustering and MPNet (Song et al., 2020) as the transformer model. Ref. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

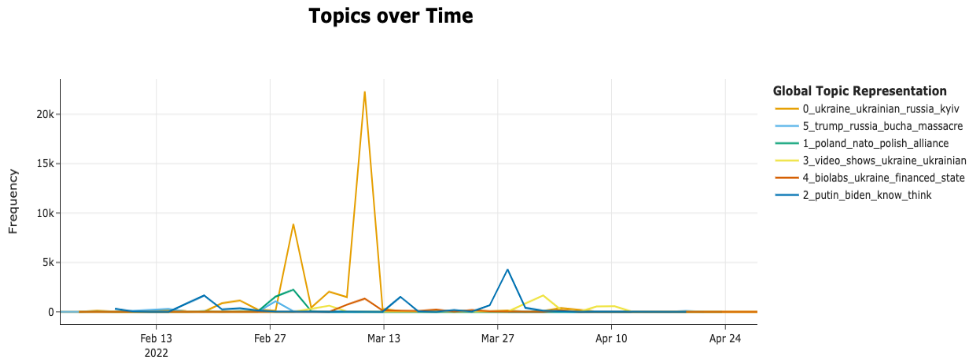


Figure 3.7: *Temporal spread of disinformation tweets in each topic cluster over time. Legend shows top five words of each cluster from Table 3.4.*

wide-spread conspiracies related to the U.S.-backed bioweapon labs in Ukraine and the US planning to send infected migratory birds to infect Russia[17]. Another cluster (one), includes the ongoing false narrative about the NATO country alliance being the real threat to Russia[18]. Table 3.4 also shows the count of corresponding disinformation tweets (and retweets) in each cluster, identifying that most of the tweets belong to cluster zero and one.

Next, we look at the temporal distribution of the disinformation tweets for each topic cluster. Figure 3.7 illustrates the line plot for topic prevalence between February and April 2022. For instance, cluster zero includes false claims related to Russia attacking Ukraine, Kyiv, neo-nazism, etc. and has two dominant peaks, one in the first week of March and another one in the second week of March (the tallest one at 10 March 2022). There is also an uptick in February suggesting that the disinformation narratives started spreading even before 24 February and then spiked later in March. Similar results were found in the EDMO's Fact-checking Briefs for February[19] where they noticed a sudden increase in posts about growing tensions between Russia and Ukraine.

Cluster one (disinformation related to NATO and western countries) spiked in the first week of March. It includes a dominant false narrative about NATO attacking countries illegitimately[20].

[17] https://euvsdisinfo.eu/report/the-us-plans-to-send-infected-migratory-birds-to-infect-russia

[18] https://www.politifact.com/factchecks/2022/feb/28/candace-owens/fact-checking-claims-nato-us-broke-agreement-again/

[19] https://edmo.eu/fact-checking-briefs

[20] https://euvsdisinfo.eu/report/nato-is-not-a-defensive-alliance-it-attac

Cluster two has multiple spikes: one in mid February; another one in mid of March; and the biggest one – at the end of March. It comprises false narratives involving Joe Biden, Vladimir Putin and Russian oil, e.g. that Biden's cancellation of the Keystone pipeline "dramatically increased Americans" dependence on Russian oil.

Cluster three comprises of videos spreading disinformation and their distribution is fairly stable, with only a slight increase in the first and last week of March.

Cluster four contains conspiracies, such as an alleged presence of US biological labs in Ukraine[21] and release of infected migratory birds to infect Russia[22]. Figure 3.7 shows that these type of conspiratorial narratives spiked during the second week of March. This is also coherent with the findings of EDMO's Fact-checking Briefs for March 2022.

Cluster five contains disinformation related to Trump and the Bucha massacre and has comparatively time-limited span, with only a small peak at the end of February 2022.

3.7 Limitations and Future Work

Our work should be seen in the light of the following limitations. First, as described in Section 3.3, the study uses only tweets which contain explicit links to known disinformation or debunk articles. While this makes the dataset highly accurate and does not require additional human annotation, it also means that tweets that spread false claims or debunk them without citing a reference link could not be included. Second, this paper only discusses results on aggregate data, without looking at whether the tweets are from real or bot accounts. Lastly, user data, such as their country, is dependent on self-declared information in the user profiles, which is missing for many tweets. Nevertheless, the sample size is sufficiently large and robust to yield useful insights.

In the future, we want to analyse the spread of disinformation and debunks before and after the start of the Russia-Ukraine war. We might have different answers for the research questions raised in the paper, which would potentially provide some insights into how an emergency event changes the spreading paradigms of disinformation and debunks. We also want to find ways to automatically detect disinformation tweets which don't explicitly mention links or where the disinformation links mentioned are different from the ones present in our dataset. Lastly,

ks-countries-illegitimately

[21] https://www.bbc.co.uk/news/60711705

[22] https://euvsdisinfo.eu/report/the-us-plans-to-send-infected-migratory-birds-to-infect-russia

the Granger causality test deals with linear relationship. Hence, in future, we plan to experiment with other non-linear tests like Hiemstra and Jones non-linear Granger causality (Hiemstra and Jones, 1994) and Convergent Cross Mapping test (Tsonis et al., 2018).

3.8 Conclusion

This study carried out a comparative analysis of the spread of Ukraine-related false claims and debunks on Twitter between February and April 2022. In particular, our comparative engagement analysis found that tweets spreading disinformation are shared and retweeted significantly more as compared to those containing debunks. With respect to debunks, we also established that around 18% are focused on false claims for which debunks have already been posted in a different country or language. This finding is particularly important, as it points out two opportunities going forward. Firstly, since many platforms, such as Facebook, already offer machine translation tools to their users, they could use that technology themselves to translate and match debunks automatically, so a false narrative spreading in one language can be flagged as false, based on an authoritative fact-check in another language. Secondly, fact-checkers themselves can benefit from using cross-lingual search and machine translation technologies to find such debunks, which they can then cite as a source or re-publish in translation and thus reduce the time elapsed between a false narrative starting to spread widely online and the time their debunk is published.

Another key finding is that the publication of debunks does ultimately lead to limiting the spread of Ukraine-related disinformation, albeit not immediately. In addition, FEVD results show substantial predictive dependencies between the spread of disinformation and debunk tweets. Lastly, our data-driven analysis uncovered also the dominant themes in Ukraine-related disinformation and their temporal intensity. In conclusion, these findings have immediate relevance for a wide range of stakeholders, including digital platforms, fact-checkers, and online information users. The dataset used for analysis is available at https://doi.org/10.5281/zenodo.6992686.

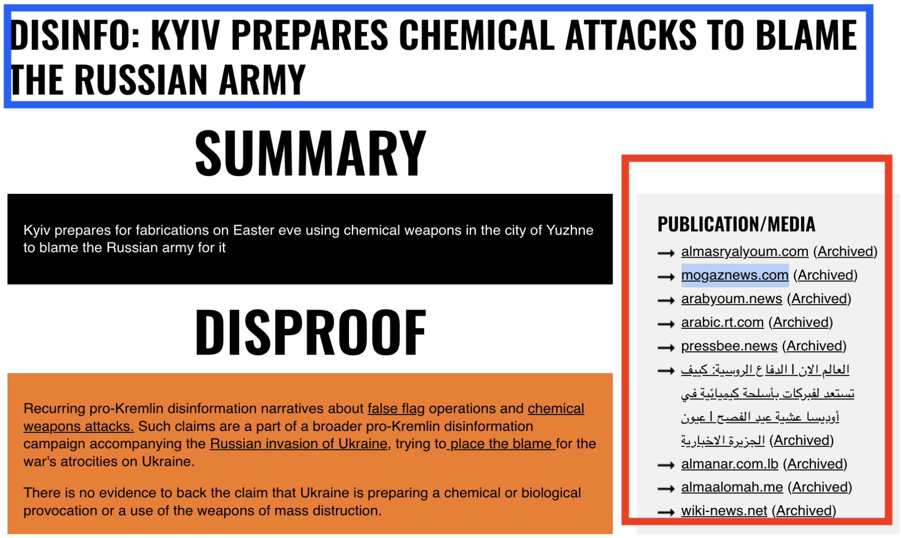
3.8.0.1 Acknowledgements.

This research has been partially supported by a European Union – Horizon 2020 Program, grant no. 825297 (WeVerify), the European Union – Horizon 2020 Program under the scheme “INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities” and Grant Agreement n.871042 (“SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” (http://www.sobigdata.eu)).

3.9 Appendix

3.9.1 Data Collection

As described in Section 3.3, we collect Ukraine-related debunks from EUvsDsinfo and ClaimReview. In order to collect the disinformation links, 1) the debunks indexed in ClaimReview schema has the `itemReviewed`[23] object which includes disinformation links that are being debunked by fact-checking organisation and debunked claim statement is present in `claimReviewed` object; 2) the debunks on EUvsDsinfo explicitly mention disinformation links on their website. Figure 3.8 shows the screenshot of one of the EUvsDsinfo debunks. The section enclosed in the red box contains disinformation links and the blue box represents the debunked claim statement.



DISINFO: KYIV PREPARES CHEMICAL ATTACKS TO BLAME THE RUSSIAN ARMY

SUMMARY

Kyiv prepares for fabrications on Easter eve using chemical weapons in the city of Yuzhne to blame the Russian army for it

DISPROOF

Recurring pro-Kremlin disinformation narratives about false flag operations and chemical weapons attacks. Such claims are a part of a broader pro-Kremlin disinformation campaign accompanying the Russian invasion of Ukraine, trying to place the blame for the war's atrocities on Ukraine.

There is no evidence to back the claim that Ukraine is preparing a chemical or biological provocation or a use of the weapons of mass distruction.

PUBLICATION/MEDIA

→ almasryalyoum.com (Archived)
→ mogaznews.com (Archived)
→ arabyoum.news (Archived)
→ arabic.rt.com (Archived)
→ pressbee.news (Archived)
→ العالم الان | الدفاع الروسية: كييف تستعد لفبركات بأسلحة كيميائية في أوديسا عشية عيد الفصح | عيون الجزيرة الاخبارية (Archived)
→ almanar.com.lb (Archived)
→ almaalomah.me (Archived)
→ wiki-news.net (Archived)

Figure 3.8: *Screenshot of one of the EUvsDsinfo debunks. Section enclosed in the red box contains disinformation links and the blue box represents the debunked claim statement.*

3.9.2 Heatmap

Figure 3.9 illustrates the heatmap of cluster similarity. The results show that except clusters one and two, most of the clusters are distinct in terms of the topics they cover. This indicates reasonable separation between the clusters found in Section 3.6.

[23] https://schema.org/ClaimReview

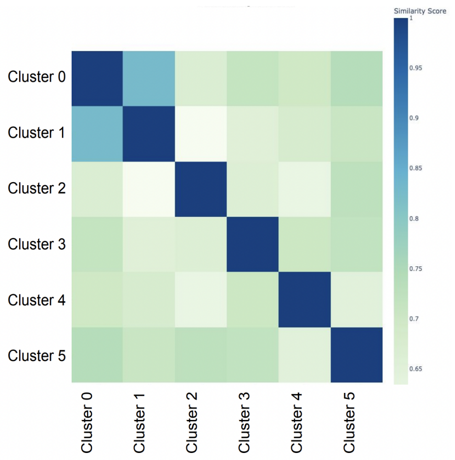


Figure 3.9: *Heatmap for topic cluster similarity. The description of clusters can be found in Section 3.6.*

Chapter 4

Multistage BiCross encoder for multilingual access to COVID-19 health information

Iknoor Singh, Carolina Scarton and Kalina Bontcheva
Department of Computer Science, The University of Sheffield, UK

Abstract

The Coronavirus (COVID-19) pandemic has led to a rapidly growing 'infodemic' of health information online. This has motivated the need for accurate semantic search and retrieval of reliable COVID-19 information across millions of documents, in multiple languages. To address this challenge, this paper proposes a novel high precision and high recall neural Multistage BiCross encoder approach. It is a sequential three-stage ranking pipeline which uses the Okapi BM25 retrieval algorithm and transformer-based bi-encoder and cross-encoder to effectively rank the documents with respect to the given query. We present experimental results from our participation in the Multilingual Information Access (MLIA) shared task on COVID-19 multilingual semantic search. The independently evaluated MLIA results validate our approach and demonstrate that it outperforms other state-of-the-art approaches according to nearly all evaluation metrics in cases of both monolingual and bilingual runs.

4.1 Introduction

The COVID-19 pandemic has, to date, infected more than 135M people worldwide. It has also been accompanied by what the World Health Organisation has dubbed an ‘infodemic’, in reference to the challenge people face in navigating and absorbing the continuously growing volumes of information on the origin, treatment, prevention, and public policies related to COVID-19 that get published online by numerous sources (some authoritative and some not), in multiple languages and countries. This has prompted the need for new efficient and accurate multilingual semantic search and retrieval methods for health information. To facilitate the comparative evaluation of existing approaches and research on new methods, the COVID-19 Multilingual Information Access (MLIA) shared task (Casacuberta et al., 2021) released benchmark datasets of COVID-19 related information spanning all EU languages and beyond. These datasets address three key information access tasks: Information Extraction (Task 1), Multilingual Semantic Search (Task 2), and Machine Translation (Task 3). Here we focus specifically on the multilingual semantic search task, which provides a large multilingual dataset to support research and a comparative evaluation of semantic search approaches (with the purpose of analysing and improving the retrieval of relevant COVID-19 documents from a given user query).

In the past few years, the large pre-trained transformer models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLM (Conneau et al., 2020; Liu et al., 2019) and others have achieved state-of-the-art performance on a wide range of natural language processing tasks from semantic text similarity to question answering. Recently, these have also been applied to information retrieval tasks (Akkalyoncu Yilmaz et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021; Nogueira and Cho, 2019). In this paper, we present a novel multistage BiCross encoder method and demonstrate that it outperforms other state-of-the-art retrieval methods for COVID-19 multilingual semantic search, according to independent comparative evaluation on the MLIA shared task 2 dataset.

Multistage BiCross encoder is a sequential three-stage ranking pipeline, composed of: 1) a BM25 retrieval stage 2) a neural refinement stage, and 3) a neural reranking stage. Our approach exploits both bi-encoder and cross-encoder transformer architectures to compute the document-level relevance score by aggregating sentence-level relevance scores. For document retrieval, cross-encoders tend to attain significantly higher accuracy (Nogueira and Cho, 2019), due to the rich interactions and self-attention over the query and document pair. On the other hand, when the number of documents to be re-ranked is large, cross-encoder recomputes encoding each time during inference (Vaswani et al., 2017) which makes them very resource-intensive when compared to bi-encoder which can make use of cached representations for faster inference. Since the performance gains come at a steep computational cost, we use both bi-encoder and cross-encoder with the former having more documents to re-rank as compared to the latter. This way we are able to utilise the benefits of both bi-encoder and cross-encoder neural models on top of the BM25 lexical model. To the best of our knowledge, ours is the first paper to investigate this for document retrieval. The main research question addressed in this paper is: how to improve the architecture and performance of state-of-the-art neural models for document retrieval to make them better suited to retrieving COVID-19 health information in multiple languages?

The key contributions of this paper are:

- Multistage BiCross encoder method, which is a three-stage ranking pipeline that uses the Okapi BM25 retrieval algorithm and state-of-the-art multilingual transformer-based bi-encoder and cross-encoder by aggregating sentence-level relevance scores for the task of COVID-19 multilingual semantic search.
- We experiment with different types of search queries in order to establish the best performing ones for retrieving COVID-19 health information across millions of documents, in multiple languages. We also present ways to combine scores from different stages using various rank fusion algorithms.
- An extensive comparison of our runs with other participant runs to demonstrate the effectiveness of our methods in achieving high precision for top ranked documents, as well as high recall for all retrieved documents in both monolingual and cross-lingual search settings.

Prior to introducing the proposed approach (Section 4.4), we first introduce the multilingual COVID-19 semantic search task and the accompanying dataset provided by the organisers of the MLIA shared task evaluation challenge (Section 4.2). Next, Section 4.3 discusses previous work on neural methods for semantic search. Section 4.4 gives a detailed description of the underlying architecture of our Multistage BiCross Encoder and our training methodology. In addition, it also describes the use of external training datasets which improved the model's performance further and helped it achieve the best reported scores on the MLIA COVID-19 semantic search task, according to the independent comparative evaluation reports by the shared task organisers[1] (Di Nunzio et al., 2021). Section 4.5 provides details of all our experimental runs and settings, as well as information on the other participating systems in MLIA. The evaluation results are presented in Section 4.6, followed by a conclusion in Section 4.7.

[1] https://bitbucket.org/covid19-mlia/organizers-task2/src/master/

4.2 MLIA COVID-19 semantic search task

Multilingual Information Access (MLIA) COVID-19 is a shared evaluation task run by an independent consortium of researchers and is endorsed by the European Commission and the European Language Resource Coordination (ELRC). In this section, we describe the COVID-19 MLIA dataset and introduce the specifics of the multilingual semantic search task 2. The core challenge of this task is to improve information exchange about COVID-19 in both monolingual and cross-lingual search settings.

4.2.1 Dataset

The dataset consists of a corpus of 3,750,588 documents and a set of 30 query topics, both available in multiple languages: English, French, German, Greek, Italian, Spanish, Swedish and Ukrainian. These languages are spoken in countries where there was a rapid spread of COVID-19 or the pandemic was managed differently at the beginning of 2020. For each language, the corpus consists of general health-related articles collected from different websites, out of which the majority of articles come from the Medical Information System (MEDISYS)[2]. Table 4.1 shows the number of documents in each language.

Table 4.1: *Count of documents for each language in the MLIA corpus.*

Language	Count
English (en)	1,452,240
Spanish (es)	833,763
Italian (it)	662,789
French (fr)	326,599
German (de)	273,761
Greek (el)	147,658
Swedish (sv)	38,196
Ukrainian (uk)	15,582
Total	3,750,588

There are a total of 30 query topics that are used for querying the dataset. Each query topic comprises of three fields: (i) a keyword field: a set of relevant keywords related to the query; (ii) a conversational field: the query in the form of a question; and (iii) an explanation field: a more detailed description of information that is needed in the retrieved documents. For our experiments, we only used the keyword and conversational fields as the explanation field is more useful

[2] https://medisys.newsbrief.eu/medisys/clusteredition/en/24hrs.html

for assessing the relevance of the document at evaluation time. Table 4.2 shows the keyword, conversational, and explanation fields for one of the English query topics on the use of ultraviolet light to kill coronavirus.

Table 4.2: *The keyword, conversational, and explanation fields for the MLIA query topic about the use of ultraviolet light to kill coronavirus.*

Topic Field	**Text**
Keyword	uv light to kill coronavirus
Conversational	Is uv light effective to kill coronavirus?
Explanation	Seeking studies that discuss whether ultraviolet light is an effective way to sanitise against COVID-19

4.2.2 Task description

In the MLIA multilingual semantic search (task 2) (Di Nunzio et al., 2021), participating systems need to search the growing information related to the novel coronavirus, in different languages and with different levels of knowledge about a specific topic. The task follows a CLEF-style (Peters, 2000) evaluation methodology where participants are provided with a collection of documents and a set of topics that are used as queries to produce various runs that could either be monolingual or bilingual, depending on the language of the query and the retrieved documents. There are two subtasks: subtask 1 is focused on high precision whilst subtask 2 is oriented towards high-recall systems. Each participating team could submit a maximum of five monolingual and five bilingual runs for each language for each subtask. In monolingual runs, the language of both query and documents is the same whereas for bilingual runs, the language of both query and documents is different.

In order to carry out a comparative evaluation of the participating systems on unseen data, the organisers created an additional pool of around 6000 to 8000 documents for each language by selecting the top k documents from all the runs. These pools of documents were then manually annotated by the experts to produce relevance judgements. The organisers then evaluated all runs using established information retrieval metrics, including recall, precision (P@5 & P@10), R-precision (RPrec), average precision (AP), and normalised discounted cumulative gain (NDCG).

4.3 Related work

Neural models for information retrieval are generally used in a two-stage pipeline architecture where re-ranking is done only on the top results retrieved by traditional ranking methods such TF-IDF or BM25 (Akkalyoncu Yilmaz et al., 2019; Nogueira and Cho, 2019; Karpukhin et al., 2020), as the computational cost of running neural models over the entire dataset can be prohibitively high (Hofstätter and Hanbury, 2019). BM25 (Jones et al., 2000) is a bag-of-words retrieval model that retrieves documents based on lexical overlap with the query terms. Nogueira and Cho (2019) are the first to demonstrate that the BERT model can also be used for fine-tuning passage re-ranking tasks and it has shown to be effective for ad-hoc document ranking. They use a sequence of tokens by concatenating the query tokens and the passage tokens, separated by a [SEP] token as an input to the BERT model and then the output embedding of the [CLS] token is passed to a single layer neural network to obtain the probability of the passage being relevant to the query. On the other hand, Karpukhin et al. (2020) use dual-encoder (or bi-encoder) architecture to apply FAISS (Johnson et al., 2017) on the encoded BERT representations of questions and passages. Although FAISS makes it fast, due to separate representations, it lacks attention between the question and passage tokens which makes it less accurate (Nogueira and Cho, 2019).

In addition, researchers tried various methods to improve the effectiveness of neural models in document re-ranking tasks. Nogueira et al. (2019a) use a query generator model to expand the document before indexing to get additional gains on the retrieval performance. Other work uses rank fusion methods (Fox and Shaw, 1994; Cormack et al., 2009) to combine various runs in order to improve the performance of retrieval systems. Clipa and Di Nunzio (Clipa and Di Nunzio, 2020) analyse and compare various state-of-the-art information retrieval methods and ranking fusion approaches for the domain of medical publication retrieval. Pradeep et al. (2021) formulate it as a pointwise and pairwise classification problem using the sequence-to-sequence T5 model (Raffel et al., 2020a) to show its effectiveness in neural re-ranking. Akkalyoncu Yilmaz et al. (2019) use sentence-level evidence to compute document-level relevance scores using a cross-encoder model for ad hoc retrieval. Previous study (Zhang et al., 2020b) also shows that the highest scoring sentence in a document is a good indicator of the relevance of the document and it helps in achieving high recall. But applying sentence-level inference on all the documents is also not possible given the computational overhead of transformer models. As far as sentence-pair scoring tasks are concerned, due to the transformer attention mechanism (Vaswani et al., 2017; Humeau et al., 2019), cross-encoder are more accurate than bi-encoder. On the other hand, cross-encoders recompute encoding each time during inference as compared to bi-encoder which can make use of cached representations for faster inference. Hence, instead of applying a cross-encoder model on each sentence during inference, we use a bi-encoder to compute the document-level relevance score by aggregating sentence-level relevance scores using cached sentence representations. This is done in the neural refinement stage (Section 4.4.2) which succeeds BM25 retrieval stage. In the last stage (Section 4.4.3), a cross-encoder architecture is used which exploits the self-attention of the transformer model to re-rank a subset of candidate documents so as to make relevant documents rank higher. This way we can take advantage of both bi-encoder and cross-encoder for finding semantically relevant documents from the initially retrieved documents from BM25 lexical model.

Furthermore, to produce good sentence representations from transformer models, Reimers and Gurevych (2019) propose SBERT, where they train BERT-based models using siamese network architecture to get semantically meaningful sentence representations. These can be leveraged for other tasks such as semantic search where the sentence representations can be compared using the sentence-pair scoring function. We used this training methodology to train the bi-encoder in the neural refinement stage to bring both the query and the relevant document into the proximity of each other in the high dimensional vector space (Section 4.4.2).

The outbreak of COVID-19 has led to ever-expanding research in deep learning for COVID-19 (Iwendi et al., 2021) and healthcare (Bhattacharya et al., 2021). Recently, the TREC-COVID challenge (Roberts et al., 2020) invited participants to develop information retrieval systems for scientific literature containing tens of thousands of scholarly articles related to COVID-19. Although, both the TREC-COVID challenge and MLIA task 2 involve the development of information retrieval systems for COVID-19 information, the domain of the corpora in both tasks is different: the former is targeted towards scientific scholarly papers whilst the latter is targeted towards systems that provide general health-related articles of relevance to citizens and public health. In addition, the TREC-COVID challenge only had English query and documents whereas MLIA is a multilingual task in which both query and documents are in multiple languages as discussed in Section 4.2.

4.4 Multistage BiCross encoder

Multistage BiCross encoder is a three-stage ranking pipeline that includes an initial lexical retrieval stage followed by two neural-based semantic retrieval stages. Fig 4.1 illustrates the architecture of our approach where q is the query, s_i is the ith sentence of the document, model $M1$ is used as a bi-encoder in neural refinement stage (Section 4.4.2) and model $M2$ is used as cross-encoder in the neural re-ranking stage (Section 4.4.3).

For lexical retrieval, Okapi BM25 (Jones et al., 2000) is used to reduce the search space from a large number of documents (e.g. 1.4M in the case of English docu-

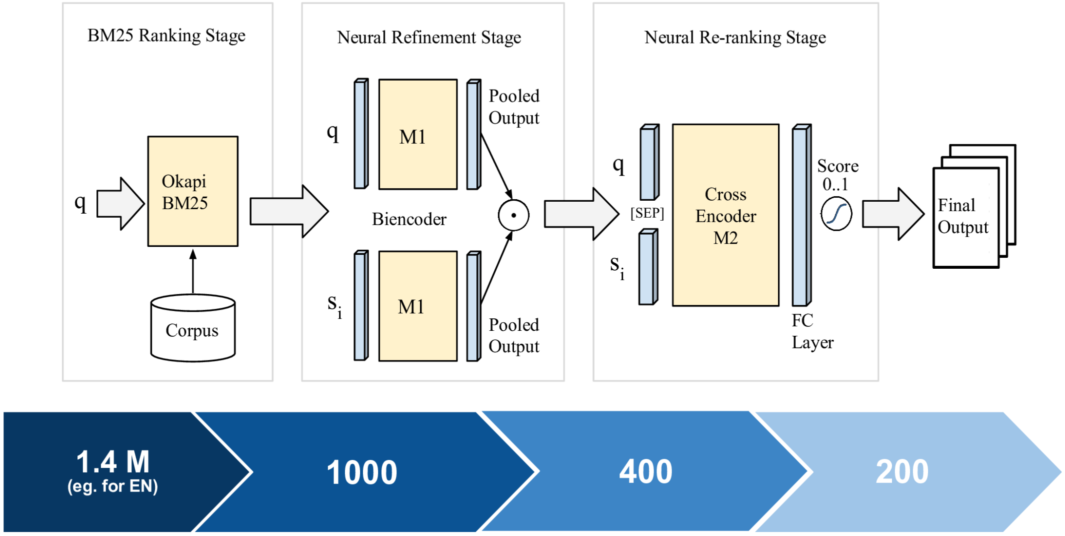


Figure 4.1: *Overview of Multistage BiCross encoder. The blue bar below shows top k candidate documents ranked in each stage.*

ments) to a small set of possibly relevant documents. In the second stage (referred to as the neural refinement stage), we leverage a transformer-based bi-encoder model to encode both query and document individually into deep contextualised representations and use them to efficiently re-rank the retrieved documents based on their relevance. The final neural re-ranking stage uses a transformer-based cross-encoder (Nogueira and Cho, 2019) to re-rank a subset of top-ranked candidate documents output from the neural refinement stage. We also explore various rank fusion techniques to combine the output from the different stages to get a single relevance score which is used for sorting the final list of documents. The blue bar in Fig 4.1 shows top k candidate documents ranked in each stage. The detailed description of all three stages can be found in the subsequent subsections.

4.4.1 BM25 retrieval stage

The initial set of candidate documents are retrieved using the traditional Okapi BM25 (Jones et al., 2000) lexical retrieval model as shown in Fig 4.1. First, we preprocess all documents in the corpus and index them using Elasticsearch[3]. Documents belonging to different languages are indexed separately. In our case, we only considered English, Spanish, French and German documents since working on all languages was not feasible within the very constrained timeline set by the organisers for submitting runs.

[3] https://www.elastic.co/elasticsearch/

The corpus provided by the organisers contains documents in the XML format and we have only used the text inside the <p> tags (all boilerplate tags are removed). Text pre-processing methods such as stopword removal and lemmatisation were applied before indexing the documents. As described in Section 4.2, each query topic has three different fields expressing the information needed in various levels of detail. Our experiments use a concatenation of the keyword and conversational fields (*key_conv*) as a query to retrieve matching documents from the Elasticsearch indexes using BM25.

We also used the keyword and conversational field (*key_conv*) to generate three more queries using sequence-to-sequence T5-base doc2query model (Nogueira et al., 2019a) and we concatenate these three generated queries with the *key_conv* to form a single query, called hereafter *t5_query*. The idea behind the *t5_query* is to assess the performance of concatenated reworded queries on the MLIA corpus. For example, *t5_query* for the topic query mentioned in Table 4.2 is "uv light to kill coronavirus Is uv light effective to kill coronavirus? Does uv light kill coronavirus? Can uv lights kill coronavirus? Is uv light effective against coronavirus?". In addition, we also tried the *Udels* query from TREC-COVID (Zhang et al., 2020a) to evaluate its effectiveness on the MLIA dataset. The *Udels* query (Zhang et al., 2020a) is made up of non-stopwords from the keyword field and the named entities mentioned in the conversational field. For instance, *Udels* query for the topic query mentioned in Table 4.2 is "uv light kill coronavirus effective kill coronavirus".

The BM25 ranking stage will filter out all lexically dissimilar documents concerning the query. The score of a document d using Okapi BM25 algorithm is formulated as

$$\text{BM25Score}(d) = \sum_{i=1}^{n} \text{IDF}\left(q_i\right) \cdot \frac{f\left(q_i, d\right) \cdot \left(k_1 + 1\right)}{f\left(q_i, d\right) + k_1 \cdot \left(1 - b + b \cdot \frac{d}{\text{avgdl}}\right)} \tag{4.1}$$

where $BM25Score(d)$ is the BM25 relevance score of document d, q_i are token keywords of the given query, $IDF(q_i)$ is the inverse document frequency of the query term q_i, $f(q_i, d)$ is the frequency of query term in the document, d is the number of words in the document d, $avgdl$ is the average length of documents in the complete database and the rest are the default parameters ($k1 = 1.2$ and $b = 0.75$) as set in Elasticsearch.

4.4.2 Neural refinement stage

In the second stage, the top 1000 documents retrieved by BM25 are re-ranked using a bi-encoder which is based on Siamese networks (Reimers and Gurevych, 2019). We use a pre-trained transformer-based model to encode both document and query separately into fixed-length deep contextualised embeddings by using mean pooling on the output layer. In the same vector space, query and relevant document lie in proximity to each other and can be efficiently retrieved using cosine similarity as shown in the neural refinement stage of Fig 4.1. As the representations are separate, the encoded representations from query and document are cached so that they can be reused for faster predictions during inference time. Following (Yang et al., 2019; Akkalyoncu Yilmaz et al., 2019), each document in the corpus is split into sentences, and we apply inference on each sentence separately to obtain a sentence-level relevance score for each pair of input query and sentence. As the documents in the MLIA corpus are long, we only consider the first N sentences for inference, where N denotes average number of sentences in documents of the corpus. Moreover, previous research (Li et al., 2018; Hammache and Boughanem, 2020) shows that any relevant document is likely to contain relevant sentences at the beginning of the document. The document-level relevance score is determined by aggregating the top k scoring sentences in the document as follows:

$$BiScore(d) = \sum_{i=1}^{k} w_i \cdot S_{Biencoder_i} \tag{4.2}$$

where BiScore(d) is the document-level relevance score for document d using the bi-encoder model. $k = 3$ as shown in (Yang et al., 2019) and $S_{Biencoder_i}$ is the i-th top sentence-level relevance score with respect to the query. Similar to (Yang et al., 2019), the parameters w_i are tuned via exhaustive grid search. Due to lack of relevance labels for the MLIA task, we have set the initial parameters such that $w_1 > w_2 > w_3$, because we want to give more weight to the sentence which is more relevant as compared to the less relevant sentences. In other words, high scoring sentences contribute more to the final relevance score of the document.

Since there are no relevance labels available for the MLIA task, we generate pseudo-qrels using external datasets specific to the COVID-19 domain which we used to train our models. We prepared the TC+IFCN data which is a combined version of the TREC-COVID challenge (TC) (Roberts et al., 2020) dataset and the IFCN dataset (Song et al., 2021). The TREC-COVID dataset has 69,318 relevance assessments for 50 different topics. Although the corpus used in the TREC-COVID challenge consists of scientific scholarly articles, we use this dataset to transfer knowledge to our model and test its performance on the MLIA corpus which comprises general health-related articles. On the other hand, the IFCN dataset consists of around 7000 COVID-19 misinformation claims debunked by members of the International Fact-Checking Network (IFCN). In this case, we consider each claim as the pseudo-query and its corresponding fact-checked article body as its relevant document. As we apply inference on each sentence separately, we prepare a sentence-level dataset in which, for each query, we extracted sentences from the document that share something meaningful with the query. For this, we use state-of-the-art models trained on the Semantic Textual Similarity (STS) (Cer et al., 2017) data to generate both positive and negative sample sentences from the document and we assume that these are relevant with respect to the query. This ensures that training is carried out on the optimal information signal to bring the query and the relevant sentence in the document together. Finally, we also develop a cross-lingual dataset (Cross_TC+IFCN), where we augmented the TC+IFCN dataset by translating the query and document pairs to Spanish, French, and German using OPUS-MT (Tiedemann and Thottingal, 2020). Cross_TC+IFCN is used to fine-tune multilingual models which are employed for runs that involve documents in languages other than English.

For training the bi-encoder, we utilised the models provided by the sentence-transformers[4] library, which includes BERT-based models (Reimers and Gurevych, 2019) fine-tuned using siamese and triplet networks to get semantically meaningful sentence representations. We further fine-tuned the SBERT models on our domain-specific dataset. The details of the models used in our experiments are as follows:

- For monolingual English runs, we try two models: 1) msmarco-distilroberta-base-v2, RoBERTa (Liu et al., 2019) base model trained on MSMARCO passage ranking dataset and 2) stsb-roberta-large, a RoBERTa large model trained on natural language inference and semantic textual similarity dataset. We use these as base models to fine-tune on the TC+IFCN dataset with a regression objective function (Reimers and Gurevych, 2019). The final models are referred to as TCIN-msmarco-distilroberta-base and TCIN-stsb-roberta-large.
- For bilingual runs, we use paraphrase-xlm-r-multilingual-v1 (Reimers and Gurevych, 2019) which is a xlm-roberta-base (Liu et al., 2019) model trained on a large scale paraphrase dataset of more than 50 languages. Transfer learning is used to fine-tune this model on Cross_TC+IFCN dataset using the same objective function as mentioned above and the final multilingual model is named as CrossTCIN-xlm-r-paraphrase.

We call this the neural refinement stage since it helps to filter out all semantically unrelated documents and it also works much faster when compared to a cross-encoder-based approach where a pair of sentences are passed together to the model every time during inference.

[4] https://github.com/UKPLab/sentence-transformers

4.4.3 Neural re-ranking stage

In the third stage, the top 400 documents retrieved by the neural refinement stage are re-ranked using a cross-encoder architecture. In this, both query tokens and the document tokens separated by [SEP] token are passed to the transformer-based model to perform full self-attention over the given input and the output of [CLS] token is passed to the linear layer with sigmoid activation to get a relevance scores from 0 to 1 (Nogueira and Cho, 2019) as illustrated in the neural re-ranking stage of Fig 4.1. Similar to the neural refinement stage, here also we split the document into sentences and apply sentence-level inference using the query. The final relevance score for each document is determined by combining the top k scoring sentences of the document i.e.

$$CrossScore(d) = \sum_{i=1}^{k} w_i \cdot S_{Cross_i} \tag{4.3}$$

where $CrossScore(d)$ is the document-level relevance score for document d using the cross-encoder model, S_{Cross_i} is the i-th top sentence score and all other parameters are kept the same as in Equation 4.2. Although cross-encoder is more accurate than bi-encoder, they are compute-intensive and time-consuming when compared to bi-encoder and this is the reason we give fewer documents to re-rank in the neural re-ranking stage. The description of models used as cross-encoders is as follows

- For monolingual English runs, we use ELECTRA model fine-tuned on the MSMARCO dataset from (Li et al., 2020a) as it was among the top positions in the second round of TREC-COVID task. The model was further fine-tuned on the TC+IFCN data using binary cross-entropy loss. We call this model as TCIN-electra-msmarco.
- For bilingual runs, currently, there does not exist any multilingual model trained on the MSMARCO passage ranking dataset. Hence, we use state-of-the-art multilingual transformer-based models such as xlm-roberta-base (Liu et al., 2019) and distilbert-base-multilingual (Devlin et al., 2018) as a base model for fine-tuning on the MSMARCO passage dataset for an epoch with a learning rate of 1e-5, batch size of 16, and a maximum of 512 input sequence length. The models are further fine-tuned on Cross_TC+IFCN dataset using sigmoid cross-entropy loss and the final models are referred to as CrossTCIN-xlm-roberta-msmarco and CrossTCIN-distilbert-multilingual-msmarco.

In the case of bilingual runs, where query and documents are in a different language, we simply use Google Translate to translate the query into the target language of the document and apply the same methods as described above in a monolingual setting. For some runs, we also combine scores from different stages using various rank fusion algorithms. These can be broadly classified into score-based and rank-based fusion algorithms. For the score-based method, we use weighted CombSUM which is a slight modification of CombSUM (Fox and Shaw, 1994) algorithm where we add the weighted scores from different ranking models. The equation is as follows

$$\begin{aligned} wCombSUM(d) = \alpha \cdot norm(CrossScore(d)) + \beta \cdot norm(BiScore(d)) \\ +(1-\alpha-\beta) \cdot norm(BM25Score(d)) \end{aligned} \tag{4.4}$$

where $wCombSUM(d)$ is the weighted CombSUM score of the document d, $norm$ is the min-max normalisation of relevance scores. $BM25Score(d)$, $BiScore(d)$ and $CrossScore(d)$ are the relevance scores of document d from first, second and third stage respectively. The parameters α and β are such that $\alpha > \beta$ and we have fixed $\alpha = 0.5$ and $\beta = 0.4$, giving more weight to cross-encoder ranking followed by bi-encoder and BM25 ranking respectively. For rank-based fusion methods, we tried Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) and Borda Fusion (Aslam and Montague, 2001). In this, the fused score of the document simply relies on the rank of the document from different ranking models, hence in our work we only use the output of neural refinement and neural re-ranking stage. The equation of RRF and Borda fusion is as follows

$$RRFScore(d) = \frac{1}{k + R_{Cross}(d)} + \frac{1}{k + R_{Biencoder}(d)} \tag{4.5}$$

$$BordaScore(d) = \frac{N - R_{Cross}(d) + 1}{N} + \frac{N - R_{Biencoder}(d) + 1}{N} \tag{4.6}$$

where $RRFScore(d)$ and $BordaScore(d)$ are the RRF and Borda fusion scores. $R_{Biencoder}(d)$ and $R_{Cross}(d)$ are the ranks of the document d from neural refinement and neural re-ranking stage. For the RRF method, we set the constant $k = 60$ default as mentioned in their respective paper (Cormack et al., 2009). In Borda fusion, N is the total number of documents during fusion.

4.5 Experimental details

4.5.1 GATENLP runs

We implement BiCross encoder to check its effectiveness in both monolingual and bilingual search settings. Our system is designed such that it aims at achieving both high precision as well as high recall. We submitted a total of 22 monolingual runs and 15 bilingual runs. The runs differ in terms of models, type of query, and rank fusion methods. The description of all the runs is given below and all runs retrieve 200 documents for each query. All our runs have `gatenlp_` as a prefix in the name of the run. The experiments were conducted on NVIDIA Titan RTX GPU.

- `gatenlp_run1 / gatenlp_run2 / gatenlp_run3` : In run 1, *Udels* method is used to generate the query, *t5_query* is used in run 2 and a concatenation of the keyword and conversational (*key_conv*) field in run 3. The bi-encoder is TCIN-msmarco-distilroberta-base and cross-encoder is TCIN-electra-msmarco. The encoder models are kept the same in all the runs to see which type of query gives the best results.

- `gatenlp_run5 / gatenlp_run7` : In run 5 and run 7, we use a different bi-encoder i.e. TCIN-stsb-roberta-large and the cross-encoder model is TCIN-electra-msmarco. The only difference between both the runs is that run 5 uses *key_conv* query and run 7 uses *Udels* query.

- `gatenlp_es_run25 / gatenlp_fr_run26 / gatenlp_de_run27 / gatenlp_es_run28 / gatenlp_fr_run29 / gatenlp_de_run30` : These are monolingual Spanish, French and German runs. Here, we use multilingual models where bi-encoder is CrossTCIN-xlm-r-paraphrase and cross-encoder is CrossTCIN-distilbert-multilingual-msmarco. In case of run 25, 26 and 27, weighted CombSum fusion is used whereas for run 28, 29 and 30, RRF fusion is used to get the relevance score for each document. In all these runs and the following monolingual runs, *key_conv* query is used to retrieve the documents. The evaluation results of these runs will depict the best performing rank fusion algorithm.

- `gatenlp_es_run31 / gatenlp_fr_run32 / gatenlp_de_run33 / gatenlp_es_run34 / gatenlp_fr_run35 / gatenlp_de_run36 / gatenlp_es_run37 / gatenlp_fr_run38 / gatenlp_de_run39` : In the above runs, we use XLM-based model i.e. CrossTCIN-xlm-roberta-msmarco as a cross-encoder and CrossTCIN-xlm-r-paraphrase as bi-encoder. Here, we use rank-based fusion where run 31, 32 and 33 uses RRF and run 34, 35 and 36 uses Borda fusion. For run 37, 38 and 39, no fusion method is used and we directly get the output from cross-encoder in the neural re-ranking stage.

- `gatenlp_en2es_run43 / gatenlp_en2fr_run44 / gatenlp_en2de_run45 / gatenlp_en2es_run46 / gatenlp_en2fr_run47 / gatenlp_en2de_run48` : These are all bilingual runs where the language of query is English and the language of documents is depicted by ISO 639-1 code after the `gatenlp_en2` identifier in the run name. Here we use multilingual models where the bi-encoder is CrossTCIN-xlm-r- paraphrase and cross-encoder is CrossTCIN-xlm-roberta-msmarco. All runs use the final output from the cross-encoder, however, *key_conv* query is used in run 43, 44 and 45 and *Udels* query in run 46, 47 and 48.

- `gatenlp_en2es_run49 / gatenlp_en2fr_run50 / gatenlp_en2de_run51 / gatenlp_en2es_run52 / gatenlp_en2fr_run53 / gatenlp_en2de_run54 :` For these runs, we use CrossTCIN-xlm-r-paraphrase as bi-encoder and CrossTCIN-distilbert-multilingual-msmarco as cross-encoder. Run 49, 50 and 51 uses *Udels* query and run 52, 53 and 54 uses *key_conv* as query for retrieving the documents.

4.5.2 Other participant runs

Besides our team, there are three more participants in the MLIA task 2, 1) Sinai (Universidad de Ja´en, Spain) 2) Cunimtir (Charles University, Czech Republic) 3) Ims (University of Padua, Italy) and there are in total 109 monolingual runs and 66 bilingual runs submitted.

In particular, Sinai (Martin-Valdivia, 2020) specifically focused on Spanish language and used Lucene to do BM25 search using different fields of topics as a query on the index of different XML tag contents of the documents. They employed keywords (`sinai1` & `sinai5`), conversational (`sinai2`), explanation (`sinai3`), and a combination of all fields as a query (`sinai4`).

On the other hand, Cunimtir's (Saleh et al., 2020) monolingual runs employed language based Dirichlet model (`cunimtir_run1`), Per-Field Normalisation Weighting (Pl2F) model (`cunimtir_run2`), Dirichlet model for conversational field as a query (`cunimtir_run5`) and two famous query expansion models i.e. Bose-Einstein (Bo2) model (`cunimtir_run3`) and Kullback-Leibler divergence (KLD) correct (`cunimtir_run4`). For multilingual runs, they used neural machine translation models for translating the query into the document language before performing the retrieval.

Finally, Ims (Di Nunzio et al., 2020) submitted multiple runs. Firstly, for each language, they submitted runs that use BM25 with default Lucene parameters where `ims_bm25` uses the keyword field of the topic as a query and `ims_c-bm25` uses keyword and conversational formulation as a query. In addition to this, they also submitted run `ims_csum` which is a one-stage CombSUM fusion of all the lexical runs, using only the keyword formulation of the query. In `ims_v-csum`, they used a two-stage fusion to merge runs associated with query reformulations. For English runs, they submitted a few more additional runs such as `ims_nlex`, a three-stage fusion using the topic formulations, lexical runs and neural runs, and ims_nsle which uses a SLEDGE model (MacAvaney et al., 2020) fine-tuned on the medical subset of the MSMARCO dataset to re-rank the documents. We refer readers to the original papers (Martin-Valdivia, 2020; Saleh et al., 2020; Di Nunzio et al., 2020) for further details.

4.6 Results and discussion

In this section, we explore the effectiveness of our approach. We evaluate the performance of Multistage BiCross Encoder using the relevance assessments provided by the MLIA organisers. All the submitted runs are evaluated using precision and normalised discounted cumulative gain (NDCG) that focus on top-ranked documents and, recall and R-precision (RPrec) whose focus is more on finding as many relevant documents as possible in all the retrieved documents.

In all the results tables (Table 4.3 – 4.7), the notation within the brackets following each run ID serves as a descriptor for the query type and the method employed. The first component before the hyphen indicates the type of query used, while the rest of the components after the hyphen denote the specific method utilised. The notations used are as as follows: 1) Query type: keywords (K), conversational (C), explanation (E), *t5_query* (T5), *key_conv* (KC), *Udels* (UD); 2) Bi-encoder model: TCIN-msmarco-distilroberta-base (TMDB), TCIN-stsb-roberta-large (TSRL), CrossTCIN-xlm-r-paraphrase (CXP); 3) Cross-encoder model: TCIN-electra-msmarco (TEM), CrossTCIN-xlm-roberta-msmarco (CXRM), CrossTCIN-distilbert-multilingual-msmarco (CDMM).

For instance, gatenlp_run5 (KC-TSRL-TEM) signifies a GATENLP run conducted using a combination of keyword and conversational field as a query (KC), employing the TCIN-stsb-roberta-large (TSRL) as bi-encoder and TCIN-electra-msmarco (TEM) as cross-encoder. Overall, these notations aid in categorising and understanding the variations in retrieval performance observed across different query types and methods.

4.6.1 Monolingual Runs

Table 4.3 shows the results of the monolingual English runs. The first part of the table contains runs which retrieve 200 documents per query as these are a part of subtask 2 and the second part of the table contains runs which retrieve 1000 documents per query as this is a requirement for subtask 1. We focus on subtask 2 where we aim at achieving both high recall as well as high precision values for the least number of retrieved documents per topic query. Over and above, we include runs from both the subtasks so as to do a fair comparison of performance of our runs with all the submitted runs in the MLIA task.

As shown in Table 4.3, GATENLP runs outperform all other participant runs in all metrics by a significant margin (p-value<0.001 using paired t-test for all metrics) for subtask 2 runs. Amongst our runs, gatenlp_run5 gives the highest scores, followed by gatenlp_run3 and other runs shown in the table. The gatenlp_run5 uses *key_conv* as a query and TCIN-stsb-roberta-large (TSRL) as a bi-encoder model. This suggests that the use of bi-encoder model pre-fine-tuned on STS data proved to be beneficial when compared to the ones pre-fine-tuned on MSMARCO dataset for monolingual English runs. Regarding the type of query, the results show that employing *key_conv* as a query achieves large gains as compared to *Udels* query and *t5_query* method. Furthermore, *t5_query* yields comparatively lower results for most of the metrics, suggesting that retrieval using concatenated reworded queries introduces noise in retrieved documents. Even though subtask 1 runs retrieve 1000 documents per query, our runs still perform equally well and in some cases even surpass subtask 1 runs despite the fact that we only retrieve 200 documents for each query. This indicates that retrieving more documents leads to high recall but there is minimal difference in performance for precision focused metrics. At last, the results also show that neural methods perform far better than the lexical-based BM25 baselines such as the ones used in ims and sinai runs.

For monolingual Spanish (Table 4.4), our runs outperformed all other submissions, to a statistically significant degree (paired t-test p-value<0.001 for all metrics). The gatenlp_run37 (KC-CXP-CXRM) scores highest in precision, whereas gatenlp_run25 (KC-CXP-CDMM-CSum) gave the best results for NDCG and recall. Conversely, gatenlp_run31 (KC-CXP-CXRM-RRF) secured the highest scores for MAP and Rprec. Similarly, Table 4.5 and Table 4.6 show the results for monolingual French and monolingual German respectively. Although ours is the only team that submitted monolingual French runs, we achieve highly competent scores. Overall, we find that the runs which use weighted CombSUM on bi-encoder CrossTCIN-xlm-r-paraphrase (CXP) and cross-encoder CrossTCIN-distilbert-multilingual-msmarco (CDMM) give top scores for recall and NDCG. We also find that the best performing run in each monolingual case attains P@5≥0.8, depicting that there is at least an average of 80% chance of getting a relevant document in the top 5 retrieved documents. Apart from this, we couldn't find any single method which performs well for all languages and metrics, as different methods yield distinct results and there is considerable variability in performance.

Additionally, we also speculate on the performance of multilingual models across various languages. We find that, aside from English, for most metrics, the scores of German runs are comparatively higher, followed by French and Spanish runs respectively. These differences might arise from intrinsic language variations in the pre-training of multilingual transformer models (eg. mBERT, XLM-RoBERTa etc) or due to differences in the document processing pipeline of different languages

Table 4.3: *Results for monolingual English runs. Our runs have* `gatenlp_` *as a prefix in the name of the run. The first part of the table contains runs which retrieve 200 documents per query (subtask 2) and the second part of the table contains runs which retrieve 1000 documents for each query (subtask 1). Best overall scores are highlighted in bold.*

Run ID	P@5	P@10	MAP	NDCG@10	NDCG	Rprec	Recall
gatenlp_run5 (KC-TSRL-TEM)	**0.9333**	**0.9000**	**0.2944**	**0.8331**	**0.5187**	**0.3486**	0.4382
gatenlp_run3 (KC-TMDB-TEM)	0.9200	0.8900	0.2912	0.8223	0.5155	0.3484	0.4375
gatenlp_run2 (T5-TMDB-TEM)	0.9000	0.7967	0.2560	0.7775	0.4925	0.3215	0.4278
gatenlp_run1 (UD-TMDB-TEM)	0.8867	0.8633	0.2776	0.8139	0.5067	0.3310	**0.4411**
gatenlp_run7 (UD-TSRL-TEM)	0.8667	0.8800	0.2719	0.8212	0.5014	0.3305	0.4292
cunimtir_run1 (K-Dirichlet)	0.5933	0.4800	0.1145	0.4254	0.2802	0.1976	0.2613
cunimtir_run3 (K-Bo2)	0.3600	0.3233	0.0609	0.2712	0.1444	0.1046	0.1278
cunimtir_run4 (K-KLD)	0.3533	0.3267	0.0530	0.2688	0.1422	0.0940	0.1239
ims_bm25_1k (K-BM25)	0.3067	0.2433	0.0688	0.2391	0.2418	0.1579	0.2595
ims_bm25_2k (K-BM25)	0.2400	0.1833	0.0478	0.1744	0.1789	0.1277	0.2028
ims_bm25_3k (K-BM25)	0.2067	0.1633	0.0396	0.1413	0.1582	0.1075	0.1930
ims_bm25_4k (K-BM25)	0.1933	0.1533	0.0367	0.1546	0.1483	0.1037	0.1677
ims_nlex (KCE-Three-Stage)	**0.8933**	**0.9000**	**0.3055**	**0.8365**	0.5740	0.3408	0.5593
ims_c-bm25 (KC-BM25)	0.8600	0.8267	0.2771	0.7592	0.5945	0.3089	0.6482
ims_v-csum (KCE-Two-Stage)	0.8533	0.8233	0.2999	0.7693	**0.6092**	**0.3450**	**0.6516**
ims_bm25 (K-BM25)	0.7200	0.6900	0.2269	0.6202	0.5264	0.2673	0.6079
cunimtir_run5 (C-Dirichlet)	0.6867	0.6900	0.1908	0.5780	0.4574	0.2364	0.5160
cunimtir_run1 (K-Dirichlet)	0.6800	0.5033	0.1659	0.4928	0.4450	0.2223	0.5148
ims_nsle (KCE-SLEDGE)	0.5067	0.5133	0.1595	0.4084	0.4145	0.2205	0.4837
cunimtir_run3 (K-Bo2)	0.4800	0.3367	0.0882	0.2944	0.2500	0.1221	0.2986
cunimtir_run2 (K-Pl2F)	0.4667	0.3033	0.0646	0.3005	0.2379	0.1163	0.2683
cunimtir_run4 (K-KLD)	0.4267	0.3400	0.0658	0.2809	0.2200	0.1051	0.2662

or by both of these factors.

Table 4.4: *Results for monolingual Spanish runs. Our runs have* `gatenlp_` *as a prefix in the name of the run. The first part of the table contains runs which retrieve 200 documents per query and the second part of the table contains runs which retrieve 1000 documents for each query. Best overall scores are highlighted in bold.*

Run ID	P@5	P@10	MAP	NDCG@10	NDCG	Rprec	Recall
gatenlp_run37 (KC-CXP-CXRM)	**0.8333**	**0.7933**	0.2043	0.7263	0.3705	0.2806	0.3086
gatenlp_run25 (KC-CXP-CDMM-CSum)	0.8133	0.7767	0.2154	0.7455	**0.3808**	0.2795	**0.3111**
gatenlp_run28 (KC-CXP-CDMM-RRF)	0.8067	0.7767	0.2113	**0.7478**	0.3768	0.2758	0.3086
gatenlp_run34 (KC-CXP-CXRM-Borda)	0.7933	0.7833	0.2173	0.7383	0.3769	0.2858	0.3086
gatenlp_run31 (KC-CXP-CXRM-RRF)	0.7933	0.7867	**0.2246**	0.7362	0.3790	**0.2873**	0.3086
sinai_sinai1 (K-BM25)	0.5200	0.4867	0.0900	0.4629	0.2177	0.1557	0.1767
sinai_sinai2 (C-BM25)	0.4400	0.4067	0.0631	0.3868	0.1835	0.1284	0.1537
sinai_sinai4 (KCE-BM25)	0.3600	0.3067	0.0535	0.2904	0.1738	0.1243	0.1594
sinai_sinai3 (E-BM25)	0.2267	0.1733	0.0284	0.1820	0.1121	0.0786	0.1011
sinai_sinai5 (K-BM25)	0.2267	0.1733	0.0155	0.1832	0.0634	0.0407	0.0444
ims_bm25_1k (K-BM25)	0.2067	0.1867	0.0577	0.1812	0.1944	0.1366	0.2142
ims_bm25_2k (K-BM25)	0.2000	0.1800	0.0591	0.1745	0.2003	0.1402	0.2275
ims_bm25_3k (K-BM25)	0.1733	0.1433	0.0444	0.1359	0.1744	0.1196	0.2072
ims_bm25_4k (K-BM25)	0.0867	0.0800	0.0309	0.0793	0.1535	0.1046	0.1900
ims_c-bm25 (KC-BM25)	**0.7000**	0.6933	0.1654	0.6346	**0.3993**	0.2224	**0.4084**
ims_v-csum (KCE-Two-Stage)	0.6867	**0.7133**	0.1697	**0.6604**	0.3797	0.2171	0.3612
ims_csum (K-One-Stage)	0.6800	0.6200	**0.1720**	0.5822	0.3769	**0.2259**	0.3779
ims_bm25 (K-BM25)	0.6133	0.5800	0.1458	0.5263	0.3540	0.2020	0.3740
sinai_sinai1 (K-BM25)	0.5200	0.4867	0.1000	0.4629	0.2839	0.1560	0.2928
sinai_sinai2 (C-BM25)	0.4400	0.4067	0.0715	0.3868	0.2436	0.1285	0.2618
sinai_sinai4 (KCE-BM25)	0.3600	0.3067	0.0626	0.2904	0.2368	0.1247	0.2689
sinai_sinai5 (K-BM25)	0.2267	0.1733	0.0157	0.1832	0.0693	0.0408	0.0550
sinai_sinai3 (E-BM25)	0.2267	0.1733	0.0342	0.1820	0.1644	0.0788	0.1906

Table 4.5: *Results for monolingual French runs. All runs retrieve 200 documents per query as there were no runs submitted by any team which retrieve 1000 documents per query. Best overall scores are highlighted in bold.*

Run ID	P@5	P@10	MAP	NDCG@10	NDCG	Rprec	Recall
gatenlp_run26 (KC-CXP-CDMM-CSum)	**0.8800**	**0.7533**	**0.3505**	**0.7490**	**0.5672**	**0.3773**	**0.5267**
gatenlp_run29 (KC-CXP-CDMM-RRF)	0.8600	0.7400	0.3302	0.7324	0.5406	0.3651	0.4926
gatenlp_run32 (KC-CXP-CXRM-RRF)	0.8133	0.7367	0.3161	0.7180	0.5297	0.3593	0.4926
gatenlp_run35 (KC-CXP-CXRM-Borda)	0.8133	0.7267	0.3125	0.7116	0.5268	0.3541	0.4926
gatenlp_run38 (KC-CXP-CXRM)	0.7867	0.6400	0.2752	0.6436	0.5030	0.3269	0.4926

4.6.2 Bilingual Runs

Table 4.7 compares the performance of our different bilingual runs. These include English to Spanish (en2es), English to French (en2fr) and English to German

Table 4.6: *Results for monolingual German runs. Our runs have `gatenlp_` as a prefix in the name of the run. The first part of the table contains runs which retrieve 200 documents per query and the second part of the table contains runs which retrieve 1000 documents for each query. Best overall scores are highlighted in bold.*

Run ID	**P@5**	**P@10**	**MAP**	**NDCG@10**	**NDCG**	**Rprec**	**Recall**
gatenlp_run30 (KC-CXP-CDMM-RRF)	**0.9067**	**0.8767**	0.4537	**0.8234**	0.6403	0.4794	0.6253
gatenlp_run27 (KC-CXP-CDMM-CSum)	0.9000	0.8667	**0.4629**	0.8211	**0.6488**	0.4858	**0.6339**
gatenlp_run36 (KC-CXP-CXRM-Borda)	0.8733	0.8267	0.4442	0.7772	0.6377	0.4843	0.6253
gatenlp_run33 (KC-CXP-CXRM-RRF)	0.8733	0.8300	0.4531	0.7793	0.6399	**0.4972**	0.6253
gatenlp_run39 (KC-CXP-CXRM)	0.7733	0.7700	0.4227	0.7078	0.6200	0.4601	0.6253
ims_bm25_1k (K-BM25)	0.1667	0.1633	0.0700	0.1475	0.2288	0.1413	0.3063
ims_bm25_2k (K-BM25)	0.1667	0.1600	0.0793	0.1515	0.2176	0.1388	0.2769
ims_bm25_4k (K-BM25)	0.1467	0.1433	0.0629	0.1396	0.1967	0.1120	0.2589
ims_bm25_3k (K-BM25)	0.1400	0.1367	0.0650	0.1276	0.1924	0.1163	0.2488
ims_v-csum (KCE-Two-Stage)	**0.7267**	**0.6733**	**0.3447**	**0.6341**	**0.6174**	**0.3737**	0.7080
ims_csum (K-One-Stage)	0.6267	0.5700	0.3072	0.5315	0.5731	0.3507	0.6940
ims_c-bm25 (KC-BM25)	0.6133	0.5633	0.2890	0.5150	0.5667	0.3131	**0.7114**
ims_bm25 (K-BM25)	0.5933	0.5333	0.2869	0.4912	0.5572	0.3173	0.6924

(en2de) runs. The best performing run in each bilingual case attains P@10≥0.7 even when the language of the query and document is different. As all runs retrieve a total of 200 documents for each query, the recall value remains similar for all three bilingual cases, which shows that the MLIA corpus does not contain many relevant documents in Spanish, French, and German language.

If we compare the performance of run 43, 44 and 45 with run 46, 47 and 48, we see that the former runs, which use *key_conv* as query, give better results than the latter ones which use *Udels* query. We see similar results for run 49, 50 and 51 which use *Udels* query and run 52, 53 and 54 which use *key_conv* as a query. This shows that the use of keyword and conversational formulation as a query gives the best results for bilingual runs (p-value≤0.05 for paired t-test). Furthermore, we find that using CrossTCIN-xlm-r-paraphrase (CXP) as the bi-encoder and CrossTCIN-distilbert-multilingual-msmarco (CDMM) as the cross-encoder consistently yields the highest scores for all bilingual cases across all metrics. This suggests the superiority of cross-encoder fine-tuned using distilbert-base-multilingual (CDMM) compared to xlm-roberta-base (CXRM) (see Table 4.7). Although ours were the only runs submitted for the above given bilingual pairs, the evaluation results show that using BiCross encoder by machine translating query into document language helped in attaining competitive baselines for future research. It is important to note that only the first stage (BM25) requires translation; both the refinement and reranking stages have the ability to encode documents in multiple languages in the same vector space (see Section 4.4.2 & 4.4.3).

Table 4.7: *Results for bilingual Spanish (es), French (fr) and German (de) runs. Here the language of the query is English and the language of documents is depicted by ISO 639-1 code after the `gatenlp_en2` identifier in the run name. All runs retrieve 200 documents per query. Best overall scores are highlighted in bold.*

Run ID	P@5	P@10	MAP	NDCG@10	NDCG	Rprec	Recall
gatenlp_en2es_run49 (UD-CXP-CDMM)	**0.8533**	0.7367	0.1579	0.7042	0.3214	0.2273	**0.2565**
gatenlp_en2es_run52 (KC-CXP-CDMM)	0.8200	**0.7700**	**0.1666**	**0.7368**	**0.3287**	0.2286	**0.2565**
gatenlp_en2es_run43 (KC-CXP-CXRM)	0.8000	0.6867	0.1538	0.6555	0.3155	**0.2287**	**0.2565**
gatenlp_en2es_run46 (UD-CXP-CXRM)	0.7733	0.6367	0.1439	0.6330	0.3120	0.2231	**0.2565**
gatenlp_en2fr_run53 (KC-CXP-CDMM)	**0.8400**	**0.7467**	**0.2870**	**0.7245**	**0.4993**	**0.3220**	**0.4452**
gatenlp_en2fr_run44 (KC-CXP-CXRM)	0.7667	0.6667	0.2527	0.6622	0.4801	0.3107	**0.4452**
gatenlp_en2fr_run47 (UD-CXP-CXRM)	0.7400	0.6300	0.2378	0.6234	0.4633	0.2980	0.4360
gatenlp_en2fr_run50 (UD-CXP-CDMM)	0.7133	0.6700	0.2506	0.6521	0.4712	0.3054	0.4360
gatenlp_en2de_run54 (KC-CXP-CDMM)	**0.7733**	**0.7267**	**0.2680**	**0.7007**	**0.4484**	**0.3221**	**0.3950**
gatenlp_en2de_run51 (UD-CXP-CDMM)	0.7200	0.6867	0.2475	0.6568	0.4334	0.3029	0.3907
gatenlp_en2de_run45 (KC-CXP-CXRM)	0.7133	0.6700	0.2474	0.6444	0.4349	0.3076	**0.3950**
gatenlp_en2de_run48 (UD-CXP-CXRM)	0.6867	0.6433	0.2292	0.6099	0.4187	0.2978	0.3907

On the whole, domain-specific fine-tuning of transformer models on TC+IFCN dataset gave a boost in performance. Also, the results suggest that fine-tuning multilingual models such as mBERT and XLM-RoBERTa on Cross_TC+IFCN can make models quickly adapt to the domain-specific data and transfer relevance matching across languages. This is coherent with the previous work (Shi and Lin, 2019). In spite of the fact that our training dataset consists mainly of scientific scholarly research papers, our fine-tuned models were able to transfer knowledge to articles about general COVID-19 health-related content and thereby accomplishing promising results on the MLIA corpus. It is also worth emphasising the effectiveness of BiCross encoder to refine and re-rank the candidate documents which has the dual advantage of high precision and high recall values in both monolingual and bilingual runs. Regarding the computational complexity of BiCross encoder, it depends on the number of documents to be re-ranked by the cross-encoder in the neural re-ranking stage and this can be controlled by restricting the count of top n documents retrieved from neural refinement stage. The benefit of the neural refinement stage is that the representation from the bi-encoder can be stored locally which obviates the need for an inference pass to the encoder model during re-ranking and this process can be expedited with GPU based implementation of similarity search such as FAISS (Johnson et al., 2017).

4.7 Conclusion

This paper proposed a novel Multistage BiCross Encoder developed for the MLIA COVID-19 multilingual semantic search (task 2). As detailed above, the multistage BiCross encoder is a three-stage approach consisting of an initial retrieval using Okapi BM25 algorithm followed by a transformer-based bi-encoder and cross-encoder to effectively rank the documents using sentence-level score aggregation with respect to the query. Our method exploited transfer learning, by finetuning large pre-trained transformer models on domain-specific data for retrieving COVID-19 health-related articles in multiple languages. While the approach is conceptually simple, the independently evaluated MLIA results demonstrate that the use of bi-encoder and cross-encoder along with BM25 is highly effective in outperforming other state-of-the-art methods according to a wide range of metrics, and it has the twofold benefit of high precision in retrieving the top-ranked documents (P@5$\geq$0.8 for best performing run), as well as a high recall for all retrieved documents. We also find that employing keyword and conversational formulation as a query gives the highest scores in both monolingual and bilingual search settings. We hope that our research will help improve multilingual access to reliable COVID-19 health information thereby mitigating the impact of the 'infodemic' as a consequence of the ongoing COVID-19 pandemic.

Future work will experiment with further hyperparameter tuning and making additional improvements to the neural architecture. We also plan to test the Multistage BiCross Encoder on other document retrieval and similar information access tasks.

4.8 Acknowledgement

This work has been partially supported by the WeVerify and SoBigData++ projects (EU H2020, grant agreements: 825297 and 871042).

Chapter 5

Breaking Language Barriers with MMTweets: Advancing Cross-Lingual Debunked Narrative Retrieval for Fact-Checking

Iknoor Singh, Xingyi Song, Kalina Bontcheva and Carolina Scarton
Department of Computer Science, The University of Sheffield, UK

Abstract

Finding previously debunked narratives involves identifying claims that have already undergone fact-checking. The issue intensifies when similar false claims persist in multiple languages, despite the availability of debunks for several months in another language. Hence, automatically finding debunks (or fact-checks) in multiple languages is crucial to make the best use of scarce fact-checkers' resources. Mainly due to the lack of readily available data, this is an understudied problem, particularly when considering the cross-lingual scenario, i.e. the retrieval of debunks in a language different from the language of the online post being checked. This study introduces cross-lingual debunked narrative retrieval and addresses this research gap by: (i) creating Multilingual Misinformation Tweets (MMTweets): a dataset that stands out, featuring cross-lingual pairs (claims in 4 different languages, and debunks in 11 different languages), images, human anno-

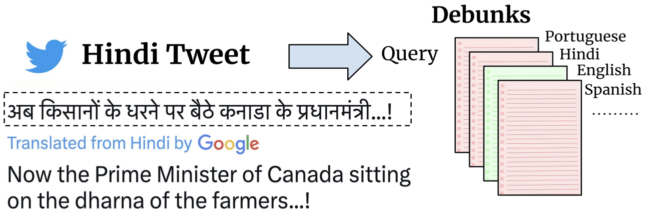

Hindi Tweet
अब किसानों के धरने पर बैठे कनाडा के प्रधानमंत्री...!
Translated from Hindi by Google
Now the Prime Minister of Canada sitting on the dharna of the farmers...!
Query
Debunks
Portuguese
Hindi
English
Spanish
....

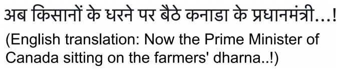
Figure 5.1: *Cross-lingual debunked narrative retrieval: Query tweet is in Hindi and the relevant debunk is in English.*

tations, and fine-grained labels, making it a comprehensive resource compared to its counterparts; (ii) conducting an extensive experiment to benchmark state-of-the-art cross-lingual retrieval models and introducing multistage retrieval methods tailored for the task, achieving an NDCG@5 score of 0.669 and an MRR score of 0.795; and (iii) comprehensively evaluating retrieval models for their cross-lingual and cross-dataset transfer capabilities within MMTweets, sensitivity to negative samples, and retrieval latency analysis. We find that MMTweets presents challenges for cross-lingual debunked narrative retrieval, highlighting areas for improvement in retrieval models. Nonetheless, the study provides valuable insights for creating MMTweets datasets and optimising debunked narrative retrieval models to empower fact-checking endeavours. The dataset and annotation codebook are publicly available at https://doi.org/10.5281/zenodo.10637161.

5.1 Introduction

Automated fact-checking systems play a vital role in both countering false information on digital media and alleviating the burden on fact-checkers (Nielsen and McConville, 2022; Shang et al., 2023; Wu et al., 2022; Guo et al., 2022; Zhang et al., 2022a; Zeng et al., 2021). A key task of these systems is the detection of previously fact-checked similar claims – an information retrieval problem where claims serve as queries to retrieve from a corpus of debunks (Nakov et al., 2022c, 2021b; Shaar et al., 2020b). This task aims to detect claims that spread even after they have already been debunked by at least one professional fact-checker. Previous work has focused on training retrieval models, primarily focusing on monolingual retrieval, where the language of the query claim matches the language of the debunk (Nakov et al., 2022c, 2021b; Shaar et al., 2020b; Kazemi et al., 2021). Moreover, these monolingual retrieval models assume that the debunks exist exclusively in one language. However, previous studies (Singh et al., 2022, 2021a; Reis et al., 2020) demonstrate that similar false claims continue to spread in multiple languages, despite the availability of debunks for several months in another

Table 5.1: *Sample query tweets and their corresponding debunks from the MMTweets dataset.*

Fields	**Hindi Query Tweet - English Debunk**	**English Query Tweet - Spanish Debunk**
Tweet	 अब किसानों के धरने पर बैठे कनाडा के प्रधानमंत्री....! (English translation: Now the Prime Minister of Canada sitting on the farmers' dharna..!)	I Sultue you Sir. You are So intelligent. RUSSIA: Vladimir Putin has Dropped 800 tigers and Lions all over the Country to push people to stay Home...Stay Safe Everyone!!
Debunk title	Old Photo Passed Off As Justin Trudeau Sitting In An Anti-Farm Laws Protest	La foto del león en la calle fue tomada en Sudáfrica en 2016 y no tiene relación con la pandemia del COVID-19
Debunk claim	Justin Trudeau sits in protest in support of the protesting farmers.	Publicaciones compartidas más de 35.000 veces en redes sociales desde el 22 de marzo último aseguran que Rusia liberó..

language. Hence, automatically finding debunks in multiple languages is crucial to make the best use of scarce fact-checkers' resources.

In this study, we define the task of **cross-lingual Debunked Narrative Retrieval (X-DNR)** as a cross-lingual information retrieval problem where a claim is used as a query to retrieve from a corpus of debunks in multiple languages (see Figure 5.1). In this paper, we use the term "debunked narrative retrieval" over the previously used term "fact-checked claim retrieval" because the term "debunked narrative" better captures the range of false narratives or stories related to a claim that has already been debunked. This term acknowledges that a single claim can have multiple narratives, all needing debunking, unlike fact-checked claim retrieval, which focuses narrowly on verified claims without addressing their diverse associated narratives. Therefore, the term "debunked narrative retrieval" is more fitting for this task, as the primary objective of X-DNR is to aid fact-checkers in identifying debunked narratives across multiple languages. Our main contributions are:

- The **M**ultilingual **M**isinformation **Tweets** (**MMTweets**): a novel benchmark that stands out, featuring cross-lingual pairs, images, and fine-grained human annotations, making it a comprehensive resource compared to its counterparts (see Section 5.3.4). In total, it comprises $1,600$ query tweet claims (in Hindi, English, Portuguese & Spanish) and $30,452$ debunk corpus (in 11 different languages) for retrieval. Table 5.1 shows dataset examples.
- An extensive evaluation of state-of-the-art (SOTA) cross-lingual retrieval models on the MMTweets dataset. We also introduce two multistage retrieval methods (*BE+CE* and *BE+GPT3.5*) adapting earlier approaches to effectively address the

cross-lingual nature of the X-DNR task. Nevertheless, the results suggest that dealing with multiple languages in the MMTweets dataset poses a challenge, and there is still room for improvement in models.
- A comprehensive evaluation aims to investigate: 1) cross-lingual transfer and generalisation across languages within MMTweets; 2) how challenging it is for models trained on existing datasets to transfer knowledge to the MMTweets test set; 3) the impact of the type and count of negative pairs on the model's performance; and 4) insights into the retrieval latency of different models (see Section 5.5).

In the following section, we discuss the related work. Section 5.3 details the MMTweets dataset. Section 5.4 presents the various experimental details related to the X-DNR task. The results are presented in Section 5.5 and we conclude the paper in Section 5.6.

5.2 Related Work

Multiple multilingual datasets have been proposed in the literature for various fact-checking tasks such as evidence retrieval (Huang et al., 2022; Dementieva et al., 2023; Hammouchi and Ghogho, 2022; Köhler et al., 2022; Dementieva and Panchenko, 2021), claim classification (Gupta and Srikumar, 2021b; Huang et al., 2022; Nielsen and McConville, 2022; Zheng et al., 2022; Li et al., 2020b; Shahi and Nandini, 2020), and claim detection (Nakov et al., 2022a; Dutta et al., 2022; Alam et al., 2020; Panchendrarajan and Zubiaga, 2024). However, this paper specifically focuses on datasets for debunked narrative retrieval, which differs sufficiently from existing tasks, as demonstrated previously by Shaar et al. (2020a).

In order to minimise the spread of misinformation and speed up professional fact-checking, the initial verification step often involves searching for fact-checking articles that have already debunked similar narratives (Singh et al., 2023; La Gatta et al., 2023; Shaar et al., 2022, 2020a). Several benchmark datasets have been created for this task (Nakov et al., 2022c, 2021b; Shaar et al., 2021; Mansour et al., 2023, 2022; Sheng et al., 2021; Chakraborty et al., 2023; Kazemi et al., 2021, 2022; Pikuliak et al., 2023). For instance, Shaar et al. (2020a) release a dataset of English claims and fact-checking articles from Snopes (Snopes, 2024) and PolitiFact (PolitiFact, 2024). On the other hand, Vo and Lee (2020) release a multimodal English dataset of tweet claims collected from Snopes and PolitiFact and investigates the use of images in tweets to retrieve previously fact-checked content. The CLEF CheckThat! Lab evaluations (Shaar et al., 2020b; Nakov et al., 2021b, 2022c; Barrón-Cedeño et al., 2023) focus on a fully automated pipeline of fact-checking claims, where fact-checked claim retrieval is one of the steps in the claim verification workflow. They release a dataset of claims collected from Snopes, PolitiFact and AraFacts (Ali et al., 2021) and ClaimsKG (Tchechmedjiev et al., 2019). However, the aforementioned work only focuses on monolingual scenarios where the claim and debunk share the same language. In contrast, our MMTweets dataset includes cross-lingual cases, making it more challenging. For a detailed comparison of different datasets with our MMTweets, please refer to Section 5.3.4. We also test domain overlap between MMTweets and other datasets in Section 5.5.3.

Prior work on claim matching (Kazemi et al., 2021) release a dataset of claims collected from tiplines on WhatsApp and conduct retrieval experiments. Although they present results for multiple languages, their dataset only includes monolingual pairs (Kazemi et al., 2021), thereby hindering the development of retrieval models capable of detecting debunked narratives in multiple languages. Finally, the closest match to our work (Kazemi et al., 2022; Pikuliak et al., 2023) focuses on cross-lingual claim matching. They release a dataset of debunked tweets sourced from the International Fact-Checking Network (IFCN) (IFCN, 2024), Google Fact-check Explorer[1] and some other fact-checking aggregators (Kazemi et al., 2022; Pikuliak et al., 2023). However, Kazemi et al. (2022) dataset lacks diverse cross-lingual pairs (see Section 5.3.4), and claims are automatically extracted from debunk articles (Pikuliak et al., 2023), which can result in false positives. In contrast, our dataset has diverse cross-lingual pairs, and each tweet in MMTweets undergoes manual annotation to ensure high-quality data (see Section 5.3). Moreover, prior work (Kazemi et al., 2022) does not train custom debunked narrative retrieval models or perform cross-lingual and cross-dataset transfer testing, a gap that we address in this paper with a specific focus on the MMTweets dataset (Section 5.5).

Furthermore, Kazemi et al. (2021) found that multistage retrieval (Nogueira and Cho, 2019) using BM25 and XLM-RoBERTa transformer (Conneau et al., 2020) re-ranking can beat the competitive BM25 baseline for debunked narrative retrieval. However, the use of multistage retrieval with BM25 and transformer model re-ranking, as demonstrated in prior work (Kazemi et al., 2021; Shaar et al., 2020a; Nogueira and Cho, 2019; Thakur et al., 2021), introduces translation overhead for BM25 in cross-lingual scenarios where the query claim and document languages differ. To address this, this paper introduces translation-free multistage retrieval methods, employing both bi-encoders and cross-encoders for the X-DNR task (Section 5.4.1). Additionally, due to dataset limitation, much of the prior research (Shaar et al., 2020a; Nakov et al., 2021b, 2022c) trains retrieval models using debunks available from a single fact-checking organisation. In contrast, our MMTweets dataset involves debunks from multiple fact-checking organisations (Section 5.3). This enables the development of retrieval models that are agnostic to debunk structure, a crucial aspect for X-DNR, as relevant debunks can originate from any fact-checking organisation.

[1] https://toolbox.google.com/factcheck/explorer

5.3 MMTweets Dataset

MMTweets is a new dataset of misinformation tweets annotated with their corresponding debunks (or fact-checks), both available in multiple languages. MMTweets primarily comprises tweets related to COVID-19 misinformation in English, Hindi, Portuguese and Spanish. The languages of tweets were selected based on two criteria: 1) these are the most frequent languages in previous publicly available COVID-19 misinformation datasets (Li et al., 2020b; Singh et al., 2021a); 2) the chosen languages are among some of the most widely spoken ones worldwide. The dataset was built in two steps: first, the raw data was collected, followed by manual data annotation.

5.3.1 Raw Data Collection

First, we collect debunk narratives published by different fact-checking organisations covering our target languages. For this, we collect a total of 30, 452 debunk articles from the following organisations (language in brackets): Boomlive (English) (Boomlive, 2024), Agence France-Presse (AFP) (German, English, Arabic, French, Spanish, Portuguese, Indonesian, Catalan, Polish, Slovak and Czech) (, AFP), Agencia EFE (Spanish) (EFE, 2024) and Politifact (English) (PolitiFact, 2024). For each debunk article, we collect the following information fields: the article title, the debunked claim statement and the article body.

Next, we select a sample of 1, 600 debunk articles from the corpus of 30, 452 debunk articles based on two specific criteria. Firstly, we focus on debunks published between January 2020 and March 2021, allowing for temporal and topical diversity as the COVID-19 pandemic unfolded. This approach, given the global nature of the pandemic, maximises the chance of including similar narratives spreading in multiple languages. Secondly, our aim is to maximise instances where the language of the potential misinformation tweets mentioned in the debunk articles differs from that of the debunk article itself. For example, while Boomlive publishes debunk articles in English, the associated tweets may be in Hindi. Overall, this careful selection of debunks ensures comprehensive cross-lingual coverage within the MMTweets dataset (Section 5.3.5.1).

Finally, following the previous work (Shaar et al., 2020a; Kazemi et al., 2022), we extract all the tweets (in Hindi, English, Portuguese & Spanish language) mentioned in the debunk article body. We use Twitter API (API, 2024) to get tweet details including tweet text and attached media (if any). We chose Twitter because of its easy open access as compared to other social media platforms at the time of this study.

Table 5.2: *Details of the MMTweets dataset: class count, Fleiss Kappa and textual misinformation ratio. Please note that the class count does not sum up to the total tweet count due to the overlap between textual and non-textual misinformation cases.*

Language	Tweet Count	Class Count				Fleiss Kappa	Textual Misinformation Ratio
		Textual Misinformation	Non-textual Misinformation	Debunk	Other		
Hindi	400	328	254	11	27	0.53	0.86
Portuguese	400	310	200	5	30	0.59	0.77
English	400	247	166	68	82	0.79	0.61
Spanish	400	291	233	14	62	0.57	0.70
Total	1600	1176	853	98	201	Average: 0.62	Average: 0.74

5.3.2 Data Annotation – Tweet Classification

The approach described in Section 5.3.1 does not guarantee that the extracted tweets from debunk articles contain text-based misinformation. We found that some contained only images or videos, while others made general comments or debunked the misinformation itself. Therefore, the extracted tweets were classified manually to create gold-standard data for evaluation. In particular, we recruited 12 student volunteers[2] who were native speakers of either English, Hindi, Portuguese or Spanish (three native speakers per language). The annotators were shown tweets along with debunk information fields. Machine translation was used wherever the language of the debunk information field was different from that of the native speakers. Finally, the annotators were asked to classify the tweets into one of three classes:

- **Misinformation tweets:** with two sub-classes – **A) Textual misinformation**, if the textual part of a tweet expresses the false claim which is being debunked by the fact-checking article. **B) Non-textual misinformation**, if a tweet contains misinformation in image or video only. Please note that a tweet can have both text and non-textual misinformation. For such cases, annotators were asked to label the tweet as having both "textual misinformation" and "non-textual misinformation".
- **Debunk tweets:** If the tweet does not express misinformation uncritically, but instead exposes the falsehood of the claim.
- **Other tweets:** If the tweet is neither "misinformation" nor "debunk", then it is classified as "other". For instance, this can be a general comment or a general enquiry relevant to the false claim that is being debunked.

[2] The dataset annotation received ethical approval from the University of Sheffield Ethics Board (Application ID 040156). This paper only discusses analysis results in aggregate, without providing examples or information about individual users.

Please refer to the annotation codebook[3] for examples of misinformation, debunk, and other tweets. To ensure data quality, we first conducted training sessions with the annotators and went through several examples to familiarise them with the task. We also had a final adjudication step, where problems and disagreements flagged by the annotators were resolved by domain experts. For instance, there were some tweets which agreed with the misinformation but did not state it directly or the annotator was unsure about the claim's veracity. All such cases were considered "other" due to the chosen narrower definition of misinformation tweets.

A total of $1,600$ tweets were annotated, resulting in approximately 400 tweets per language (see Table 5.2). Following previous methodology (Sheng et al., 2022; Kazemi et al., 2022), a total of 400 tweets (100 per language) were triple annotated to compute inter-annotator agreement (IAA) and the final category was chosen by majority voting. Table 5.2 reports Fleiss Kappa scores which indicate moderate to substantial IAA for all languages. Table 5.2 also shows the textual misinformation ratio (i.e. the proportion of tweets annotated as "textual misinformation" out of all annotated tweets) for each language. The ratio is variable due to the varied nature of the debunks in each language and the different ways in which fact-checkers refer to misinformation-bearing tweets. On average, textual misinformation comprised 74% of all the classified tweets in the dataset.

5.3.3 Data Annotation – Claim Matching

The annotations gathered in Section 5.3.2 only pertain to tweets mentioned in the debunk articles, indicating a one-to-one relationship between tweets and debunks. However, prior research (Singh et al., 2021a, 2022) demonstrates that there can be various potential debunks for the same misinformation. To address this and establish a one-to-many relationship between misinformation tweets and debunks, we conduct a subsequent round of annotations to identify comparable debunks. However, annotating relevance judgments between tweets and all the previously collected 30,452 debunks is not feasible. Therefore, we take debunked claim statements linked to each tweet and compute cosine similarity[4] with all 30,452 debunked claim statements in the hope of finding similar debunked claim statements. To ensure this, we select the *top-k* matching claim statements for annotation, with a depth of seven as per previous work (Voorhees et al., 2021). We also retain only those claim pairs with a similarity score exceeding the 0.6 threshold to exclude irrelevant claim pairs from the annotations. We acknowledge that selecting claim pairs based on a specific model may lead to pooling bias and incomplete

[3] https://doi.org/10.5281/zenodo.10637161

[4] We use Sentence-transformer model *all-mpnet-base-v2* on English-translated statement. https://huggingface.co/sentence-transformers/all-mpnet-base-v2

relevance assessment. However, in a deliberate effort to address this concern, we employ the powerful general purpose Sentence-transformer model *all-mpnet-base-v2*, which has been trained on an extensive dataset of over 1 billion sentence pairs. It consistently achieves top scores across multiple datasets, ensuring robust and comprehensive results[5]. Finally, annotators were asked to classify 4,594 pairs of debunked claim statements into exact match, partial match, or irrelevant (3-level) using previously published annotation guidelines (Kazemi et al., 2021). Examples for each class can be found in the annotation codebook[6].

The annotations were conducted on the GATE Teamware annotation tool (Bontcheva et al., 2013) – refer to the annotation codebook for examples of the tool's user interface. A total of 14 PhD researchers were recruited to manually annotate pairs of debunked claim statements. To ensure high-quality annotations, we conducted a pre-annotation phase. An initial annotator training session familiarises them with the instructions. Subsequently, annotators were asked to annotate a certain number of test samples. We then review these annotations and only those annotators who correctly classify at least 80% of the samples proceed with further annotations. Based on prior research (Mu et al., 2023), we also ask annotators to provide a confidence score for each annotation, and we further discard annotations with low confidence scores to maintain data quality. Finally, following prior works (Voorhees et al., 2021; Hu et al., 2023; Bonisoli et al., 2023), we find the IAA Kappa to be 0.5 on a subset of the data using triple annotations, suggesting a moderate level of agreement among the annotators. All annotators were paid at a standard rate of 15 GBP per hour for their work.

Table 5.3 presents a summary of the complete MMTweets dataset, including the number of query tweets and the count of query tweet and debunk pairs for 3-level relevance annotations. Specifically, it includes 2,716 exact matches, 1,542 partial matches, and the remaining are categorised as irrelevant (see Section 5.3.5.1 for count in different language pairs). The average word count in query tweets is 28 ± 14.3 (1 std). There are a total of 1,600 tweets in MMTweets, and on average, each tweet is linked with 2.7 ± 2.0 (1 std) debunks, either exact or partial match. Please note that the one-to-many relation between query tweets and the debunks enriches our dataset to include cases beyond the tweets mentioned in the debunk articles. Additionally, the fine-grained classification of debunks into exact and partial matches serves as fine-grained labels for our subsequent information retrieval experiments (see Section 5.4.1).

[5] https://www.sbert.net/docs/pretrained_models.html

[6] https://doi.org/10.5281/zenodo.10637161

5.3.4 Comparison to Existing Datasets

Table 5.4 provides a comparison between MMTweets and the existing datasets, revealing favourable query claim counts in our dataset compared to the majority of other existing datasets. Notably, MMTweets stands out with 43% cross-lingual instances across various language pairs (see Section 5.3.5.1). This is in stark contrast to the cross-lingual dataset by Kazemi et al. (2022), which only comprises 10% of Hindi-English pairs, where the claim is in Hindi and the debunk is in English. On the other hand, the recently released big MultiClaim (Pikuliak et al., 2023) has 13% cross-lingual instances; however, their dataset lacks fine-grained labels and images for multimodal detection. Additionally, all tweets in MMTweets undergo manual annotation to produce a gold-standard dataset, unlike other existing datasets (Hardalov et al., 2022; Pikuliak et al., 2023; Kazemi et al., 2022; Shaar et al., 2020a), where social media posts are automatically extracted from fact-check articles, potentially leading to false positives. Moreover, automated extraction of tweets also leads to missing one-to-many connections between claims and debunks as shown in prior work (Singh et al., 2023). Furthermore, MMTweets provides 3-level graded relevance scores (fine-grained) for query-passage pairs, unlike prior work which uses binary relevance scores (coarse-grained) (Shaar et al., 2020a; Nakov et al., 2022c, 2021b; Pikuliak et al., 2023).

Among other datasets, Shaar et al. (2020a) and CLEF variants lack cross-lingual pairs, images, and fine-grained labels. The larger Vo and Lee (2020) dataset incorporates images and human annotations but lacks fine-grained labels. CrowdChecked (Hardalov et al., 2022) contains a massive volume of claims but lacks crucial features like manual annotations and cross-lingual pairs. Although prior work (Kazemi et al., 2022, 2021) provide multilingual support, it's impossible to replicate or conduct comparative experiments on their datasets because they do not release the corpora of debunks used in the retrieval experiments – only the query claims are released. Moreover, it lacks images, manual annotations and fine-grained labels (Kazemi et al., 2022). In contrast, our MMTweets dataset stands out, featuring cross-lingual pairs, images, human annotations, and fine-grained labels, making it a comprehensive resource compared to its counterparts. Additionally, we examine the domain overlap between MMTweets and other datasets, revealing a low degree of overlap (refer to Section 5.5.3).

Table 5.3: *Complete summary of the MMTweets dataset.*

Language	Hindi	Portuguese	English	Spanish	Total
Query Tweets	400	400	400	400	1600
Exact Match	518	742	812	644	2716
Partial Match	417	409	342	374	1542
Irrelevant	475	656	337	468	1936

Table 5.4: *Comparison of debunked narrative retrieval datasets: "Lang" denotes the count of different languages of claims; "Cross" indicates the presence of cross-lingual pairs; "Img" indicates whether the dataset is multi-modal and includes images; "Anot" indicates whether the dataset is human-annotated and is gold-standard; "Fine" indicates the availability of fine-grained labels.*

Dataset	Items	Lang	Cross	Img	Anot	Fine
Shaar et al. (2020a)	1,768	1	×	×	×	×
CLEF20-EN	1,197	1	×	×	✓	×
CLEF21 2A-EN	2,070	1	×	×	×	×
CLEF21 2A-AR	858	1	×	×	✓	×
CLEF22 2A-EN	2,362	1	×	×	×	×
CLEF22 2A-AR	908	1	×	×	✓	×
Vo and Lee (2020)	13,239	1	×	✓	✓	×
CrowdChecked (Hardalov et al., 2022)	330,000	1	×	×	×	×
Kazemi et al. (2021)	382	5	×	×	✓	✓
Kazemi et al. (2022)	6,533	4	✓	×	×	×
MultiClaim (Pikuliak et al., 2023)	31,305	27	✓	×	×	×
MMTweets (ours)	1,600	4	✓	✓	✓	✓

Table 5.5: *Language of tweet and debunk pairs in MMTweets. Language codes are ISO 639-1 representations for Portuguese (PT), Spanish (ES), Hindi (HI), English (EN), Indonesian (ID), Slovak (SK), Catalan (CA), Polish (PL), Czech (CS), and French (FR).*

Tweet Language	PT	ES	HI	EN	EN	EN	PT	EN	EN	ES	EN	EN	EN	PT	EN	ES	HI	PT	HI	ES	Total
Debunk Language	PT	ES	EN	EN	ES	ID	ES	PT	SK	CA	PL	CA	CS	ID	FR	ID	PT	EN	FR	EN	
Count	1045	954	925	450	332	158	80	65	53	50	30	27	22	22	17	11	7	4	3	3	4,258

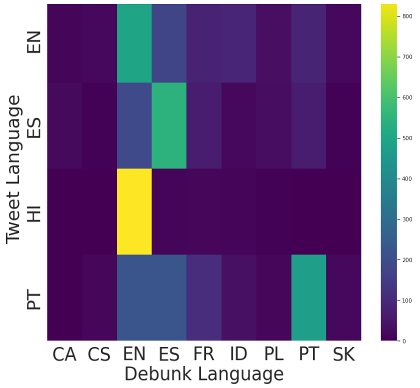


Figure 5.2: *Cross-language analysis: tweet vs. debunk.*

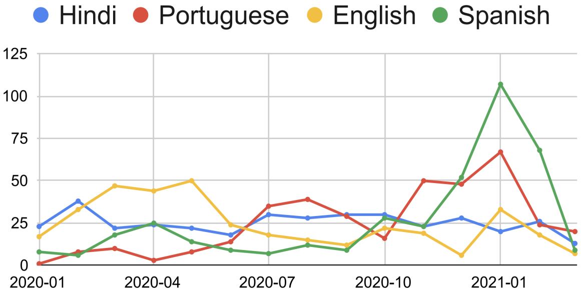


Figure 5.3: *Line plot for month-by-month breakdown of tweet counts for each language in the MMTweets dataset.*

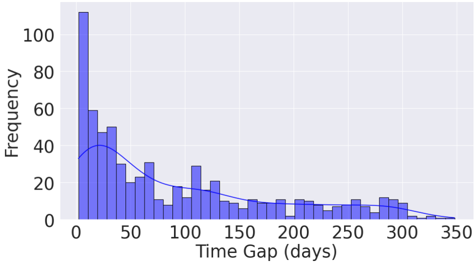


Figure 5.4: *Time gap between tweet and debunk.*

5.3.5 Data Analysis

5.3.5.1 Linguistic Diversity

Table 5.5 shows the count of query tweet and debunk pairs for different languages[7]. In particular, there are a total of 4,258 positive pairs (exact and partial matches) of tweets and their corresponding debunks. Among these, 1,809 instances (43%) are pairs where the language of tweets and debunks is different (cross-lingual). This makes our dataset the one with the highest proportion of cross-lingual instances when compared to existing datasets (see Section 5.3.4). The majority of these cross-lingual pairs have tweets in Hindi and corresponding debunks in English, followed by instances with tweets in English and debunks in Spanish.

Figure 5.2 displays the heatmap illustrating language dynamics of tweets and its related debunks in MMTweets. Notably, multiple languages exhibit near-zero associated debunks in languages besides English (e.g., Hindi), suggesting a potential gap in fact-checking coverage for specific languages. This emphasises the need to address disparities in debunk distribution and highlights opportunities for automated cross-language fact-checking methods like X-DNR.

5.3.5.2 Temporal Diversity

To assess dataset diversity, we also analyse the temporal characteristics of tweets in Figure 5.3, presenting a month-by-month breakdown of tweet counts for each language in MMTweets. We observe that Hindi and English tweets exhibit a relatively even distribution from Jan 2020 to Mar 2021. Conversely, Portuguese and Spanish tweets show a more concentrated presence, primarily emerging in late 2020 and early 2021. It's important to note that the MMTweets dataset encompasses at least one tweet for each month from Jan 2020 to Mar 2021 (spanning 15 months). Overall, we find the tweets to be temporally diverse across languages.

In examining cases where debunking precedes misinformation tweets (22.3% of cases), Figure 5.4 illustrates publication date gaps. With a median gap of 76 days, the findings reveal misinformation can persist even after relevant debunks are available. For instance, one of the false tweets about "Bill Gates launching implantable chips to track COVID-19," appeared in English on Twitter on 3 July 2020, while the earliest related debunk available was published on 13 May 2020 (, AFP) in the French language (49 days gap). This emphasises the need for effective methods, such as X-DNR, to detect the spread of already debunked narratives in multiple languages.

[7] We use *langdetect* (`https://pypi.org/project/langdetect/`) for detecting the language.

Table 5.6: *Topics captured by Latent Dirichlet Allocation.*

Language	Latent Dirichlet Allocation Topics
Hindi	Topic 1: hindu, delhi, corona, farmer, government
Hindi	Topic 2: going, people, world, temple, muslim
Hindi	Topic 3: massive, please, wipe, foreign, affair
Portuguese	Topic 1: people, bolsonaro, vaccine, world, covid
Portuguese	Topic 2: vaccine, work, mask, vote, without
Portuguese	Topic 3: vaccine, world, minister, took, abortion
English	Topic 1: deployed, mask, time, corona, epidemic
English	Topic 2: coronavirus, wuhan, like, china, year
English	Topic 3: people, work, coronavirus, hospital, covid
Spanish	Topic 1 : people, without, vaccine, go, mask
Spanish	Topic 2: day, say, first, government, usa
Spanish	Topic 3: vaccine, died, nurse, covid, netherlands

5.3.5.3 Domain Diversity

Table 5.6 presents the results of topic modelling using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), showcasing the top three topics for each tweet language in MMTweets. As expected, the topics related to coronavirus are apparent in all four languages. However, some topics are specific to events in the country where the language is spoken. In Hindi, the first topic appears to focus on a combination of religious and political elements. For instance, words such as “farmer” and “Delhi” are related to the misinformation that spread during the farmers’ protest in Delhi, India (Wikipedia, 2024). Similarly, in Portuguese, the dominant topics revolve around President Bolsonaro and vaccines. The topics related to “vaccine” are dominant in both Portuguese and Spanish tweets which is likely because the tweets for these languages are mainly from the end of 2020 (see Figure 5.3), when vaccine-related information was at its peak (Yousefinaghani et al., 2021). English topics cover diverse aspects, including misinformation related to the origin of COVID-19 and its impact on people and hospitals. Overall, the table provides insights into the diverse and multifaceted nature of claims related to COVID-19 in MMTweets.

5.4 Cross-lingual Debunked Narrative Retrieval (X-DNR)

In this section, we formally define the X-DNR task. Given a tweet claim as a query t, the X-DNR system employs a retrieval model to obtain a candidate set of debunked narratives from a larger corpus of debunks $D = \{d_i\}_{i=1}^{D}$ in multiple languages. The final trained model can be expressed as X-DNR(t, D), whose ultimate goal is to provide the most accurate fact-checking information to users in response to potential misinformation claims in any language.

In this paper, we exclusively focus on textual misinformation cases (totalling 1,176, as shown in Table 5.2). For the retrieval corpus, we utilise a collection of $30,452$ previously gathered debunks in multiple languages (refer to Section 5.3.1). Each debunk comprises a concatenated debunked claim and article title field (Section 5.3.1).

5.4.1 Cross-lingual Retrieval Models

We test the following cross-lingual retrieval models on MMTweets.

Okapi BM25. We utilise the ElasticSearch (Gormley and Tong, 2015) implementation of BM25 (Gormley and Tong, 2015) with default parameters ($k = 1.2$ and $b = 0.75$). Since BM25 is designed for monolingual retrieval, we employ machine translation using the Fairseq's `m2m100_418M` model (Fan et al., 2021) to make it applicable to cross-lingual query and document pairs. All non-English tweets and debunks are translated into English, and the complete corpus of debunks is indexed in ElasticSearch (Gormley and Tong, 2015). We then use the English-translated tweets as queries over the debunks.

xDPR Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), an early dense retrieval model, uses BERT-based encoders for queries and documents to assess relevance based on their similarity. To expand its support beyond English, we use a multilingual variant, xDPR (Yang et al., 2022; Huggingface, 2024a), which is an XLM-RoBERTa (Conneau et al., 2020) model fine-tuned on the MSMARCO dataset (Nguyen et al., 2016). We further fine-tune xDPR on our MMTweets (Yang et al., 2022).

mContriever Izacard et al. (2022) introduced mContriever, which employs contrastive loss for unsupervised pretraining of mBERT (Devlin et al., 2019), showing enhanced performance on various IR tasks. We use the provided multilingual checkpoint (Huggingface, 2024b), already fine-tuned on MSMARCO (Nguyen et al., 2016). We further fine-tune this model on MMTweets, employing the same methodology as described in Izacard et al. (2022).

Bi-Encoder (BE) We fine-tune different Multilingual Pretrained Transformer (MPT) models as bi-encoders (Reimers and Gurevych, 2019; Karpukhin et al., 2020) on pairs of query tweets and their corresponding debunks. The objective function employed is the mean squared error, measuring the disparity between the true label and the model-calculated relevance score for the tweet-debunk pair. This adjusts model parameters, aligning the embedding of a query tweet closer to its relevant debunks in the vector space. The loss equation is as follows,

$$\mathcal{L}(\theta) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left(\mathcal{Y}_i - \left(\frac{f_\theta(t_i) \cdot f_\theta(d_i)}{\|f_\theta(t_i)\|_2 \|f_\theta(d_i)\|_2} \right) \right)^2 \tag{5.1}$$

where f_θ is the shared MPT encoder for tweet t_i and debunk d_i, $\mathcal{Y}_i$ represents the true label of the i-th sample. The relevance score between tweet and debunk is computed using cosine similarity. We employ cosine similarity with the mean-pooling technique due to its proven effectiveness in prior research (Reimers and Gurevych, 2019).

We fine-tune bi-encoder using five different MPT models, namely multilingual BERT (mBERT) (Devlin et al., 2019), XLM-RoBERTa (XLMR) (Conneau et al., 2020) and Language-Agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022). Additionally, we also fine-tune two Sentence-Transformer model variants i.e. Universal Sentence Encoder (USE) (Huggingface, 2024d; Yang et al., 2020) and Masked and Permuted Pretraining for Language Understanding (MPNet) (Huggingface, 2024c; Song et al., 2020). These bi-encoder models are denoted by the prefix "BE-" in subsequent experiments (Section 5.5.1).

Multistage Retrieval Drawing inspiration from the success of multistage retrieval methods in IR tasks (Nogueira and Cho, 2019; Thakur et al., 2021; Singh et al., 2021b), we apply these techniques to the X-DNR task. Within this context, we introduce two methods that adapt earlier approaches, specifically tailored for the X-DNR task. These methods are as follows:

- **Bi-Encoder+Cross-Encoder (*BE+CE*):** In the first retrieval stage, we fine-tune an MPT model as a bi-encoder instead of the standard BM25-based lexical retrieval approach adopted in prior work (Shaar et al., 2020a; Kazemi et al., 2021). This choice is motivated by the MPT model's suitability for the cross-lingual nature of the task, eliminating the need for translation. In the second stage, we fine-tune an MPT model as a cross-encoder (Nogueira and Cho, 2019) to re-rank the top-K retrieved debunks from the first stage. Here, the model employs self-attention mechanisms on the given tweet and debunk pair to get the final relevance score. The input to the model follows the structure: $[CLS]$ $[T_1]...[T_n]$ $[SEP]$ $[DC_1]...[DC_i][DT_1]...[DT_j]$, where T_n are the tweet subword tokens and DC_i and DT_j are the debunked claim and title subword tokens, respectively. $[CLS]$ and $[SEP]$ are the default tokens to indicate "start of input” and “separator”, respectively, in the Next Sentence Prediction task (Devlin et al., 2019).

- **Bi-Encoder+ChatGPT (*BE+GPT3.5*):** Large language models like ChatGPT (*gpt-3.5-turbo*) have consistently showcased impressive capabilities across a broad spectrum of natural language processing tasks (AI, 2024). Therefore, to evaluate ChatGPT’s performance, we implement a Listwise Re-ranker with a Large Language Model (LRL) (Ma et al., 2023) to re-rank documents retrieved by the first stage ranker. The main distinctions in our approach compared to prior work (Ma et al., 2023) are: 1) we employ multilingual bi-encoders described in Section 5.4.1 as the first-stage ranker 2) each re-ranked document consists of concatenated debunk claim and title fields. Besides this, all parameters are kept same as used by Ma et al. (2023).

5.4.2 Experimental Details

5.4.2.1 Train and test sets

We divide 1,176 textual misinformation tweet queries into train and test sets. The test set consists of 400 tweet queries (100 queries per language), comprising the same triple-annotated tweets used for calculating IAA (Section 5.3.5). The remaining 776 tweet queries are used as training data, with a 10% subset used as a validation set. Please note that during test time, we do not know if a tweet has been debunked, because tweets linked with debunks in the test set do not occur in the train set. This ensures a realistic test scenario by preventing tweets linked to the same debunk from appearing in both the train and test sets.

Now, since each query tweet in the training set is linked to multiple debunks (Section 5.3.3), therefore, the final training set comprises 2,360 positive (1,420 exact matches and 940 partial matches) tweet and debunk pairs. For negative pairs, ten debunks are randomly sampled for each tweet, resulting in a total of 7,760 negative tweet and debunk pairs. For a comprehensive analysis of various methods used for getting negative tweet and debunk pairs, please refer to Section 5.5.4. We also experimented with hard negative mining and higher counts of negatives, but did not observe any significant improvements. In total, the training set consists of 10,120 fine-grained tweet and debunk pairs for training different retrieval models.

5.4.2.2 Evaluation Metrics

We employ two widely used ranking metrics (Nakov et al., 2021b, 2022c) for evaluation: Mean Reciprocal Rank (**MRR**) and Normalised Discounted Cumulative Gain (**nDCG@1** & **nDCG@5**). MRR measures the effectiveness of the system by computing the score based on the highest-ranked relevant debunk for each misinformation tweet. MRR is defined as MRR $= \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\frac{1}{rank_i}$, where $|\mathcal{T}|$ is the

Table 5.7: *Results for different cross-lingual retrieval models on the test set of MMTweets. The best scores are in bold.*

Language	Metric	BM25	xDPR	mCont	BE-mBERT	BE-XLMR	BE-USE	BE-LaBSE	BE-MPNet	BE+CE	BE+GPT3.5
MMTweets-HI	**nDCG@1**	0.263	0.435	0.240	0.135	0.160	0.210	0.525	0.320	**0.610**	0.575
	nDCG@5	0.267	0.421	0.304	0.149	0.188	0.246	0.514	0.366	**0.569**	0.527
	MRR	0.320	0.503	0.352	0.199	0.250	0.310	0.623	0.439	**0.674**	0.637
MMTweets-PT	**nDCG@1**	0.625	0.695	0.770	0.540	0.685	0.730	0.755	0.755	**0.845**	0.840
	nDCG@5	0.598	0.690	0.761	0.514	0.595	0.672	0.726	0.720	**0.765**	0.757
	MRR	0.723	0.781	0.849	0.627	0.737	0.782	0.822	0.821	**0.887**	0.880
MMTweets-EN	**nDCG@1**	0.591	0.635	0.705	0.515	0.465	0.675	0.680	0.710	**0.720**	0.715
	nDCG@5	0.572	0.625	0.670	0.475	0.472	0.638	0.650	**0.696**	0.682	0.662
	MRR	0.706	0.759	0.801	0.603	0.590	0.760	0.780	**0.814**	**0.814**	0.807
MMTweets-ES	**nDCG@1**	0.560	0.620	0.610	0.405	0.435	0.500	0.585	0.615	**0.735**	0.660
	nDCG@5	0.525	0.621	0.646	0.394	0.428	0.497	0.582	0.582	**0.662**	0.632
	MRR	0.648	0.707	0.730	0.491	0.536	0.591	0.670	0.678	**0.804**	0.741
Average	**nDCG@1**	0.510	0.596	0.581	0.399	0.436	0.529	0.636	0.600	**0.728**	0.698
	nDCG@5	0.490	0.589	0.595	0.383	0.421	0.513	0.618	0.591	**0.669**	0.644
	MRR	0.599	0.687	0.683	0.480	0.528	0.611	0.724	0.688	**0.795**	0.766

number of tweets used as a query and $rank_i$ is the rank of the top relevant debunk for the *ith* tweet. On the other hand, the nDCG@K normalises DCG@K by dividing it by ideal DCG@K, where DCG@K discounts the graded relevance value of retrieved debunks based on their ranks. DCG@K is defined as follows,

$$DCG@K = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{K}|}\frac{2^{rel_{i,k}}-1}{\log_2(rank_{i,k}+1)}, \tag{5.2}$$

where $rel_{i,k}$ is the graded relevance of the debunk at $rank_{i,k}$ for the ith query tweet. Higher MRR and nDCG scores indicate better performance.

5.4.2.3 Hyperparameters

The bi-encoder is trained for four epochs with a batch size of 32, a learning rate of $4e-5$ and maximal input sequence length of 256. The cross-encoder, trained for two epochs, uses a batch size of 16, $4e-5$ learning rate, with truncation of subword tokens beyond 512. Both models employ linear warmup, AdamW optimiser, and manual hyperparameter tuning on a validation set. Hyperparameter bounds are set as: 1) 1 to 5 epoch 2) $1e-5$ to $5e-5$ learning rate 3) 8 to 64 batch size on NVIDIA RTX 3090.

5.5 Results and Discussion

In this section, we present the results of retrieval experiments that aim to address the following five research questions:

RQ1 To what extent do the current SOTA cross-lingual retrieval models perform in addressing the specific challenges posed by the MMTweets dataset? (Section 5.5.1)

RQ2 How challenging is it for models to transfer and generalise across languages within MMTweets? (Section 5.5.2)

RQ3 Can models trained on existing datasets transfer knowledge and generalise on the MMTweets test set? (Section 5.5.3)

RQ4 How does the type and count of negative pairs impact the model's performance on MMTweets? (Section 5.5.4)

RQ5 What insights can be gained into the retrieval latency of various cross-lingual retrieval models? (Section 5.5.5)

5.5.1 Model Performance

Table 5.7 shows Mean Reciprocal Rank (MRR) and Normalised Discounted Cumulative Gain (nDCG@1 & nDCG@5) on the test set of MMTweets (HI, PT, EN & ES). The results suggest that BE-mBERT and BE-XLMR consistently show lower scores, with occasional lower performance when compared to BM25. BM25's strength lies in lexical overlap with machine-translated text, giving it an advantage over other models. However, other retrieval models outperform BM25 on several metrics. Notably, BE-LaBSE performs better than BE-MPNet, BE-USE, BE-mBERT, and BE-XLMR, even outperforming state-of-the-art models like xDPR and mContriever in average metric scores. This is attributed to LaBSE's sentence-level objective, combined with pretraining techniques involving translation and masked language modelling, as discussed in Feng et al. (2022).

The last two columns of Table 5.7 report the scores of multistage retrieval methods (*BE+CE* & *BE+GPT3.5*). In multistage retrieval, we employ LaBSE for the first stage due to its superior performance over other models (see Table 5.7). Similarly, the second stage in *BE+CE* also utilises LaBSE, with the number of re-ranked documents set to 20. Although we experimented with various MPT models and different counts of re-ranked documents in the second stage, no significant improvements were observed. The results show that *BE+CE* consistently emerges as the top performer across all datasets and metrics, achieving an average nDCG@1 score of 0.728, an average nDCG@5 score of 0.669, and an average MRR score of 0.795 (Table 5.7 – second last column). On the other hand, while *BE+GPT3.5* outperforms other models in average metric scores, its retrieval latency is the highest (see section 5.5.5). Although other models like BE-LaBSE, BE-MPNet, xDPR, and mContriever showcase competitive performance, none consistently match the performance demonstrated by multistage retrieval methods. Additionally, despite being trained on the extensive MSMARCO training dataset (Section 5.4.1), models such as xDPR and mContriever do not notably enhance performance, suggesting distinctive challenges presented by MMTweets.

For *BE+CE*, the extent of improvement varies across languages. For example,

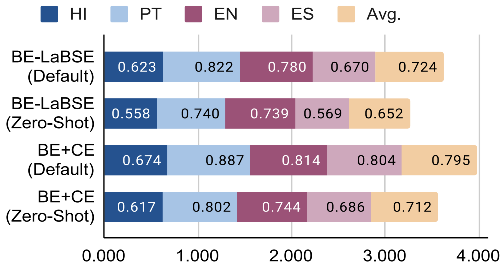


Figure 5.5: *Stacked bar plot for MRR scores for zero-shot cross-lingual transfer and the default results (from Table 5.7).*

in the case of Portuguese, *BE+CE* outperforms BM25 with increases of 132% for nDCG@1, 112% for nDCG@5, and 110% for MRR. Conversely, the improvement is relatively low for English, with increases of only 22%, 19%, and 15% for nDCG@1, nDCG@5, and MRR scores, respectively. We hypothesise that this disparity in performance across different languages may be attributed to noisy translations in the case of BM25, while *BE+CE* doesn't rely on translation. Additionally, the scores on different languages are reflective of how topics found in each language impact a model's performance. For example, the Hindi tweets have the lowest performance across all models and evaluation metrics, which suggests that the topics found in these languages (Section 5.3.5) are quite challenging for the model. Another reason for poor Hindi performance could be the change in the language script to Devanagari. This suggests that dealing with various languages in MMTweets poses a challenge, and there is still potential for improvement in retrieval models.

Furthermore, we also observe the challenge of distinguishing closely related debunks by the model. This occurs when the retrieved debunk is not entirely relevant, but still shares some degree of relevance with the query claim. For instance, consider the query claim about the sighting of crocodiles in the flooded streets of Hyderabad; the top-retrieved debunks are closely related, involving sightings of crocodiles in Mumbai, Bengaluru, Florida, etc. This highlights the need for continued refinement in retrieval models to enhance the relevance of top-ranked debunks for the X-DNR task.

In summary, these evaluations highlight performance differences among models, emphasising the consistent superiority of multistage retrieval methods across various languages and metrics. While BM25 is faster (see Section 5.5.5), the necessity of machine translation for BM25 incurs additional costs and time overheads.

Table 5.8: *Domain overlap between the test set of MMTweets and the train set of other datasets.*

Train Set	MMTweets	Snopes	CLEF 20-EN	CLEF 21-EN	CLEF 21-AR	CLEF 22-EN	CLEF 22-AR
Overalp	0.29	0.15	0.14	0.12	0.16	0.11	0.13

5.5.2 Cross-lingual Transfer

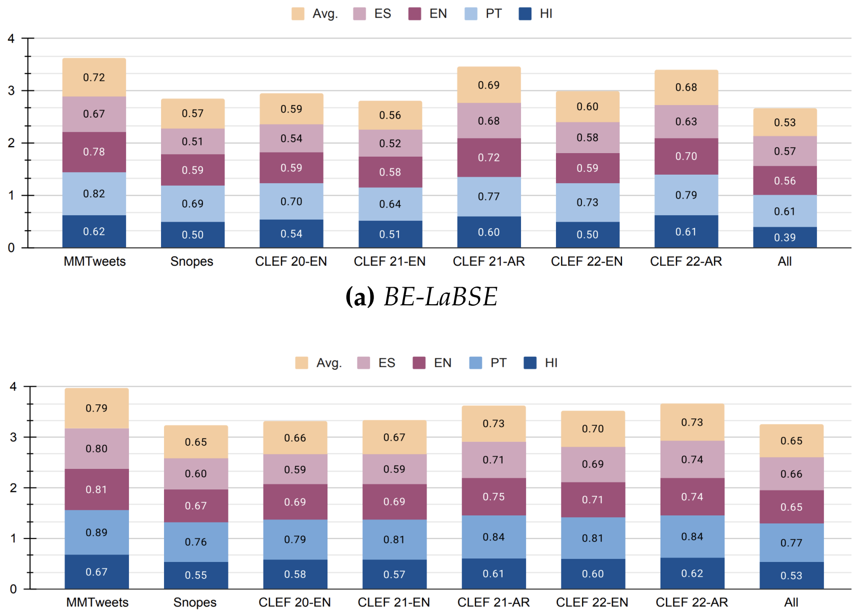
To test the zero-shot transfer capabilities, the model is trained on languages other than the one it is tested on. For instance, to test zero-shot transfer for Hindi, the models are trained on only those tweet and debunk pairs that are not in Hindi. Hence, in total four models are trained for four different languages in the MMTweets.

We evaluate the cross-lingual transfer capability of BE-LaBSE and *BE+CE*, which yield the highest average scores (Table 5.7). Figure 5.5 shows a stacked bar plot illustrating MRR scores for zero-shot cross-lingual transfer and the default results sourced from Table 5.7.

When comparing the zero-shot results with the default results, the default results consistently outperform zero-shot results for both models (BE-LaBSE and *BE+CE*) across all languages, as expected due to training on the complete dataset. Nevertheless, zero-shot models surpass several baselines, including BM25 (from Table 5.7) in this challenging setting. The results suggest that models have the potential to transfer knowledge between languages without the need for language-specific training. This also supports prior observations that MPT models, when fine-tuned on monolingual data, exhibit strong performance on a different language (Izacard et al., 2022; Conneau et al., 2018). Despite these promising outcomes, there is still room for improvement for zero-shot models to match the performance of default models.

5.5.3 Cross-dataset Transfer

To test the zero-shot cross-dataset transfer capabilities of the models, we train them on the training set of previously published datasets and subsequently evaluate their performance on the test set of MMTweets. This ensures real-life testing to assess the generalisability of the models. The previously published datasets include Snopes (Shaar et al., 2020a) and CLEF CheckThat! Lab task datasets which include CLEF 22-EN and CLEF 22-AR (Nakov et al., 2022c), CLEF 21-EN and CLEF 21-AR, (Nakov et al., 2021b) and CLEF 20-EN (Shaar et al., 2020b). Please note that CLEF 22-AR and CLEF 21-AR are Arabic datasets while other datasets are in English. In addition, we machine translate all the claims in previously published datasets into the languages of claims in MMTweets (HI, EN, PT, ES) and train the



(a) *BE-LaBSE*

(b) *BE+CE*

Figure 5.6: *Stacked bar plot for MRR scores for zero-shot cross-dataset transfer using BE-LaBSE (a) and BE+CE (b).*

retrieval model on the combined datasets (represented as *All* hereafter). This is to determine if translating existing monolingual datasets into multiple languages can help achieve reliable results on the cross-lingual MMTweets test set.

First, we assess the domain overlap to see how challenging it is for models trained on existing datasets to transfer knowledge to the MMTweets test set. For this, we use weighted Jaccard similarity (Ioffe, 2010) to compute the domain overlap between the test set of MMTweets and the train set of other datasets used for cross-dataset analysis (Table 5.8). We also report the overlap between the train and test set of MMTweets for reference. We find low domain overlap (ranging from 11-16%) with other datasets' train sets compared to MMTweets' train set (which has a 29% overlap) indicating distinct or less common instances between MMTweets and other datasets. We also conducted this analysis for each language but didn't find much variation in the results. Overall, MMTweets stands out as a unique dataset, showing low domain overlap with existing datasets.

Figure 5.6 shows MRR scores for zero-shot cross-dataset transfer using BE-LaBSE (a) and *BE+CE* (b), alongside default MMTweets trained results (from Table 5.7). Notably, models trained on CLEF 21-AR and CLEF 22-AR, despite being in Arabic, achieve the highest scores across all languages after the default MMTweets trained models. Additionally, models fine-tuned on CLEF 22-EN and 20-EN closely com-

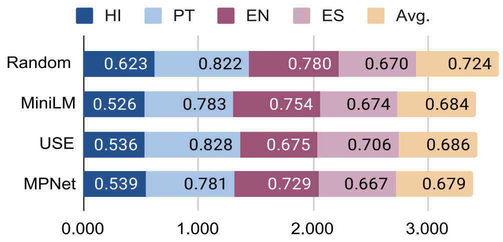


Figure 5.7: *Stacked bar plot illustrating MRR scores for different methods to get negative claim and debunk pairs.*

pete with other retrieval models (Table 5.7). Notably, while all claims in other datasets are either in English or Arabic, the MMTweets test set encompasses multiple other languages, making it even more challenging to retrieve the best matching debunk. Finally, we find that translating existing monolingual datasets into multiple languages and training models on them (represented as *All* in Figure 5.6) leads to degraded results on the cross-lingual MMTweets test set. This degradation in performance can be attributed to the noisy automatic machine translations.

Overall, the findings suggest some knowledge transfer between datasets, which is especially valuable when obtaining a domain-specific dataset for training a dedicated model is challenging. However, despite these positive outcomes, there remains potential for models to match or surpass default MMTweets trained results.

5.5.4 Negative Claim and Debunk Pairs

To train retrieval models, we employ two methods for selecting negative debunks: hard negatives using MPT models (*paraphrase-multilingual-MiniLM-L12-v2*, *distiluse-base-multilingual-cased-v2*, *paraphrase-multilingual-mpnet-base-v2*), and random sampling. Evaluation is limited to BE-LaBSE due to the practical limitations of training using different methods. Figure 5.7 shows MRR scores for different methods. While the USE method excels notably in certain languages, the random sampling method yields the highest average MRR score of 0.724. This suggests random sampling as a reliable method for the X-DNR task.

Figure 5.8 illustrates the impact of varying counts of random negatives used during training. Increasing negative counts (N=5 to N=80) shows fluctuating performance across languages. For instance, Portuguese MRR increases with more negatives, peaking at 0.864 at N=80. Conversely, Hindi's performance peaks at N=10, with little improvement beyond. Generally, higher counts of negatives do not lead to noticeable improvements.

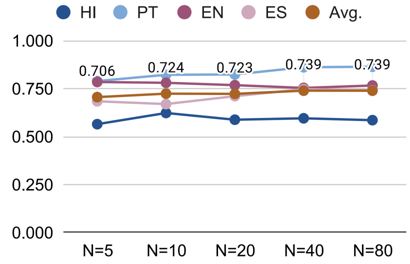


Figure 5.8: *Line plot for the impact of varying counts of random negatives on model performance. Average scores are labelled.*

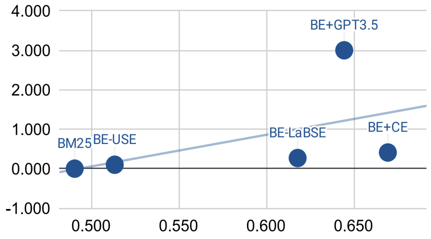


Figure 5.9: *Scatter plot for retrieval latency (in seconds) and MRR scores for various models.*

5.5.5 Retrieval Latency

Figure 5.9 shows the average MRR scores and retrieval latency for different models. Lower values in the retrieval latency indicate faster query processing by the IR model. When comparing models, *BE+CE* achieves the highest MRR score (0.669) but exhibits a latency of 0.41 seconds, indicating a comparatively longer retrieval time. BE-LaBSE follows closely with an MRR score of 0.618 and a moderate retrieval latency of 0.27 seconds, striking a balance between performance and retrieval speed. While *BE+GPT3.5* displays a competitive MRR score (0.644), its retrieval latency increases to 3 seconds, impacting its practical application in real-time scenarios. BM25, although has the fastest retrieval latency at 0.001 seconds, it compromises ranking quality with the lowest MRR score of 0.490.

Overall, BE-LaBSE provides a balanced option with reasonable performance and moderate retrieval latency, while *BE+CE* excels in ranking quality, albeit with a slightly longer retrieval latency.

5.6 Conclusion and Future Work

This paper focuses on cross-lingual debunked narrative retrieval (X-DNR) for automated fact-checking. It introduces MMTweets, a novel benchmark dataset that stands out, featuring cross-lingual pairs, human annotations, fine-grained labels, and images, making it a comprehensive resource compared to other datasets. Furthermore, initial tests benchmarking SOTA cross-lingual retrieval models reveal that dealing with multiple languages in the MMTweets dataset poses a challenge, indicating a need for further improvement in retrieval models. Nevertheless, the introduction of tailored multistage retrieval methods demonstrates superior performance over other SOTA models, achieving an average nDCG@5 of 0.669. However, it's crucial to note the trade-offs between model performance and retrieval latency, with *BE+CE* offering better ranking quality at the expense of longer retrieval times. Finally, the findings also suggest some knowledge transfer across languages and datasets, which is especially valuable in scenarios where language-specific models are not available or feasible to train. However, despite these positive outcomes, there is still room for models to match or even surpass the performance of default MMTweets trained models. To achieve this objective, the model needs an in-depth understanding of both language and context, along with the capability to differentiate among closely related debunked narratives. More sophisticated models could potentially introduce these capabilities in the future.

In future, we plan to extend the dataset to include claims from other social media platforms and domains to enhance its generalisability. Additionally, we aim to explore multimodal debunked narrative retrieval, leveraging information from various modalities.

5.7 Acknowledgements

This research is supported by a UKRI grant EP/W011212/1, an EU Horizon 2020 grant (agreement no.871042) ("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)) and a European Union grant INEA/CEF/ICT/A2020/2381686 (European Digital Media Observatory (EDMO) Ireland, https://edmohub.ie/).

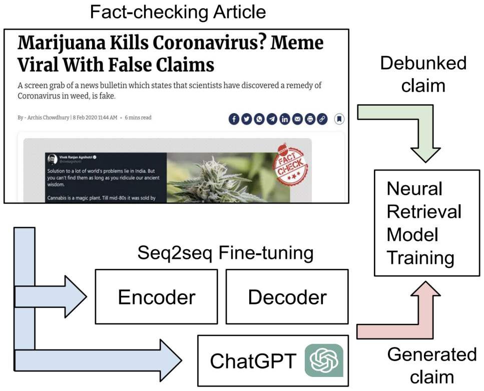
Chapter 6

Unsupervised Method for Training Debunked Narrative Retrieval Models

Iknoor Singh, Carolina Scarton and Kalina Bontcheva
Department of Computer Science, The University of Sheffield, UK

Abstract

A key task in the fact-checking workflow is to establish whether the claim under investigation has already been debunked or fact-checked before. This is essentially a retrieval task where a misinformation claim is used as a query to retrieve from a corpus of debunks. Prior debunk retrieval methods have typically been trained on annotated pairs of misinformation claims and debunks. The novelty of this paper is an Unsupervised Method for Training Debunked-Narrative Retrieval Models (`UTDRM`) in a zero-shot setting, eliminating the need for human-annotated pairs. This approach leverages fact-checking articles for the generation of synthetic claims and employs a neural retrieval model for training. Our experiments show that `UTDRM` tends to match or exceed the performance of state-of-the-art methods on seven datasets, which demonstrates its effectiveness and broad applicability. The paper also analyses the impact of various factors on `UTDRM`'s performance, such as the quantity of fact-checking articles utilised, the number of synthetically generated claims employed, the proposed *entity inoculation* method, and the usage of large language models for retrieval.



Marijuana Kills Coronavirus? Meme Viral With False Claims

A screen grab of a news bulletin which states that scientists have discovered a remedy of Coronavirus in weed, is fake.

By - Archis Chowdhury | 8 Feb 2020 11:44 AM • 6 mins read

Vivek Ranjan Agnihotri

Solution to a lot of world's problems lie in India. But you can't find them as long as you ridicule our ancient wisdom.

Cannabis is a magic plant. Till mid-80s it was sold by

Figure 6.1: *End-to-end pipeline for `UTDRM`: a two-step method involving the generation of topical claims and the training of a neural retrieval model.*

6.1 Introduction

Automated fact-checking systems are pivotal not only for combatting false information on digital media but also for reducing the workload of fact-checkers (Procter et al., 2023; Shaar et al., 2020b). A key functionality of these systems is the retrieval of already debunked narratives for misinformation claims, which essentially means retrieving previously fact-checked similar claims (Nakov et al., 2022c, 2021a; Shaar et al., 2020b). This function is accomplished by training debunked-narrative retrieval models that utilise misinformation claims as queries to retrieve relevant debunked narratives.

Previous methods for training debunked-narrative retrieval models heavily rely on annotated pairs of misinformation claims and debunks (Shaar et al., 2020b; Nakov et al., 2021a; Kazemi et al., 2021). However, the process of manually creating annotated pairs is time-consuming, labour-intensive, and often limited in scale, which can impede the performance of the retrieval models.

In this paper, we propose an **U**nsupervised method for **T**raining **D**ebunked-Narrative **R**etrieval **M**odels (`UTDRM`) that utilises synthetic claims to overcome the limitation of relying on manual annotations (see Figure 6.1). Moreover, we hypothesise that `UTDRM` has the potential to detect topical misinformation by generating claims from incoming topical fact-checks, thereby expanding its overall impact. Furthermore, our proposed *entity inoculation* method (Section 6.6.3) addresses the pressing challenge of similar false narratives evolving with different entities (Singh et al., 2021a). Our inspiration for this approach stems from an independent analysis, noting similar misinformation claims involving distinct entities. For example, misinformation about crocodile sightings during floods vary across locations – Hyderabad[1], Patna[2], Bengaluru[3], and Florida[4] (see Appendix 6.10.4 for more examples). By replacing named entities in generated claims, *entity inoculation* enhances the robustness of our UTDRM method, directly addressing the issue of narrative adaptability (see Section 6.6.3).

In particular, the research question addressed in this study is: how to train efficient debunked-narrative retrieval models without relying on human-annotated data?

The main contributions of this paper are:

- UTDRM, a two-step method for training debunked-narrative retrieval models that achieves comparable or superior retrieval scores to supervised models, all without relying on annotations. Figure 6.1 illustrates the UTDRM's end-to-end pipeline.
- A **large-scale dataset of synthetic topical claims** created using topical claim generation techniques based on text-to-text transformer-based models and large language models (LLMs).
- A **comprehensive performance evaluation** of UTDRM on seven publicly available datasets, demonstrating its effectiveness and generalisability in retrieving accurate debunks for misinformation in tweets, political debates, or speeches.
- **Extensive ablation experiments** that assess the impact of different factors on UTDRM's performance. This includes: 1) the volume of fact-checking articles utilised, 2) the number of synthetically generated claims used for training, 3) the proposed *entity inoculation* method, and 4) the usage of LLMs, such as Large Language Model Meta AI (LLaMA 2) and Chat Generative Pre-trained Transformer (ChatGPT), for retrieval.

In the following sections, we discuss related work (Section 6.2) and our proposed UTDRM method (Section 6.3). Section 6.4 presents the various experimental methods and the datasets used for evaluation. The results and ablation experiments are presented in Section 6.5 and Section 6.6 respectively. Finally, we conclude the paper in Section 6.8.

[1] https://factcheck.afp.com/no-footage-has-circulated-2019-reports-about-crocodile-west-india

[2] https://www.boomlive.in/crocodile-spotted-during-bihar-floods-video-from-gujarat-shared-as-patna/

[3] https://www.indiatoday.in/fact-check/story/fact-check-crocodile-spotted-waterlogged-bengaluru-viral-video-mp-1997133-2022-09-06

[4] https://factcheck.afp.com/doc.afp.com.32KT6D7

6.2 Related work

Information retrieval involves the search and retrieval of relevant documents from a collection in response to a query. Initially, conventional lexical methods such as, Okapi Best Match 25 (BM25) (Robertson et al., 2009), Term Frequency-Inverse Document Frequency (TF-IDF) weighting (Salton and Buckley, 1988), Query Likelihood model (QL) (Ponte and Croft, 2017), and Divergence From Randomness (DFR) (Amati and Van Rijsbergen, 2002), were the primary information retrieval techniques, which demonstrated the effectiveness of lexical and statistical approaches. However, these traditional approaches faced challenges in addressing lexical gaps and semantic issues in relevance matching (Berger et al., 2000). In response to these challenges, recent Transformer-based methods (Vaswani et al., 2017) aim to harness the power of deep learning to enhance performance (Thakur et al., 2021). In the following sections, we review related work in two main areas: supervised and unsupervised methods for debunked-narrative retrieval.

6.2.1 Supervised Training Methods

Many existing methods for training debunked-narrative retrieval models rely on supervised learning techniques which typically leverage annotated pairs of misinformation claims and fact-checking articles as training data (Nakov et al., 2022c, 2021b; Hardalov et al., 2022; Shaar et al., 2020a; Sheng et al., 2021; Bhatnagar et al., 2022). For instance, Shaar et al. (Shaar et al., 2020a) train a pairwise learning-to-rank model for identifying debunked narratives. They also release Snopes and Politifact datasets (Shaar et al., 2020a), which we use for evaluation in this paper (Section 6.4.1). Similarly, Vo and Lee (2020) train a ranking model that incorporates both textual and visual features to retrieve previously fact-checked content, while Shaar et al. (2022) employ the Transformer-XH (Zhou et al., 2019) to examine the role of context in political debates. On the other hand, Kazemi et al. (Kazemi et al., 2022, 2021) address the task of debunked-narrative retrieval as a binary classification problem and train support vector machines model to classify misinformation tweets. However, formulating it as a classification problem is computationally not scalable due to its quadratic complexity.

The Conference and Labs of the Evaluation Forum (CLEF) CheckThat! Lab shared tasks 2020, 2021 and 2022 (Shaar et al., 2020b; Nakov et al., 2021b, 2022c; Barrón-Cedeño et al., 2023) focus on debunked-narrative retrieval task and release different datasets for training and testing. In this paper, we utilise all of these CLEF test datasets for evaluation (Section 6.4.1). Teams in CLEF 22 use diverse methods, such as Sentence-T5 and GPT-Neo for re-ranking (Shliselberg and Dori-Hacohen, 2022), Simple Contrastive Learning of Sentence Embeddings (SimCSE) (Gao et al., 2021), and data augmentation like back translation (Frick and Vogel, 2022). We utilise the state-of-the-art performance demonstrated by the shared task winners as a benchmark for comparing against our `UTDRM` method (Section 6.4.2).

While supervised training approaches require annotated training data, which can be costly and time-consuming to collect, this research proposes an alternative novel approach. By utilising fact-checking articles from professional fact-checking organisations, our method generates high-quality training data without the need for annotations. This methodology yields high scores in debunked-narrative retrieval (Section 6.5).

6.2.2 Unsupervised Training Methods

In recent years, unsupervised training methods for information retrieval have gained significant interest (Lee et al., 2019; Wang et al., 2021, 2022; Chang et al., 2020; Thakur et al., 2021). Our proposed `UTDRM` method falls within this category. These unsupervised methods aim to overcome the challenges associated with acquiring annotated training data by utilising large corpora of unlabeled documents. For example, Lee et al. (2019) introduce the Inverse Cloze Task (ICT) for training models using synthetic query-passage pairs by uniformly sampling sentences from random passages. Alternatively, Tranformer-based Denoising AutoEncoder (TSDAE) (Wang et al., 2021) encodes sentences with randomly deleted 60% of the tokens and the decoder to reconstruct the original sentences. Similarly, methods like SimCSE (Gao et al., 2021) and Contrastive Tension (Carlsson et al., 2021) focus on minimising the distance between embeddings from the same sentence. ICT, TSDAE, and SimCSE are among the unsupervised methods employed for comparison with our proposed `UTDRM` method (as discussed in Section 6.4.2).

Other lines of unsupervised methods explore query generation as an alternative to improve retrieval performance. For example, Nogueira et al. (Nogueira et al., 2019a,b) enhance traditional BM25 search by expanding passages with synthetic queries. On the other hand, Ma et al. (2021) propose a zero-shot learning approach for passage retrieval using synthetic question generation, while Wang et al. (2022) introduce Generative Pseudo Labelling (GPL), an unsupervised domain adaptation method that combines a T5-based query generator with pseudo labelling from a cross-encoder. However, these methods are not suitable for our specific use case since generating claims from fact-checking articles is a novel task in itself, and therefore, relying on pre-trained query generation models trained for different purposes is not appropriate. Additionally, the use of Margin Mean Squared Error (MarginMSE) (Hofstätter et al., 2020) in GPL, which relies on a cross-encoder trained on Microsoft Machine Reading Comprehension (MSMARCO) data, may not be effective for our specific debunked-narrative retrieval task. This is because our task differs from general information retrieval tasks that typically require general queries as input, while the task in this paper specifically focuses on false claims on social media and political debates (Section 6.4.1).

While existing unsupervised methods show promising results, there is still room for improvement in retrieval performance and applicability. UTDRM aims to address these challenges by utilising unsupervised learning techniques tailored specifically for training debunked-narrative retrieval models. It focuses on generating high-quality topical misinformation claims from fact-checking articles (Section 6.3.1) which, to the best of our knowledge, has not been explored in previous work. These generated claims are employed to train the retrieval model in a zero-shot setting (Section 6.3.2).

Finally, this study is the first to assess the performance of LLMs (LLaMA 2 and ChatGPT) as listwise re-rankers on seven publicly available debunked-narrative retrieval datasets (Section 6.6.4). This assessment is conducted to examine how LLMs perform in comparison to other unsupervised methods, including our UTDRM.

6.3 UTDRM: Unsupervised Method for Training Debunked-Narrative Retrieval Models

Debunked-narrative retrieval is a key task in a typical fact-checking workflow, where the verification professionals determine whether the claim or content that they need to verify has already been debunked in a publicly available debunking article posted by another fact-checking organisation. This is essentially a retrieval, where a misinformation claim serves as the query to extract relevant debunked claims (or fact-checked claims) from a database of already published publicly available debunking articles. It must be noted that if a claim has not already been debunked in a published article, there may not be suitable matches.

This section presents our proposed UTDRM method, which consists of two steps: i) generation of topical claims (Section 6.3.1); and ii) training of a debunked-narrative retrieval model (Section 6.3.2). Figure 6.1 illustrates the end-to-end pipeline for UTDRM.

6.3.1 Topical Claim Generation

We synthetically generate topical claims that resemble misinformation claims based on the debunked information provided by professional fact-checkers. To accomplish this, we propose two novel methods: the use of Text-to-Text Transfer Transformer (T5) and ChatGPT as claim generators. In this work, we specifically investigate the zero-shot scenario, where annotated pairs of social media posts and debunked claim pairs are unavailable, and only a large corpus for fact-checks is available.

6.3.1.1 T5 Claim Generator

The T5 claim generator is a sequence-to-sequence model based on the text-to-text transfer transformer (T5) (Raffel et al., 2020b). We choose T5 model because of its proven effectiveness in various sequence-to-sequence tasks in prior research (Raffel et al., 2020b; Nogueira et al., 2019a; Wang et al., 2022). T5 is used to generate claims from fact-checking articles by framing the task as an encoder-decoder problem. The encoder is trained to understand and represent the fact-checking articles, while the decoder generates potential misinformation claims that can be effectively debunked using the corresponding fact-checking articles.

To train the T5 claim generator, first, we create a corpus of fact-checking articles published by different fact-checking organisations, namely Boomlive[5], Agence France-Presse (AFP)[6] and Politifact[7]. We choose these fact-checking websites for their wide topic coverage, deferring the comparison of claim generators trained on different websites for future research. A total of 23,901 fact-checking articles were collected. For each fact-checking article, we collect the debunked claim statement, the title and the main body of the article. During fine-tuning, the input to the T5 model consists of the title and the main body of the fact-checking article, and the model is trained to generate the debunked claim statement. Since the generated claims are conditioned on the fact-checking article, they remain closely related to the actual claims being debunked in the fact-checking article. Please refer to Appendix 6.10.1 for hyperparameter details.

6.3.1.2 ChatGPT Claim Generator

We use ChatGPT (*gpt-3.5-turbo*)[8] to generate tweets that are relevant to the debunked claims of fact-checking articles collected above. To achieve this, we provide an input prompt instructing the model to generate five different tweets about the text, ensuring that the generated tweets are not fact-checks or debunks. Additionally, we encourage the diversity of hashtags in the generated tweets to enhance their variability. For this, we use the input prompt as:

```
Generate ten different tweets about the text delimited by triple backticks.
Make sure that generated tweets should not be a fact-check or a debunk.
Also, tweets should have different hashtags. ‘‘‘{Debunked Claim}‘‘‘
```

In summary, we use ChatGPT in conjunction with the T5 claim generator due to our observation that ChatGPT generates claims that are more diverse (Table 6.2) and closely resemble actual tweet claims (Section 6.3.1.3). Additionally, both T5

[5] https://www.boomlive.in/

[6] https://www.afp.com/

[7] https://www.politifact.com/

[8] https://platform.openai.com/docs/models

Table 6.1: *Sample claims generated from T5 and ChatGPT claim generator (Section 6.3.1). The debunked claim statements from the fact-checking articles used to generate claims are also mentioned.*

Debunked Claims	T5 Generated Claims	ChatGPT Generated Claims
Sen. Kamala Harris “may have broken laws” when she visited a voting site in Ohio.	1. Kamala harris “said “this weekend, yells at people in line to vote in ohio, may have broken laws.” 2. Sen. Kamala Harris broke ohio election laws. 3. Kamala harris broke ohio election laws by speaking to people in line to vote outside a polling site.	1. Did Sen. Kamala Harris cross the line during her Ohio voting site visit? Investigation underway. #votingrights #legalissues 2. Sen. Kamala Harris is facing accusations of breaking the law during a recent visit to an Ohio voting site. Stay tuned for updates! 3. Did Sen. Kamala Harris break the law during a recent visit to an Ohio voting site? The controversy continues to unfold. #VoterFraud
Dr Kafeel Khan driving a tractor at the farmers tractor rally in Delhi on Republic Day.	1. Dr Kafeel Khan was present at farmers tractor rally in Delhi. 2. Dr Kafeel Khan sitting on a red tractor in Delhi 3. Dr Kafeel Khan Was Present On Tractor Rally By Farmers On Republic Day	1. A picture worth a thousand words - Dr. Kafeel Khan driving a tractor at the farmers rally in Delhi on Republic Day. #Solidarity #India 2. Dr. Kafeel Khan showing his solidarity with farmers as he drives a tractor at the Republic Day rally in Delhi. #FarmersProtest #India 3. A powerful image of Dr. Kafeel Khan joining the farmers tractor rally in Delhi on Republic Day. #Solidarity #FarmersRights
Pondicherry University student found the cure for covid-19.	1. Pondicherry university student found a coronavirus cure by mixing ginger, black pepper powder and honey for five days. 2. "An Indian student from Pondicherry university found a home remedy cure for COVID-19 which is for the very first time accepted by WHO." 3. BJP member of parliament, Manaskar Singh, found "a home remedy cure for COVID-19, which is for the very first time accepted by WHO".	1. #Inspiration This is the kind of news that inspires us all. A student from Pondicherry University has made a groundbreaking discovery in the fight against COVID-19. Let’s support and celebrate our young innovators! 2. The world is one step closer to defeating COVID-19 thanks to a brilliant mind from Pondicherry University. Let’s celebrate this incredible achievement! #science #innovation 3. #Hope A glimmer of hope in these trying times! A student from Pondicherry University has uncovered a potential cure for COVID-19. Let’s keep our fingers crossed for a breakthrough!
Photos show cash and gold recovered in income tax raids at MK Stalins daughters house.	1. Photos show money seised from the Income Tax raids conducted at the house of MK Stalin’s daughter Senthamarai 2. Photos shows the money seized at the house of MK Stalins daughter Senthamarai Sabareesan 3. Photos show cash and gold taken in Income Tax raids at the house of MK Stalins daughter Senthamarai	1. The rich get richer? Income tax raids at MK Stalin’s daughter’s house expose a hoard of cash and gold. Will they be held accountable? #corruption #India 2. Shocking revelation! Income tax raids at MK Stalin’s daughter’s house reveal cash and gold - photos don’t lie! #corruption #IndiaPolitics 3. The truth is out! Income tax raids at MK Stalin’s daughter’s house reveal a stash of cash and gold. Will justice prevail? #corruption #India
Countries were exporting COVID-19 “diagnostic test instruments” in 2018.	1. The world bank created a list of COVID-19 tests before the coronavirus pandemic. 2. “Coronavirus tests exist in 2018, predates the pandemic.” 3. A list of exports of COVID-19 diagnostic tests "predates" the coronavirus.	1. Did you know that countries were already exporting #COVID19 diagnostic test instruments as early as 2018? Interesting discovery. 2. This is mind-boggling: countries were exporting COVID-19 "diagnostic test instruments" in 2018! How did we not see this coming? 3. It’s surprising to learn that #COVID19 diagnostic test instruments were already being exported by countries in 2018. What does this tell us?

and ChatGPT claim generator can address emerging topics by generating claims from incoming topical fact-checks. These generated claims serve as valuable inputs for training our neural retrieval model (Section 6.3.2).

6.3.1.3 Generated Claims

Table 6.1 showcases sample claims generated from T5 and ChatGPT. We present five random instances of debunked claims alongside three generated claims from each model. In the first example, T5 produces three claims pertaining to Senator Kamala Harris potentially violating laws during a visit to an Ohio voting site, while ChatGPT generates alternative claims with similar themes. Similarly, for the other examples, T5 and ChatGPT generate diverse variations of claims related to Dr Kafeel Khan’s involvement in a farmers’ rally in Delhi and a supposed COVID-19 cure by a Pondicherry University student.

In summary, both T5 and ChatGPT generate different types of claims with variations in wording, focus, and emphasis, while still conveying similar information related to the original debunked claims. Moreover, our analysis reveals that the claims generated by T5 exhibit simplicity and a higher level of similarity to the debunked claims. On the other hand, the claims generated by ChatGPT demonstrate greater diversity and closely resemble actual tweets, often incorporating hashtags (as shown in Table 6.2 – Section 6.3.1.4). Notably, some of the ChatGPT generated claims ask questions while stating the debunked claim (last example in Table 6.1). Finally, by using both T5 and ChatGPT, we can capture a broader range of claim styles and ensure comprehensive coverage for training debunked-narrative retrieval models.

6.3.1.4 Quality and Diversity

Table 6.2 evaluates the generated claims using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) and self Bilingual Evaluation Understudy (selfBLEU) (Shu et al., 2019) metrics. Following previous work (Nogueira et al., 2019a; Wang et al., 2022), our evaluation does not involve human assessment. Instead, we rely on automatic metrics to assess the quality of generated claims. ROUGE measures the proximity of the generated claims to the reference debunked claims, while selfBLEU assesses the diversity among the generated claims. The choice of these metrics is justified by their close alignment with our research objectives, emphasising both quality and diversity as crucial evaluation criteria. We generate a total of six claims (three from each claim generator) from the collected fact-checking articles (Section 6.3.1), as it yields the best scores during experiments (see Section 6.6.2). The results in Table 6.2 indicate that T5 outperforms ChatGPT in ROUGE scores across all n-gram levels, indicating higher overlap with the reference debunked claims. This performance difference can be attributed to the fine-tuning of T5 in the T5 claim generator. Further evaluation of retrieval models trained on generated claims will provide insights into the claim quality and their alignment with task requirements (Section 6.5).

Table 6.2 also presents the selfBLEU scores, which computes the similarity between the generated claims, with lower scores indicating higher diversity. T5 exhibits higher selfBLEU scores across all N-gram levels, indicating more similarity among its generated claims. In contrast, ChatGPT achieves lower selfBLEU scores, suggesting greater diversity and distinctiveness in its generated claims.

6.3.2 Neural Retrieval Model

The neural retrieval model is a transformer model fine-tuned on the generated claim and the original debunked claim statement pairs using multiple negatives ranking loss (MNRL) (Oord et al., 2018; Henderson et al., 2017). In this, consider a dataset of synthetically generated claims $g = (g_1, ..., g_N)$ along with their corresponding debunked claim statements $d = (d_1, ..., d_N)$. During fine-tuning, each batch of size K contains one generated claim g_i and one corresponding relevant

Table 6.2: *ROUGE and selfBLEU scores for claims generated from T5 and ChatGPT claim generator. Lower selfBLEU scores indicate higher diversity, while higher ROUGE scores indicate greater overlap with the reference debunked claims.*

Metrics	ROUGE-1	ROUGE-2	ROUGE-L	selfBLEU1	selfBLEU2	selfBLEU3
T5	0.563	0.423	0.541	0.553	0.493	0.444
ChatGPT	0.272	0.119	0.237	0.250	0.142	0.085

debunked claim statement d_i, which is the same debunked claim used for generating g_i. The remaining $K-1$ elements in the batch are irrelevant debunked claim statements which are the hard negatives mined using a pretrained retrieval model. Every debunked claim statement d_j is a negative candidate for generated claim g_i if $i \neq j$. The loss for a single batch of size K is defined as,

$$-\frac{1}{K}\sum_{i=1}^{K}\log\frac{exp(Sim(f_\theta(g_i),f_\theta(d_i)))}{\sum_{j=1}^{K}exp(Sim(f_\theta(g_i),f_\theta(d_j)))} \tag{6.1}$$

where f_θ is the sentence encoder using the transformer model and Sim is the similarity between the encoded embeddings. We employ cosine similarity function with the mean-pooling technique due to its proven effectiveness in prior research (Reimers and Gurevych, 2019). MNRL aims to maximise the similarity between the generated claim and its relevant debunked claim statement while minimising the similarity with irrelevant statements. Hyperparameter details are in Appendix 6.10.1.

6.4 Experimental Setup

6.4.1 Evaluation Datasets

We evaluate the models on the test set of seven publicly available datasets. The datasets are divided into two types based on whether the claims are sourced from Twitter or from political debates or speeches:

- ***Twitter-based*** **datasets**: **Snopes** (Shaar et al., 2020a) and CLEF CheckThat! Lab task datasets which include **CLEF 22 2A** (Nakov et al., 2022c), **CLEF 21 2A** (Nakov et al., 2021b) and **CLEF 20 2A** (Shaar et al., 2020b).
- ***Political-based*** **datasets**: **Politifact** (Shaar et al., 2020a) and CLEF CheckThat! Lab task datasets which include **CLEF 22 2B** (Nakov et al., 2022c) and **CLEF 21 2B** (Nakov et al., 2021b).

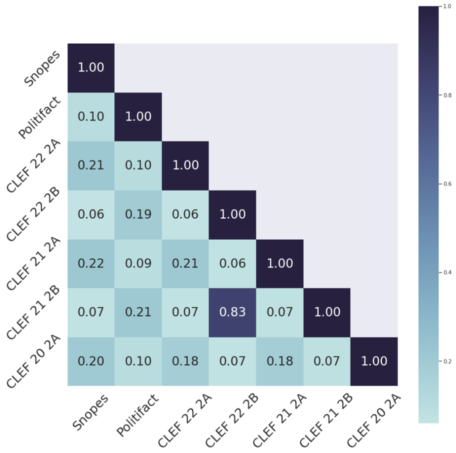


Figure 6.2: *Heatmap for dataset domain overlap.*

To assess the diversity of domains, we calculate the pairwise domain overlap between all the claims in the datasets using a weighted Jaccard similarity measure (Ioffe, 2010). Figure 6.2 shows a heatmap illustrating the pairwise weighted Jaccard similarity scores. Besides CLEF 22 2B and CLEF 21 2B, the results indicate a relatively low overlap among most datasets, suggesting that the evaluation of UTDRM is conducted on diverse data.

In order to avoid any data leakage with the fact-checking articles utilised for claim generation (Section 6.3.1), we exclude all fact-checking articles that exhibit a Jaccard similarity of 0.5 or higher between the debunked claim statements. Please note that fact-checking articles used for claim generation are removed and are not from the evaluation datasets.

6.4.2 Baselines

Okapi BM25. We use the ElasticSearch[9] (Gormley and Tong, 2015) implementation of BM25 (Jones et al., 2000), with default parameters in ElasticSearch ($k = 1.2$ and $b = 0.75$).

Out-of-the-box models. We use two strong out-of-the-box pre-trained models for information retrieval. We test these models in their default configuration without any supervision from the generated claims to assess their zero-shot performance. The models are: 1) *Sentence-Transformer's* model based on Masked and Permuted Pre-training for Language Understanding (**MPNet**) (Song et al., 2020)

[9] https://www.elastic.co/elasticsearch/

all-mpnet-base-v2[10] which has been trained on a large and diverse dataset of over a billion training examples. 2) Approximate Nearest Neighbor Negative Contrastive Estimation (**ANCE**), which is a RoBERTa (Liu et al., 2019) model fine-tuned on MSMARCO dataset (Nguyen et al., 2016) with hard negatives selected using approximate nearest neighbor (Xiong et al., 2021).

Unsupervised methods. We use five different unsupervised methods which utilise the same set of fact-checking articles for training, as used in the claim generation process (Section 6.3.1): 1) **ICT** (Lee et al., 2019) is employed to generate pseudo-claims by uniformly sampling sentences from the fact-checking articles. MNRL loss (Section 6.3.2) is then applied to train the model using the pairs of pseudo and debunked claim statements. 2) Back-Translation (**BT**) (Sennrich et al., 2016) involves translating all debunked claim statements to Hindi and then back to English. The resulting pairs of back-translated claim and the original debunked claim statement are further used for training the model using MNRL loss. 3) **SimCSE** (Gao et al., 2021) encodes the same debunked claim statement twice with different dropout masks and utilises MNRL loss for training. 4) **TSDAE** (Wang et al., 2021) pre-trains a retrieval model using a denoising autoencoder. It encodes debunked claim statements with randomly deleted 60% of the tokens and the decoder reconstructs the original debunked claim statements (Wang et al., 2021). All unsupervised methods employ a distilled version of the RoBERTa-base (Liu et al., 2019)[11] as the underlying model. Hyperparameter details are in Appendix 6.10.1.

Supervised methods. We also report previous State-Of-The-Art (SOTA) performance achieved by the winners of the shared tasks on the test set, as published in their respective papers (Shaar et al., 2020a,b; Nakov et al., 2021b, 2022c). Please note that these supervised methods benefit from annotated training data, which enables them to utilise specific information pertaining to real-world instances of misinformation claims and their corresponding debunks.

For example, the winning team of CLEF 22 2A (Shliselberg and Dori-Hacohen, 2022) use Sentence-T5 (Ni et al., 2022) for candidate selection and GPT-Neo (Gao et al., 2020) for re-ranking. The winning team in CLEF 22 2B (Hövelmeyer et al., 2022) employ a combination of semantic and lexical similarity features between claims and debunks for retrieval. In CLEF 21 2A (Nakov et al., 2021b), the top-performing team utilise a combination of TF-IDF, Sentence-BERT, and Lambda Multiple Additive Regression Trees (LambdaMART) for ranking (Chernyavskiy et al., 2021), while the winning team in CLEF 21 2B (Mihaylova et al., 2021) combines the Sentence-BERT model with a custom neural network to get the final list

[10] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[11] https://huggingface.co/distilroberta-base

of sorted debunks based on relevance. The top-performing team in CLEF 20 2A (Bouziane et al., 2020) use a fine-tuned RoBERTa model for retrieval.

Lastly, for Snopes and Politifact, we directly report scores from Shaar et al. (2020a), who utilise a pairwise learning-to-rank model for debunk retrieval.

6.4.3 Experimental Details

UTDRM is tested on two models: a distilled version of the RoBERTa-base model (UTDRM-RoBERTa) and the MPNet model (UTDRM-MPNet) (Section 6.5). We generate six topical claims (three from each claim generator) for all the collected 23,901 fact-checking articles (Section 6.3.1), as this approach yields the best scores during experiments (Section 6.6.2). Following previous work (Wang et al., 2022), we employ nucleus sampling during generation, using a *Top-k* value of 25 and a *Top-p* value of 0.95. For the ChatGPT claim generator, we keep all API parameters at their default values, except for the temperature, which is set to 0.7 to ensure diversity. The total cost of using ChatGPT to generate the claims was 14 GBP. Finally, a total of 1,43,406 (23,901x6) generated claims are used for training the neural retrieval model.

6.4.4 Evaluation Metrics

For evaluation, we employ two widely used ranking metrics (Nakov et al., 2021b, 2022c): Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP). MRR computes the score based on the highest-ranked relevant debunk for each misinformation tweet and is defined as $\mathrm{MRR} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{1}{\mathrm{rank}_i}$, where $|C|$ is the number of input claims used as query and $rank_i$ is the rank of the relevant debunk for the ith claim. The higher the MRR score the better. MAP, on the other hand, measures the precision of the system in returning relevant results for a given query. We use two variations of MAP: MAP@1 and MAP@5, which evaluate the top one and top five retrieved documents, respectively. A higher MAP@k score indicates better performance.

6.5 Results and Discussion

Table 6.3 reports the results of UTDRM evaluation divided into two parts: the top part presents the individual and average results for *Twitter-based* datasets (Snopes, CLEF 22 2A-EN, CLEF 21 2A-EN & CLEF 20 2A-EN), while the bottom part showcases the individual and average results for *political-based* datasets (Politifact, CLEF 22 2B-EN & CLEF 21 2B-EN).

Table 6.3: *Performance of BM25, out-of-the-box, unsupervised and SOTA supervised models. The first part of the table shows the individual and average results for Twitter-based datasets, while the second part shows the individual and average results for political-based datasets.* UTDRM *results are highlighted in blue. The highest scores for each dataset and metric are in* ***bold****.*

Datasets	Metrics	Elastic	Out-of-the-box		Unsupervised Methods						Supervised
		BM25	**MPNet**	**ANCE**	**BT**	**ICT**	**SimCSE**	**TSDAE**	**UTDRM-RoBERTa**	**UTDRM-MPNet**	**Prev SOTA**
Snopes	**MAP@1**	0.557	0.776	0.662	0.333	0.627	0.545	0.458	0.716	**0.831**	0.691
	MAP@5	0.690	0.840	0.752	0.406	0.737	0.643	0.532	0.811	**0.889**	0.782
	MRR	0.786	0.843	0.759	0.418	0.745	0.652	0.548	0.815	**0.890**	0.788
CLEF 22 2A	**MAP@1**	0.823	0.866	0.761	0.368	0.756	0.589	0.469	0.823	0.933	**0.943**
	MAP@5	0.856	0.898	0.800	0.425	0.797	0.661	0.520	0.857	0.946	**0.956**
	MRR	0.862	0.899	0.807	0.444	0.804	0.674	0.539	0.861	0.948	**0.957**
CLEF 21 2A	**MAP@1**	0.797	0.837	0.767	0.332	0.762	0.644	0.510	0.817	**0.906**	0.861
	MAP@5	0.844	0.881	0.815	0.386	0.819	0.694	0.564	0.863	**0.933**	0.883
	MRR	0.849	0.885	0.823	0.406	0.825	0.704	0.579	0.869	**0.936**	0.884
CLEF 20 2A	**MAP@1**	0.834	0.884	0.869	0.372	0.769	0.673	0.578	0.874	**0.945**	0.897
	MAP@5	0.869	0.924	0.893	0.416	0.836	0.711	0.642	0.913	**0.961**	0.929
	MRR	0.878	0.925	0.896	0.436	0.840	0.722	0.653	0.915	**0.961**	0.927
Average	**MAP@1**	0.753	0.841	0.765	0.351	0.729	0.613	0.504	0.808	**0.904**	0.848
Twitter-based	**MAP@5**	0.815	0.886	0.815	0.408	0.797	0.677	0.565	0.861	**0.932**	0.888
	MRR	0.844	0.888	0.821	0.426	0.804	0.688	0.580	0.865	**0.934**	0.889
Politifact	**MAP@1**	0.467	0.413	0.428	0.387	0.445	0.355	0.424	0.426	0.516	**0.531**
	MAP@5	0.503	0.494	0.499	0.446	0.512	0.404	0.477	0.507	**0.600**	0.588
	MRR	0.541	0.524	0.532	0.464	0.543	0.431	0.504	0.539	**0.627**	0.608
CLEF 22 2B	**MAP@1**	0.308	0.285	0.331	0.277	0.254	0.254	0.269	0.369	0.392	**0.408**
	MAP@5	0.371	0.344	0.368	0.295	0.336	0.276	0.309	0.408	0.431	**0.459**
	MRR	0.419	0.374	0.413	0.337	0.377	0.319	0.349	0.459	0.467	**0.475**
CLEF 21 2B	**MAP@1**	0.285	0.247	0.272	0.222	0.215	0.209	0.241	0.310	**0.348**	0.304
	MAP@5	0.343	0.308	0.310	0.239	0.282	0.226	0.276	0.340	**0.392**	0.346
	MRR	0.377	0.333	0.344	0.268	0.317	0.262	0.307	0.386	**0.422**	0.350
Average	**MAP@1**	0.353	0.315	0.344	0.295	0.305	0.273	0.311	0.368	**0.419**	0.414
Political-based	**MAP@5**	0.406	0.382	0.392	0.327	0.377	0.302	0.354	0.418	**0.474**	0.464
	MRR	0.446	0.410	0.430	0.357	0.412	0.337	0.387	0.461	**0.505**	0.478

BM25 and out-of-the-box models. These models consistently achieve high retrieval scores across all metrics, with MPNet outperforming the others (Table 6.3 column 3–5). This indicates that leveraging models trained on other information retrieval datasets can improve retrieval effectiveness (Section 6.4.2). However, it is important to note that there are variations in performance among the datasets, suggesting that the models' effectiveness might depend on the specific characteristics of the dataset.

Among the *Twitter-based* datasets, MPNet stands out as the best-performing model with the highest average scores. It achieves an average MAP@1 score of 0.841, MAP@5 score of 0.886, and MRR score of 0.888. In contrast, when considering the *political-based* datasets (Politifact, CLEF 22 2B-EN, and CLEF 21 2B-EN), BM25 emerges as the top-performing model with the average MAP@1 score of 0.353, MAP@5 score of 0.406, and MRR score of 0.446 , indicating its effectiveness in retrieving relevant information from political speech datasets. Overall, the average scores suggest that the models perform better on the *Twitter-based* datasets compared to the *political-based* datasets. This difference in performance can be attributed to the fact that *political-based* claims pose greater challenges for the models.

Unsupervised methods. Table 6.3 reports the results of the unsupervised methods, including the baselines BT, ICT, SimCSE, TSDAE (columns 6–9), as well as the proposed `UTDRM-RoBERTa` (Table 6.3 columns 10). All these methods utilise a distilled RoBERTa model, as described in Section 6.4.2. Among the baselines, ICT achieves the highest scores across all metrics, followed by SimCSE and TSDAE. However our proposed `UTDRM-RoBERTa` achieves the highest average scores for both *Twitter-based* and *political-based* datasets, followed by ICT and SimCSE. Additionally, the table reveals that each method has its own strengths and weaknesses on different datasets. For instance, `UTDRM-RoBERTa` performs well on all datasets except Politifact, where it is surpassed by ICT.

Furthermore, given the impressive performance of the out-of-the-box MPNet model, we also test `UTDRM` on the MPNet model (Table 6.3 column 11). `UTDRM-MPNet` outperforms all other methods, achieving the highest scores across all evaluation metrics. It obtains an average MAP@1, MAP@5, and MRR of 0.904, 0.932, and 0.934, respectively, for *Twitter-based* datasets. For *political-based* datasets, it achieves an average MAP@1, MAP@5, and MRR of 0.419, 0.474, and 0.505, respectively. Overall, `UTDRM-MPNet` consistently achieves the highest scores across all datasets, demonstrating its effectiveness. `UTDRM-RoBERTa` also performs well, albeit slightly lower than `UTDRM-MPNet`.

Supervised methods. Table 6.3 (last column) reports the results for the previous SOTA methods (Section 6.4.2). These methods benefit from annotated training data, allowing them to leverage specific information about real-life misinformation claims and debunked claim statements (Section 6.4.2). In contrast, the `UTDRM` does not have access to any annotated training data. Surprisingly, the `UTDRM-MPNet` model, despite being an unsupervised method, achieves comparable or even superior retrieval scores compared to the SOTA supervised models. This demonstrates the effectiveness of `UTDRM` without the need for any annotations.

Summary. We find that the choice of method depends on specific requirements, data availability, and the desired performance-resource trade-off. `UTDRM-RoBERTa` and `UTDRM-MPNet` consistently yield the highest retrieval scores, while the out-of-the-box models offer viable alternatives without the need for any training data whatsoever for debunked-narrative retrieval. Additionally, our proposed method, `UTDRM`, has the potential to detect topical misinformation claims by generating claims from incoming topical fact-checks; thus allowing it to address emerging topics and contribute to the timely detection of misinformation.

6.6 Analysis

6.6.1 Influence of Fact-checking Articles

Table 6.4 shows the results UTDRM-MPNet when trained using different numbers of fact-checking articles (1K, 5K, 10K, and *All*). Due to space limitations, the table reports only the MAP@1 and MRR metrics.

The results suggest that the size of the corpus does have a positive effect on the performance of UTDRM, but the extent of the improvement may vary depending on the specific dataset and corpus size being used (Table 6.4). For instance, the CLEF 21 2A (*Twitter-based* dataset) shows an increasing trend until the number of fact-checking articles reaches 10K, after which it becomes relatively constant. On the other hand, for *political-based* datasets, the average performance continues to increase as the number of fact-checking articles increases, suggesting that a larger corpus of fact-checking articles has a more pronounced impact on improving retrieval performance.

6.6.2 Influence of the Generated Claims

Table 6.5 shows results of UTDRM-MPNet using different numbers of generated claims for training N: 2, 6, 10, 20. It should be noted that the proportion of claims generated using T5 and ChatGPT is kept the same for all cases. The individual performance of models trained on T5 and ChatGPT generated claims separately is generally lower (Appendix 6.10.3 and 6.10.2).

Table 6.5 demonstrates an overall improvement in performance as the number of generated claims increases from $N = 2$ to $N = 6$ and $N = 10$ across most datasets. However, performance either declines or stabilises beyond $N = 10$. For instance, in the Snopes dataset, MAP@1 and MRR scores show a slight decline from $N = 6$ to $N = 20$. Similar trends are observed in the CLEF 22 2A, CLEF 21 2A, and CLEF 20 2A datasets, where MAP@1 performance peaks at $N = 6$ and then plateaus or slightly decreases. In contrast, the CLEF 22 2B and CLEF 21 2B datasets reach their peak performance at $N = 10$. In general, the results suggest that $N = 6$ is the optimal value for the number of generated claims, as it yields the highest average retrieval performance, while going beyond this range may introduce noise and decrease performance.

Table 6.4: *Influence of fact-checking articles on `UTDRM`. The highest scores for each dataset and metric are in **bold**.*

Datasets	**Metrics**	**Fact-checking Articles**			
		1K	**5K**	**10K**	**All**
Snopes	**MAP@1**	0.750	0.794	0.803	**0.831**
	MRR	0.810	0.842	0.865	**0.890**
CLEF 22 2A	**MAP@1**	0.871	0.914	0.919	**0.933**
	MRR	0.904	0.932	0.936	**0.948**
CLEF 21 2A	**MAP@1**	0.851	0.891	**0.906**	**0.906**
	MRR	0.895	0.925	**0.937**	0.936
CLEF 20 2A	**MAP@1**	0.894	0.940	0.935	**0.945**
	MRR	0.931	0.957	0.957	**0.961**
Average	**MAP@1**	0.842	0.885	0.891	**0.904**
Twitter-based	**MRR**	0.885	0.914	0.924	**0.934**
Politifact	**MAP@1**	0.428	0.484	0.508	**0.516**
	MRR	0.541	0.602	0.618	**0.627**
CLEF 22 2B	**MAP@1**	0.285	0.362	0.377	**0.392**
	MRR	0.381	0.436	0.450	**0.467**
CLEF 21 2B	**MAP@1**	0.259	0.323	0.335	**0.348**
	MRR	0.346	0.390	0.402	**0.422**
Average	**MAP@1**	0.324	0.390	0.407	**0.419**
Political-based	**MRR**	0.423	0.476	0.490	**0.505**

Table 6.5: *Influence of the generated claims on* `UTDRM`*. The highest scores for each dataset and metric are in* ***bold****.*

Datasets	**Metrics**	**Generated Claims**			
		N=2	**N=6**	**N=10**	**N=20**
Snopes	**MAP@1**	0.821	**0.831**	0.830	0.829
	MRR	0.881	**0.890**	**0.890**	0.889
CLEF 22 2A	**MAP@1**	0.914	**0.933**	**0.933**	**0.933**
	MRR	0.934	0.948	0.948	**0.949**
CLEF 21 2A	**MAP@1**	0.906	**0.906**	0.901	0.896
	MRR	0.936	**0.936**	0.932	0.931
CLEF 20 2A	**MAP@1**	0.935	**0.945**	**0.945**	**0.945**
	MRR	0.957	0.961	0.963	**0.964**
Average	**MAP@1**	0.894	**0.904**	0.902	0.901
Twitter-based	**MRR**	0.927	**0.934**	0.933	0.933
Politifact	**MAP@1**	0.508	**0.516**	0.500	0.496
	MRR	0.616	**0.627**	0.619	0.615
CLEF 22 2B	**MAP@1**	0.362	0.392	**0.400**	0.392
	MRR	0.441	0.467	**0.473**	0.468
CLEF 21 2B	**MAP@1**	0.323	0.348	**0.354**	0.335
	MRR	0.394	0.422	**0.424**	0.416
Average	**MAP@1**	0.397	**0.419**	0.418	0.408
Political-based	**MRR**	0.484	**0.505**	**0.505**	0.499

Table 6.6: *Influence of entity inoculation on `UTDRM`. `UTDRM` is the deafult `UTDRM-MPNet` performance from Table 6.3. The highest scores for each dataset and metric are in* ***bold****.*

Datasets	**Metrics**	*Entity Inoculation*				**UTDRM**
		GPE	**PERSON**	**ORG**	**Combine**	**Default**
Snopes	**MAP@1**	0.831	0.831	**0.841**	0.821	0.831
	MRR	0.889	0.891	**0.893**	0.881	0.890
CLEF 22 2A	**MAP@1**	0.928	0.919	0.923	0.919	**0.933**
	MRR	0.942	0.936	0.942	0.935	**0.948**
CLEF 21 2A	**MAP@1**	**0.916**	0.901	0.901	0.906	0.906
	MRR	**0.940**	0.929	0.932	0.932	0.936
CLEF 20 2A	**MAP@1**	0.940	0.930	0.940	0.935	**0.945**
	MRR	0.957	0.955	0.958	0.955	**0.961**
Average	**MAP@1**	**0.904**	0.895	0.901	0.895	**0.904**
Twitter-based	**MRR**	0.932	0.928	0.931	0.926	**0.934**
Politifact	**MAP@1**	0.492	**0.527**	0.512	0.512	0.516
	MRR	0.613	**0.637**	0.631	0.633	0.627
CLEF 22 2B	**MAP@1**	0.415	0.400	0.415	**0.423**	0.392
	MRR	0.482	0.471	0.482	**0.495**	0.467
CLEF 21 2B	**MAP@1**	0.367	0.354	0.367	**0.373**	0.348
	MRR	0.433	0.423	0.433	**0.442**	0.422
Average	**MAP@1**	0.425	0.427	0.431	**0.436**	0.419
Political-based	**MRR**	0.509	0.510	0.515	**0.524**	0.505

6.6.3 Influence of *Entity Inoculation*

We propose an *entity inoculation* method, which involves replacing a random named entity in the generated claims with another random named entity to simulate real-world scenarios where similar misinformation narratives spread with different entities (see Appendix 6.10.4 for examples). By training the model with these modified claims, it is expected to become more robust in retrieving debunked narratives regardless of the specific entities involved. Table 6.6 presents the results of *entity inoculation* using different entity types: geopolitical entities (GPE), person (PERSON), and organisation name (ORG), as well as a combined approach that uses all types. The *Default* column represents the performance of `UTDRM-MPNet` without *entity inoculation* (from Table 6.3).

Entity inoculation shows positive results on *political-based* datasets with an average increase of two MRR points with the combined approach as compared to the *Default* performance without *entity inoculation*. This indicates the effectiveness of *entity inoculation* in handling misinformation narratives in political contexts. On the other hand, for *Twitter-based* datasets, the impact of *entity inoculation* is less pronounced. While *entity inoculation* shows benefits in making models' entities agnostic, we hypothesise that its effectiveness may be limited to datasets that contain cases where similar narratives are spread with different entities. Examples of such false narratives can be found in Appendix 6.10.4.

6.6.4 Influence of Large Language Models (LLMs)

Large Language Models (LLMs) have consistently demonstrated impressive performance across a wide range of natural language processing (NLP) tasks (Touvron et al., 2023; Köpf et al., 2023). However, their application in information retrieval tasks remains an ongoing area of research, with the aim of optimising their ability to retrieve relevant information from large corpora in response to a given input query (Ai et al., 2023; Ma et al., 2023). Therefore, to assess the performance of LLMs in comparison to our `UTDRM` method, we employ a Listwise Re-ranker with a Large Language Model (LRL) (Ma et al., 2023) to re-rank the *Top-k* documents retrieved by the initial stage ranker. In this context, the LLM is provided with the following instruction template:

```
Passage1 = {Debunk_1}
...
PassageM = {Debunk_M}
Query = {Claim}
Passages = [Passage1, ..., PassageM]
Sort the Passages by their relevance to the Query.
Sorted Passages = [
```

Table 6.7: *Influence of large language models (LLaMA 2 and ChatGPT) as a second stage retriever to re-rank the top candidate claims retrieved by BM25 and `UTDRM`. `UTDRM` is the default `UTDRM-MPNet` performance from Table 6.3. UTDRM+ChatGPT signifies that `UTDRM-MPNet` performs the initial ranking, and ChatGPT conducts the second-stage ranking. The highest scores for each dataset and metric are in* ***bold****.*

Datasets	Metrics	BM25+LLaMA2	UTDRM+LLaMA2	BM25+ChatGPT	UTDRM+ChatGPT	UTDRM
Snopes	**MAP@1**	0.460	0.657	0.667	**0.841**	0.831
	MRR	0.659	0.728	0.862	**0.890**	**0.890**
CLEF 22 2A-EN	**MAP@1**	0.794	0.890	0.895	0.919	**0.933**
	MRR	0.835	0.913	0.916	0.936	**0.948**
CLEF 21 2A-EN	**MAP@1**	0.782	0.911	0.906	**0.926**	0.906
	MRR	0.836	0.939	0.927	**0.949**	0.936
CLEF 20 2A-EN	**MAP@1**	0.673	0.729	0.849	0.925	**0.945**
	MRR	0.724	0.762	0.894	0.948	**0.961**
Average	**MAP@1**	0.679	0.819	0.822	0.895	**0.904**
Twitter-based	**MRR**	0.777	0.860	0.902	0.925	**0.934**
Politifact	**MAP@1**	0.260	0.293	0.512	**0.561**	0.516
	MRR	0.333	0.417	0.607	**0.680**	0.627
CLEF 22 2B-EN	**MAP@1**	0.285	0.346	**0.400**	**0.400**	0.392
	MRR	0.383	0.426	0.486	**0.493**	0.467
CLEF 21 2B-EN	**MAP@1**	0.266	0.310	**0.361**	**0.361**	0.348
	MRR	0.347	0.388	**0.445**	0.425	0.422
Average	**MAP@1**	0.270	0.316	0.424	**0.441**	0.419
Political-based	**MRR**	0.354	0.411	0.513	**0.533**	0.505

Please note that due to LLM memory constraints, input sequences may exceed the maximum input sequence length. In such cases, we implement progressive re-ranking (*M*=*20*) following the approach of Ma et al. (2023). This technique re-ranks *M* debunks at a time and incrementally shifts the window by *M*/*2* towards the beginning of the retrieved debunks, leading to an enhancement in the top-ranked results. In this work, we test two types of LLMs: 1) the open-sourced LLaMA 2 13B (Touvron et al., 2023; Köpf et al., 2023)[12]; and 2) the private LLM ChatGPT (*gpt-3.5-turbo*). LLaMA 2 was hosted on our local server (2x24GB NVIDIA GeForce RTX 3090) and for ChatGPT, we use OpenAI API[13]. The total cost of testing using ChatGPT was 20 GBP.

Table 6.7 shows the results of LRL using BM25 and `UTDRM-MPNet` as first-stage rankers. For example, "BM25+ChatGPT" (column 5 - Table 6.7) signifies that BM25 performs the first-stage ranking, and ChatGPT conducts the second-stage ranking. Following the methodology from prior work (Ma et al., 2023), the LLM is used to re-rank 100 documents on top of BM25 and 20 documents on top of `UTDRM`. The results indicate that ChatGPT outperforms LLaMA 2 across all datasets and

[12] We use OpenAssistant's LLaMA 2 13B model for our experiments, accessible at `https://huggingface.co/OpenAssistant/llama2-13b-orca-8k-3319`

[13] `https://platform.openai.com/docs/models`

metrics. Moreover, we find that re-ranking on top of UTDRM yields superior scores compared to re-ranking on top of BM25 (Table 6.7). Figure 6.3 visually depicts the average MRR performance of different retrieval methods.

For the *Twitter-based* datasets, although UTDRM achieves the highest average scores, UTDRM+ChatGPT outperforms UTDRM in Snopes (MAP@1) and in CLEF 21 2A-EN (MAP@1 and MRR). For the *political-based* datasets, notably, UTDRM+ChatGPT beats UTDRM and attains the highest performance in MAP@1, MAP@5, and MRR across all datasets.

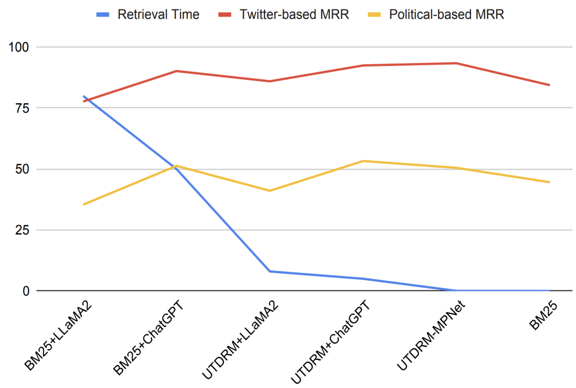


Figure 6.3: *Retrieval times (in seconds per query) and average MRR performance (scaled by a factor of 100) comparison of different retrieval methods.*

While LLMs exhibit impressive performance, it is important to consider the tradeoffs, one of which is retrieval cost and latency. We conduct experiments to measure the time taken per claim to retrieve debunks for each method and we observe notable differences in retrieval speed. Figure 6.3 shows retrieval times and average MRR performance comparison of different retrieval methods. We find that BM25+LLaMA2 and BM25+ChatGPT exhibit longer retrieval times, averaging around 80 seconds and 50 seconds per claim, respectively. In contrast, UTDRM+LLaMA2 and UTDRM+ChatGPT significantly reduce retrieval time, taking only 8 seconds and 5 seconds per claim, respectively, possibly due to the fewer number of debunks to be re-ranked. Remarkably, UTDRM-MPNet on its own achieves an exceptionally low retrieval time of just 0.04 seconds per claim. These findings underscore that, despite LLMs' impressive performance in relevance ranking, they often come at the cost of extended retrieval times, whereas our proposed UTDRM-MPNet approach offers both high relevance and exceptional retrieval speed.

6.7 Error Analysis

The evaluation of UTDRM would be incomplete without a thorough examination of the types of errors it may produce. To address this, we manually review cases where the retrieval model fails to rank the most relevant debunked claim at the top. We conduct this analysis by inspecting the retrieved debunked claims for 50 randomly selected cases from the Snopes and Politifact datasets. We find that the primary cause of such errors is when a misinformation claim is associated with multiple debunked claims (19 out of 50). For instance, the false claim "African Union warning African citizens against the safety of travelling to the United States" in Snopes has multiple relevant debunked claims. In such instances, the model assigns highly similar high scores to all relevant debunked claims, even though each misinformation claim is linked to a single debunked claim in the dataset. This highlights inconsistencies in the existing datasets and the need for further improvement.

The second type of error occurs when the retrieved debunked claim is not entirely relevant, but there is some degree of relevance to the input misinformation claim (16 out of 50). For instance, for the claim "Governor Christie has endorsed many of the ideas that Barack Obama supports, whether it is gun control or the appointment of Sonia Sotomayor", the top retrieved debunked claim discusses Governor Chris Christie and Barack Obama sharing similar views on gay marriage. This highlights the challenge of distinguishing closely related debunked claims, emphasising the need for continued refinement in retrieval models for enhanced precision. Moreover, we hypothesise that this may also be attributed to limitations in the claim generation model, where it generates claims that, while not entirely irrelevant, are only tangentially related to the intended debunked claim. Such errors suggest the propagation of errors in the retrieval process and suggest the need for improvement in the claim generation model.

The third category, accounting for 15 out of 50 cases, involves errors that occur when a misinformation claim lacks sufficient context to find the relevant debunked claim. For example, one of the misinformation claims in the Politifact dataset states "very few children" which is ambiguous and makes finding a relevant debunk challenging. Moreover, the task becomes even more challenging when misinformation claims span multiple modalities, such as combining text and images. For instance, one of the misinformation claims is a X (formerly Twitter) post stating "Botswana condemns remarks made by President Trump", along with an image containing details of the remarks. In such cases, retrieval models also require information contained in the image, as the text of the tweet alone is not sufficient. This motivates future work on multimodal debunked-narrative retrieval, where models can exploit joint information from different modalities.

6.8 Conclusion

This paper presents UTDRM, an unsupervised method for training debunked-narrative retrieval models that effectively overcomes the reliance on manually annotated training data. UTDRM introduces a novel approach to synthetically generate large-scale topical claims from fact-checking articles. A comprehensive comparison with other out-of-the-box, unsupervised, and supervised models confirm the efficacy of UTDRM in retrieving accurate debunked claims. In general, UTDRM-MPNet and UTDRM-RoBERTa consistently achieve the highest scores across all datasets, with UTDRM-MPNet exhibiting slightly better performance.

Furthermore, this study emphasises the importance of corpus size, demonstrating that larger corpora contribute to improved retrieval performance. The paper also examines how different factors, such as the quantity of synthetically generated claims used and the *entity inoculation* method, influence the performance of UTDRM. While *entity inoculation* shows benefits in making models entity agnostic, its effectiveness may be limited to cases involving narratives that adapt and propagate with different entities.

Additionally, this paper experiments with state-of-the-art LLMs as listwise rerankers and compares them to our UTDRM method. While LLMs exhibit slight performance improvements over UTDRM on some datasets, their use comes at the cost of lower computational efficiency, making UTDRM a more practical choice for real-time applications.

Finally, UTDRM allows models to adapt and learn from synthetically generated topical claims in real-time; thus providing significant benefits in combating ever-evolving topical misinformation.

6.9 Limitations and Future Work

The present work acknowledges certain limitations and identifies several avenues for future improvement. Firstly, this study focused solely on English-language datasets and did not explore cross-lingual retrieval. However, the UTDRM approach can be replicated and adapted to other languages using pre-trained multilingual language models. Conducting cross-lingual experiments would provide a more comprehensive understanding of UTDRM's performance and applicability in diverse linguistic contexts, thereby extending its potential impact in combating misinformation on a global scale. Additionally, future work can include testing on a broader range of fact-checking articles and exploring novel approaches to further improve the information retrieval models used in UTDRM.

Declarations

Ethics Statement

This research has received ethics approval by the Sheffield University Ethics Board. While this paper involves generating false claims for research purposes, its overarching goal is to develop effective techniques for identifying already debunked narratives. The synthetically generated false claims dataset is solely for evaluation and will only be made available to academic researchers following careful vetting and a signed contract, in order to prevent public harm or spreading of misinformation. Furthermore, the research demonstrates that UTDRM is an effective method for training debunked-narrative retrieval models without the need for annotations, which are often time-consuming, expensive, and limited in scale. The overall aim is to promote ethical technology use and advance misinformation debunking efforts for the benefit of fact-checkers and users in general.

Availability of data and material

The datasets supporting the conclusions of this article are publicly available: 1) Snopes and Politifact (Shaar et al., 2020a) https://github.com/sshaar/That-is-a-Known-Lie 2) CLEF 22 2A-EN and 2B-EN (Nakov et al., 2022c) https://sites.google.com/view/clef2022-checkthat 3) CLEF 21 2A-EN and 2B-EN (Nakov et al., 2021b) https://sites.google.com/view/clef2021-checkthat 4) CLEF 20 2A-EN (Shaar et al., 2020b) https://sites.google.com/view/clef2020-checkthat. To facilitate repeatability, UTDRM models and code are made publicly available at https://github.com/iknoorjobs/UTDRM.

Funding

This research has been partially supported by a UKRI grant EP/W011212/1 and the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities" - Grant Agreement n.871042("SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (http://www.sobigdata.eu)).

Abbreviations

LLMs, Large Language Models; BERT, Bidirectional Encoder Representations from Transformers; RoBERTa, Robustly Optimized BERT Approach; UTDRM, Unsupervised Method for Training Debunked-Narrative Retrieval Models; LLaMA, Large Language Model Meta AI; ChatGPT, Chat Generative Pre-trained Transformer; BM25, Best Match 25; TF-IDF, Term Frequency - Inverse Document Frequency; QL, Query Likelihood model; DFR, Divergence From Randomness; CLEF, Conference and Labs of the Evaluation Forum; T5, Text-to-Text Transfer Transformer; SimCSE, Simple Contrastive Learning of Sentence Embeddings; ICT, Inverse Cloze Task; TSDAE, Tranformer-based Denoising AutoEncoder; GPL, Generative Pseudo Labeling; MSMARCO, Microsoft Machine Reading Comprehension; MSE, Mean Squared Error; AFP, Agence France-Presse; GPT-3.5, Generative Pretrained Transformer versions 3.5; ROUGE, Recall-Oriented Understudy for Gisting Evaluation; BLEU, Bilingual Evaluation Understudy; MNRL, Multiple Negatives Ranking Loss; MPNet, Masked and Permuted Pre-training for Language Understanding; ANCE, Approximate Nearest Neighbor Negative Contrastive Estimation; BT, Back-Translation; SOTA, State-Of-The-Art; LambdaMART, Lambda Multiple Additive Regression Trees; MRR, Mean Reciprocal Rank; MAP, Mean Average Precision; GPE, Geopolitical Entities; PERSON, Person; ORG, Organisation; NLP, Natural Language Processing; LRL, Listwise Re-ranker with a Large Language Model.

6.10 Appendix

6.10.1 Hyperparameters

For the T5 claim generator, we fine-tune the base variant of the T5 model[14] using a constant learning rate of $1e-4$ for 2 epochs, with a batch size of 12. The maximum input tokens allowed is 512, and the maximum output tokens is set to 64.

The training details for the neural retrieval model are as follows. `UTDRM-RoBERTa` is fine-tuned for two epochs with a batch size of 64 and a learning rate of $4e-5$. For `UTDRM-MPNet`, we fine-tune it for one epoch with a batch size of 64 and a learning rate of $8e-7$. The maximum input sequence length is set to 350, the optimiser used is AdamW and we use linear warmup as the learning rate scheduler. Hard negatives for training the neural retrieval model are mined using the *all-mpnet-base-v2*[15] and *all-MiniLM-L12-v2*[16] models because of their demonstrated efficacy [17]. Both `UTDRM-RoBERTa` and `UTDRM-MPNet` are validated using the respective dataset's validation set, and we manually tune the hyperparameters based on the evaluation metrics (Section 6.4.3). The hyperparameter bounds are as follows: 1) Epochs range from 1 to 5, 2) Learning rate ranges from $1e-7$ to $1e-5$, and

[14] https://huggingface.co/t5-base

[15] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[16] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[17] https://www.sbert.net/docs/pretrained_models.html

3) Batch size ranges from 8 to 64, limited by the GPU requirements of the model. The training time for each epoch ranges from 10 to 15 minutes.

For the baselines, BT and ICT use the same hyperparameters as `UTDRM-RoBERTa` to ensure a fair comparison. For SimCSE and TSDAE, we use the same hyperparameters as stated by the authors in their respective papers (Gao et al., 2021; Wang et al., 2021). Finally, all experiments are conducted on a machine with a 24GB NVIDIA GeForce RTX 3090.

6.10.2 Influence of ChatGPT Claims

Table 6.8 shows the performance of the `UTDRM-MPNet` model trained using different numbers of generated claims using ChatGPT ($N = 1$, $N = 2$, $N = 6$, $N = 10$). The datasets are divided into two categories: *Twitter-based* datasets (Snopes, CLEF 22 2A, CLEF 21 2A, CLEF 20 2A) and *political-based* datasets (Politifact, CLEF 22 2B, CLEF 21 2B).

From Table 6.8, we can observe that the model generally performs better on *Twitter-based* datasets, with the highest MAP@1 and MRR values of 0.945 and 0.962 respectively, recorded on the CLEF 20 2A dataset with $N = 6$ and $N = 10$ generated claims. In contrast, performance on *political-based* datasets is comparatively lower, with the highest MAP@1 and MRR values of 0.512 and 0.612 respectively, both recorded on the Politifact dataset with six generated claims ($N = 6$). Furthermore, the performance generally tends to improve with more generated claims, however, there are exceptions. On the Snopes and CLEF 21 2A datasets, performance dips slightly when increasing generated claims from $N = 2$ to $N = 10$. Overall, these observations suggest that the optimal number of claims to generate for best performance can vary depending on the specific dataset and whether it is *Twitter-based* or *political-based*.

6.10.3 Influence of T5 Claims

Table 6.8 shows the performance of the `UTDRM-MPNet` model trained using different numbers of generated claims using T5 ($N = 1$, $N = 2$, $N = 6$, $N = 10$). On the *Twitter-based* datasets, the model reaches peak performance on the CLEF 20 2A dataset with $N = 6$ generated claims (MAP@1 = 0.935 and MRR = 0.957). On *political-based* datasets, the model achieves maximum performance on the Politifact dataset with $N = 6$ generated claims (MAP@1 = 0.516 and MRR = 0.637). In general, finding an optimal number of generated claims for the best performance varies depending on the dataset, and the pattern is different from that of the ChatGPT generated claims (Section 6.10.2).

Table 6.8: *Influence of ChatGPT generated claims. The highest scores for each dataset and metric are in* ***bold****.*

Datasets	**Metrics**	**ChatGPT Generated Claims**			
		N=1	**N=2**	**N=6**	**N=10**
Snopes	**MAP@1**	0.811	**0.826**	0.813	0.811
	MRR	0.869	**0.882**	0.880	0.879
CLEF 22 2A	**MAP@1**	0.904	0.909	0.919	**0.923**
	MRR	0.926	0.929	0.937	**0.940**
CLEF 21 2A	**MAP@1**	0.876	**0.901**	**0.901**	0.896
	MRR	0.915	**0.931**	**0.931**	0.929
CLEF 20 2A	**MAP@1**	0.940	0.935	**0.945**	**0.945**
	MRR	0.956	0.955	**0.962**	**0.962**
Average	**MAP@1**	0.883	0.893	**0.894**	**0.894**
Twitter-based	**MRR**	0.917	0.924	**0.928**	**0.928**
Politifact	**MAP@1**	0.461	0.477	**0.512**	0.496
	MRR	0.572	0.593	**0.612**	0.606
CLEF 22 2B	**MAP@1**	0.346	0.346	0.377	**0.385**
	MRR	0.423	0.424	0.448	**0.458**
CLEF 21 2B	**MAP@1**	0.310	0.310	0.335	**0.342**
	MRR	0.379	0.381	0.403	**0.411**
Average	**MAP@1**	0.372	0.378	**0.408**	0.407
Political-based	**MRR**	0.458	0.466	0.488	**0.492**

Table 6.9: *Influence of T5 generated claims. The highest scores for each dataset and metric are in* ***bold****.*

Datasets	Metrics	T5 Generated Claims			
		N=1	**N=2**	**N=6**	**N=10**
Snopes	**MAP@1**	0.811	0.821	0.846	**0.851**
	MRR	0.870	0.879	0.898	**0.900**
CLEF 22 2A	**MAP@1**	0.900	0.909	**0.928**	**0.928**
	MRR	0.923	0.930	**0.946**	0.945
CLEF 21 2A	**MAP@1**	0.886	0.901	**0.916**	0.906
	MRR	0.923	0.933	**0.938**	0.937
CLEF 20 2A	**MAP@1**	**0.935**	**0.935**	**0.935**	**0.935**
	MRR	0.954	0.955	**0.957**	0.955
Average	**MAP@1**	0.883	0.891	**0.906**	0.905
Twitter-based	**MRR**	0.917	0.924	**0.935**	0.934
Politifact	**MAP@1**	0.484	**0.516**	**0.516**	0.500
	MRR	0.598	0.628	**0.637**	0.627
CLEF 22 2B	**MAP@1**	0.392	0.408	**0.415**	**0.415**
	MRR	0.451	0.473	0.487	**0.490**
CLEF 21 2B	**MAP@1**	0.348	0.361	0.354	**0.367**
	MRR	0.403	0.420	0.431	**0.439**
Average	**MAP@1**	0.408	**0.428**	**0.428**	0.427
Political-based	**MRR**	0.484	0.507	0.518	**0.519**

Table 6.10: *Influence of T5 generated claims. The highest scores for each dataset and metric are in* ***bold****.*

Table 6.11: *Examples showcasing the variation of similar debunked claims across multiple entities and contexts, with corresponding fact-check links. The text in **bold** shows difference in named entities between the claims.*

Debunked Claim	Fact-check Link
Claim 1: Crocodile swimming on a flooded street in the south Indian city of **Hyderabad**.	Link
Claim 2: Crocodile seen during flood in **Patna, Bihar**.	Link
Claim 3: Crocodile spotted in the waterlogged streets of **Bengaluru**.	Link
Claim 4: Crocodile seen during flood in **Aligarh, Uttar Pradesh**.	Link
Claim 5: Crocodile seen during flood in **Florida**.	Link
Claim 1: Sushant Singh Rajputs father **KK Singh** has demanded a CBI inquiry into his death.	Link
Claim 2: **PM Modi** has ordered for a CBI inquiry into Sushant Singh Rajputs death.	Link
Claim 3: **Amit Shah** ordered CBI probe for investigating Sushant Singh Rajput's death.	Link
Claim 1: A video in which a **woman** suffers a seizure on the floor in an **Argentine hospital** after the woman was vaccinated against covid-19 .	Link
Claim 2: A video shows a **man** fainting after receiving the Covid-19 vaccine in **Indonesia's West Nusa Tenggara** province.	Link
Claim 1: A video shows the meeting of the **Pacific Ocean** and **Atlantic ocean**, but without that they mix.	Link
Claim 2: A video which shows a place where the **Indian Ocean** meets the **Atlantic ocean**, and the waters of the two oceans do not mix.	Link
Claim 3: A video shows the meeting of the **Gulf of Mexico** and **Mississippi River**, but without mixing.	Link

6.10.4 *Entity Inoculation* Motivation

Table 6.11 illustrates an intriguing aspect of misinformation - it tends to replicate across diverse contexts and entities, applying similar narratives or themes to varied situations. The first example shows similar claims about "Crocodiles". These falsehoods involve the sighting of crocodiles in flooded city streets but vary by location — Hyderabad, Patna, Bengaluru, Aligarh and Florida. This shows how a single false narrative can be adapted to fit multiple geographical contexts, fueling misinformation in different locations. Similarly, The second example shows claims around "Sushant Singh Rajput's Death" (Table 6.11). These false narratives revolve around the demand for a CBI inquiry into the actor's death. The narrative remains consistent but the entities change - one claim implicates Rajput's father, KK Singh, and the other brings in PM Modi and Amit Shah. These falsehoods illustrate how misinformation can persist by switching the characters.

In summary, Table 6.11 highlights the importance of our adopted approach of *entity inoculation*, as detailed in Section 6.6.3. This method involves replacing one randomly chosen named entity in the generated claims with another random named entity, with the intent to mimic real-world scenarios where similar misinformation narratives disseminate involving different entities. This emphasises both the adaptability and resilience of misinformation, underlining the need for effective methods like *entity inoculation* to detect debunked narratives.

Chapter 7

Conclusions

The collection of research papers presented in this thesis collectively addresses the challenges presented by false information, forming a cohesive piece of research that spans the COVID-19 infodemic, Ukraine-related disinformation, and the broader landscape of automated fact-checking. This chapter concludes this thesis by summarising key findings and contributions while proposing potential avenues for future research.

7.1 Summary of Thesis

The journey began with the first paper (Chapter 2), which delved into the global infodemic triggered by the COVID-19 pandemic. It uncovers 10.3% of instances where similar false narratives related to COVID-19 are independently debunked multiple times, highlighting the widespread dissemination of misinformation during the pandemic. It underscores not only the extensive efforts of fact-checkers in debunking false narratives but also highlights the redundancy and waste of resources in repeatedly debunking similar narratives across different countries and languages. It highlights the prevalence of similar false narratives related to general medical advice across multiple countries, particularly on Facebook, where users unknowingly perpetuate false narratives despite the existence of fact-check articles. Finally, it set the stage for the subsequent papers by emphasising the importance of a cross-lingual debunked narrative search tool in the fact-checking pipeline. Additionally, it emphasises the need for social media platforms to adopt similar technology at scale, thereby optimising the utilisation of scarce fact-checker resources.

The second paper (Chapter 3) complements the first by conducting a comparative analysis of the spread of Ukraine-related disinformation and debunks on Twitter.

With respect to debunks, it established that around 18% are about false claims for which debunks have already been posted by another fact-checking organisation in a different country or language. The study's findings resonate with the findings of the first paper, suggesting the need for machine translation and reliable cross-lingual search tools to quickly find and match similar debunked claims; thus, reducing the time lag between the emergence of disinformation and the publication of corresponding debunks in a different language. Furthermore, a comparative analysis of Ukraine-related disinformation and debunks on Twitter reveals that disinformation is shared and retweeted significantly more than debunks. Finally, this paper also uses statistical methods like Granger Causality to underscore the impact of debunks in limiting the spread of disinformation, albeit not instantly. This forms the basis for the rest of the papers, indicating that even if fact-checkers or some automated methods regularly debunk information online, the debunks eventually have a positive impact on reducing disinformation.

Building on the foundation laid in the first and second papers, the third paper (Chapter 4) provides a concrete method for improving multilingual access to reliable COVID-19 information, crucial for mitigating the impacts of the infodemic. It proposes a Multistage BiCross Encoder, which is a three-stage method consisting of an initial lexical retrieval using Okapi BM25 algorithm followed by a transformer-based bi-encoder and cross-encoder to effectively re-rank the documents using sentence-level score aggregation with respect to the query. The independently evaluated Multilingual Information Access (MLIA) results show that this simple hybrid method outperforms other state-of-the-art methods across various metrics. It achieves high precision (P@5$\geq$0.8 for the best-performing run) in retrieving top-ranked documents and maintains a high recall for all retrieved documents. Overall, this paper aligns with the overarching theme of enhancing cross-lingual search technologies.

Extending the discussion initiated in earlier papers, the fourth paper (Chapter 5) focuses on cross-lingual debunked narrative retrieval (X-DNR) for automated fact-checking. It introduces MMTweets, a novel benchmark dataset that stands out, featuring cross-lingual pairs, human annotations, fine-grained labels, and images, making it a comprehensive resource compared to other datasets. Initial tests benchmarking state-of-the-art cross-lingual retrieval models reveal that dealing with multiple languages in the MMTweets dataset poses a challenge, indicating a need for further improvement in retrieval models. Nevertheless, inspired by the third paper, this paper introduces two tailored multistage retrieval methods (*BE+CE* and *BE+GPT3.5*) that demonstrate superior performance over other state-of-the-art models, achieving an average MRR score of 0.795. Finally, the findings also suggest some knowledge transfer across languages and datasets, which is especially valuable in scenarios where language-specific models are not available or feasible to train.

Continuing the ongoing quest for more optimised and efficient models, the fifth paper (Chapter 6) introduces UTDRM, an unsupervised method for training debunked-narrative retrieval models that effectively overcomes the reliance on manually annotated training data. UTDRM introduces a novel approach to synthetically generate large-scale topical claims from fact-checking articles. A comprehensive comparison with other out-of-the-box, unsupervised, and supervised models confirms the efficacy of UTDRM in retrieving accurate debunked claims. It achieves an average MRR score of 0.934 and 0.505 for Twitter-based and political-based datasets, respectively. UTDRM's effectiveness in retrieving accurate debunked claims, combined with its ability to adapt and learn from synthetically generated topical claims in real-time, positions it as a practical choice for combating ever-evolving misinformation and disinformation.

7.2 Research Questions Discussion

This section provides a comprehensive discussion addressing the two primary research questions and their corresponding sub-research questions outlined in the introduction (Section 1.3).

RQ1: To what extent do false narratives propagate even after being debunked by professional fact-checkers, and how does their spread compare to the dissemination of corresponding debunks?

This research question sheds light on the persistence of false narratives even after professional debunking (Chapters 2 and 3). The spatiotemporal analysis (Chapter 2) sets the stage for understanding the persistence of false narratives, while the Ukraine-related disinformation analysis (Chapter 3) emphasises the importance of timely debunking and the subsequent impact on mitigating disinformation.

1. **Spatiotemporal Analysis (Sub-RQ1a):** Chapter 2 uncovered the spatiotemporal characteristics of the spread of debunked narratives related to COVID-19. It revealed that a considerable percentage (10.3%) of false narratives are independently debunked multiple times across languages, platforms, and modalities. The analysis highlighted the redundancy in fact-checking efforts and the need for cross-lingual tools for detecting debunked narratives.

2. **Ukraine-related Disinformation (Sub-RQ1b):** In Chapter 3, the comparative analysis of Ukraine-related disinformation and debunks on Twitter demonstrated that disinformation is shared and retweeted significantly more than debunked content. With respect to debunks, it established that around 18% are focused on false claims for which debunks have already been posted by another fact-checking organisation in a different country or

language. On the other hand, the Granger causality test provided insights into the delayed but impactful role of debunks in limiting the spread of disinformation. This analysis also emphasised the importance of timely cross-lingual debunked narrative retrieval, setting the stage for the exploration of effective detection methods.

RQ2: What are the effective ways by which we can detect and alleviate the repeated dissemination of multilingual debunked narratives?

This research question aimed at detecting the repeated dissemination of debunked narratives by framing it as a cross-lingual information retrieval problem. Chapters 4, 5, and 6 collectively addressed this question, presenting innovative solutions.

1. **Multistage BiCross Encoder (Sub-RQ2a):** Chapter 4 proposed the Multistage BiCross Encoder, a hybrid method for enhancing multilingual and cross-lingual information retrieval. Results from the Multilingual Information Access (MLIA) shared task showcased the effectiveness of the proposed approach in improving both precision and recall, offering a solution for efficient access to reliable information during infodemics.

2. **Cross-Lingual Debunked Narrative Retrieval (Sub-RQ2b):** In Chapter 5, MMTweets, a benchmark dataset for cross-lingual debunked narrative retrieval, was introduced. Additionally, two specifically designed multistage retrieval methods were presented (*BE+CE* and *BE+GPT3.5*) that surpasses other state-of-the-art models across various languages and metrics. The chapter highlights the importance of considering cross-lingual aspects in debunked narrative retrieval. The chapter also delved into challenges and opportunities, offering valuable insights into the dynamics of knowledge transfer across languages and datasets.

3. **Unsupervised Training for Debunked Narrative Retrieval (Sub-RQ2c):** Chapter 6 presented `UTDRM`, an unsupervised method for training debunked-narrative retrieval models. This approach showcased remarkable effectiveness in retrieving accurate debunked claims, offering a practical solution for real-time adaptation without relying on human annotations.

7.3 Future Work

The conclusions drawn from these papers open avenues for future research to build upon and extend the presented findings.

Improvements in Retrieval Models: The third, fourth, and fifth papers (Chapter 4, 5, 6) focus on the development of accurate and reliable cross-lingual retrieval methods. However, there is still room for improvement. Hence, future research can delve deeper into cross-lingual retrieval methods, addressing identified challenges and exploring innovative techniques to enhance performance. For instance, 1) development of more sophisticated retrieval models that have in-depth understanding of both language and context, along with the capability to differentiate among closely related debunked narratives 2) effective usage of Language Models (LLMs) and in-context learning for debunked narrative retrieval tasks, such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), usage of Toolformer (Schick et al., 2024) for factual lookup 3) integration of multimodal information into debunked narrative retrieval, including images and other content types 4) expanding on the entity inoculation approach for training (discussed in Section 6.6.3) by incorporating trending topical entities instead of random replacements, aiming to better simulate real-world scenarios. These enhancements aim to create robust systems capable of detecting debunked narratives across various formats.

Debunked Narrative Classification: The fourth and fifth papers (Chapter 5, 6) primarily focus on the retrieval of debunked narratives. However, the debunked narrative retrieval models lack the ability to determine whether the given query claim has been debunked. This limitation is inherent in the information retrieval task, as it relies on sorting a corpus of documents based on a relevance score. To this end, prior work (Kazemi et al., 2022, 2021) addresses the task of debunked-narrative retrieval as a binary classification problem and trains a support vector machine model to classify claims. However, formulating it as a classification problem is computationally not scalable due to its quadratic complexity. Therefore, future work is needed that not only retrieves but also classifies whether the query claim has been debunked and, at the same time, is computationally efficient. For instance, classifying the query claim based on the top of the retrieved debunked claim. This will not only expedite fact-checking but will also reduce the overhead of manually going through every top-ranked debunk.

Implications of Overlapping Debunks: The assumption that overlap in debunks (i.e., multiple debunks by different fact-checkers for the same claims) is an indicator of redundant effort, while valid, is not entirely foolproof. Different fact-checkers may have legitimate reasons to publish their own versions of debunks, update existing ones, or add more evidence. Additionally, while tools supporting cross-language retrieval of similar claims can expedite the creation of debunks, they may not entirely eliminate duplicated debunks due to these necessary variations. Another assumption in this research is that having debunks available in people's native languages will decrease the sharing of related misinformation. However, this is not always the case, as many individuals either do not check or are unaware of debunks. Future work should delve deeper into these aspects to assess user engagement with debunks and the actual impact on misinformation sharing behaviour, aiming to develop more effective fact-checking strategies. Furthermore, future research should also explore cases where fact-checkers produce differing opinions about the same debunked claims that are not fully aligned. In such cases, it is essential to evaluate the credibility of the fact-checkers and work towards unifying their opinions to refine methodologies and improve user trust.

Extending the Datasets: To enhance generalisability, there is a need to extend the dataset to include claims from other social media platforms and domains. Most of the existing datasets either focus on Twitter posts or are statements taken directly from political debates. However, false narratives can extend beyond social media posts to include articles that encompass a more extensive and coherent story or account, which may consist of multiple false claims. Therefore, there is a need for diverse datasets that capture the variety of formats and sources through which false claims can propagate. Expanding the dataset to include content from platforms such as Facebook, Instagram, Reddit, YouTube and online news articles will contribute to a more comprehensive understanding of the debunked narratives and improve the robustness of debunked narrative retrieval models. Additionally, incorporating claims from various domains beyond politics, such as health, science, and technology, will further enrich the dataset and enable the development of models with broader applicability across different contexts. Finally, including datasets that focus on specific types of false information, such as misinformation, disinformation, and malinformation, may provide different insights. This will enhance our understanding of how various forms of false information spread and how they can be effectively countered.

In summary, the research presented in these papers provides valuable new insights into addressing the challenges posed by false information in various contexts. We hope that the findings in this thesis benefit fellow researchers as well as fact-checkers who can make use of our resources in downstream applications to quickly debunk the ever-growing numbers of unsubstantiated claims on the internet. The future work outlined aims to further refine, generalise, and extend the proposed solutions, contributing to the ongoing efforts to combat misinformation on a global scale.

Bibliography

(AFP), A. F.-P. (2024a). Agence france-presse (AFP) — afp.com/. https://www.afp.com/. [Accessed 03-02-2024].

(AFP), A. F.-P. (2024b). Agence france-presse (AFP) — afp.com/. https://factuel.afp.com/non-bill-gates-na-pas-propose-dimplanter-une-puce-electronique-la-population. [Accessed 03-02-2024].

Aguerri, J., Santisteban, M., and Miró-Llinares, F. (2022). The fight against disinformation and its consequences: Measuring the impact of "Russia state-affiliated media" on Twitter. *SocArXiv*.

AI, O. (2024). Open ai — openai.com/. https://platform.openai.com/docs/models. [Accessed 03-02-2024].

Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F., et al. (2023). Information retrieval meets large language models: A strategic report from chinese ir community. *AI Open*.

Akkalyoncu Yilmaz, Z., Yang, W., Zhang, H., and Lin, J. (2019). Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China. Association for Computational Linguistics.

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Martino, G. D. S., Abdelali, A., Sajjad, H., Darwish, K., et al. (2020). Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. *ArXiv preprint*, abs/2007.07996.

Ali, Z. S., Mansour, W., Elsayed, T., and Al-Ali, A. (2021). Arafacts: the first large arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Anderson, J. and Rainie, L. (2017). The future of truth and misinformation online.

AP (2024). Eating alkaline foods will not kill the coronavirus.

API, T. (2024). Twitter api — twitter.com/. https://developer.twitter.com/en/docs/twitter-api. [Accessed 03-02-2024].

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., et al. (2023). Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Aslam, J. A. and Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., and Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Babacan, K. and Tam, M. S. (2022). The information warfare role of social media: Fake news in the russia-ukraine war. *Erciyes İletişim Dergisi*, (3):75–92.

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, 182:104123.

Barrón-Cedeño, A., Alam, F., Caselli, T., Da San Martino, G., Elsayed, T., Galassi, A., Haouari, F., Ruggeri, F., Struß, J. M., Nandi, R. N., et al. (2023). The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority. In *European Conference on Information Retrieval*, pages 506–517. Springer.

Barrón-Cedeno, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., et al. (2020).

Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer.

BBC (2024). Hundreds dead' because of covid-19 misinformation.

Berger, A., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–199.

Bhatnagar, V., Kanojia, D., and Chebrolu, K. (2022). Harnessing abstractive summarization for fact-checked claim detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2934–2945.

Bhattacharya, S., Maddikunta, P. K. R., Pham, Q.-V., Gadekallu, T. R., Chowdhary, C. L., Alazab, M., Piran, M. J., et al. (2021). Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey. *Sustainable cities and society*, 65:102589.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bonisoli, G., Di Buono, M. P., Po, L., and Rollo, F. (2023). Dice: a dataset of italian crime event news. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2985–2995.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47:1007–1029.

Bontcheva, K., Posetti, J., Teyssou, D., Meyer, T., Gregory, S., Hanot, C., and Maynard, D. (2020). Balancing act: Countering digital disinformation while respecting freedom of expression. Technical report, United Nation Educational, Scientific and Cultural Organization.

Boomlive (2020). Can eating alkaline foods prevent or cure covid-19? a fact check | boom.

Boomlive (2024). Boomlive — boomlive.in. https://www.boomlive.in/. [Accessed 03-02-2024].

Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., and Sadeq, A. (2020). Team buster. ai at checkthat! 2020 insights and recommendations to improve fact-checking. In *CLEF (Working Notes)*.

Brennen, J. S., Simon, F., Howard, P. N., and Nielsen, R. K. (2020). Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7(3):1.

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., and Zhilin, O. (2020). The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, 1(3).

Burel, G., Farrell, T., and Alani, H. (2021). Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter. *Information Processing & Management*, 58(6):102732.

Burel, G., Farrell, T., Mensio, M., Khare, P., and Alani, H. (2020). Co-spread of misinformation and fact-checking content during the covid-19 pandemic. In *International Conference on Social Informatics*, pages 28–42. Springer.

Carlsson, F., Gyllensten, A. C., Gogoulou, E., Hellqvist, E. Y., and Sahlgren, M. (2021). Semantic Re-tuning with Contrastive Tension. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Casacuberta, F., Ceausu, A., Choukri, K., Declerck, T., Deligiannis, M., Di Nunzio, G. M., Domingo, M., Eskevich, M., Ferro, N., García-Martínez, M., Grouin, C., Herranz, M., Jacquet, G., Papavassiliou, V., Piperidis, S., Prokopidis, P., and Zweigenbaum, P. (2021). The Covid-19 MLIA@ Eval initiative: Developing multilingual information access systems and resources for Covid-19. https://bitbucket.org/covid19-mlia/organizers-overall/src/master/report/.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Chakraborty, T., La Gatta, V., Moscato, V., and Sperlì, G. (2023). Information retrieval algorithms and neural ranking models to detect previously fact-checked information. *Neurocomputing*, 557:126680.

Chang, W., Yu, F. X., Chang, Y., Yang, Y., and Kumar, S. (2020). Pre-training Tasks for Embedding-based Large-scale Retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chen, Q., Zhang, Y., Evans, R., and Min, C. (2021). Why Do Citizens Share COVID-19 Fact-Checks Posted by Chinese Government Social Media Accounts? The

Elaboration Likelihood Model. *International Journal of Environmental Research and Public Health*, 18(19):10058.

Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2021). Aschern at CheckThat! 2021: lambda-calculus of fact-checked claims. *Faggioli et al.[12]*.

Clipa, T. and Di Nunzio, G. M. (2020). A study on ranking fusion approaches for the retrieval of medical publications. *Information*, 11(2):103.

Cocuyo, E. (2020). ¿consumir alimentos «más alcalinos» elimina el virus que causa el covid-19?

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Cormack, G. V., Clarke, C. L., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.

Dementieva, D., Kuimov, M., and Panchenko, A. (2023). Multiverse: Multilingual evidence for fake news detection. *Journal of Imaging*, 9(4):77.

Dementieva, D. and Panchenko, A. (2021). Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Di Domenico, G., Sit, J., Ishizaka, A., and Nunan, D. (2021). Fake news, social media and marketing: A systematic review. *Journal of Business Research*, 124:329–341.

Di Nunzio, G. M., Dosso, D., Fabris, A., Faggioli, G., Ferro, N., Giachelle, F., Irrera, O., Marchesin, S., Piazzon, L., Purpura, A., et al. (2020). Unipd at covid-19 mlia. *MLIA COVID-19*.

Di Nunzio, G. M., Eskevich, M., and Ferro, N. (2021). The Covid-19 MLIA@ Eval initiative: Overview of the multilingual semantic search task. https://bitbucket.org/covid19-mlia/organizers-task2/src/master/report/.

Dutta, S., Dhar, R., Guha, P., Murmu, A., and Das, D. (2022). A multilingual dataset for identification of factual claims in indian twitter. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 88–92.

Ecker, U. K., Lewandowsky, S., and Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, 38:1087–1100.

EFE, A. (2024). Agencia EFE — efe.com. https://www.efe.com/. [Accessed 03-02-2024].

Erlich, A. and Garner, C. (2021). Is pro-Kremlin Disinformation Effective? Evidence from Ukraine. *The International Journal of Press/Politics*, page 19401612211045221.

EUvsDisinfo (2024). Euvsdisinfo fact-check database.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Fatos, A. (2020). Dieta rica em alimentos alcalinos não é capaz de eliminar o coronavírus | aos fatos.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

Fox, E. A. and Shaw, J. A. (1994). Combination of multiple searches. *NIST special publication SP*, 243.

Frick, R. A. and Vogel, I. (2022). Fraunhofer SIT at CheckThat! 2022: ensemble similarity estimation for finding previously fact-checked claims. *Working Notes of CLEF*.

FullFact (2024a). Fullfact: Fact checks featuring claims seen in the uk about the ongoing conflict in israel and gaza.

FullFact (2024b). Fullfact report tracks fake covid-19 news across five countries – society of editors.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gelfert, A. (2018). Fake news: A definition. *Informal logic*, 38(1):84–117.

Gerber, T. P. and Zavisca, J. (2016). Does Russian propaganda work? *The Washington Quarterly*, 39(2):79–98.

Gormley, C. and Tong, Z. (2015). *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc.".

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Graves, L. and Cherubini, F. (2016). The rise of fact-checking sites in europe. *Digital News Project Report*.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guess, A. M., Nyhan, B., and Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature human behaviour*, 4(5):472–480.

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

Gupta, A. and Srikumar, V. (2021a). X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Gupta, A. and Srikumar, V. (2021b). X-fact: A new benchmark dataset for multilingual fact checking. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Hammache, A. and Boughanem, M. (2020). Term position-based language model for information retrieval. *Journal of the Association for Information Science and Technology*.

Hammouchi, H. and Ghogho, M. (2022). Evidence-aware multilingual fake news detection. *Ieee Access*, 10:116808–116818.

Haque, M. M., Yousuf, M., Arman, Z., Rony, M. M. U., Alam, A. S., Hasan, K. M., Islam, M. K., and Hassan, N. (2018). Fact-checking initiatives in bangladesh, india, and nepal: a study of user engagement and challenges. *arXiv preprint arXiv:1811.01806*.

Hardalov, M., Chernyavskiy, A., Koychev, I., Ilvovsky, D., and Nakov, P. (2022). CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 266–285.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *ArXiv preprint*, abs/1705.00652.

Hiemstra, C. and Jones, J. D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664.

Hofstätter, S., Althammer, S., Schröder, M., Sertkan, M., and Hanbury, A. (2020). Improving efficient neural ranking models with cross-architecture knowledge distillation. *ArXiv preprint*, abs/2010.02666.

Hofstätter, S. and Hanbury, A. (2019). Let's measure run time! extending the ir replicability infrastructure to include performance aspects. *ArXiv preprint*, abs/1907.04614.

Hövelmeyer, A., Boland, K., and Dietze, S. (2022). SimBa at CheckThat! 2022: lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way. *Working Notes of CLEF*.

Hu, X., Guo, Z., Chen, J., Wen, L., and Yu, P. S. (2023). Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2901–2912.

Huang, K.-H., Zhai, C., and Ji, H. (2022). CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Huggingface (2024a). Huggingface — huggingface.co/. `https://huggingface.co/eugene-yang/dpr-xlm-align-engtrained`. [Accessed 03-02-2024].

Huggingface (2024b). Huggingface — huggingface.co/. `https://huggingface.co/facebook/mcontriever-msmarco`. [Accessed 03-02-2024].

Huggingface (2024c). Huggingface — huggingface.co/. `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`. [Accessed 03-02-2024].

Huggingface (2024d). Huggingface — use variant which supports around 50 languages. `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2`. [Accessed 03-02-2024].

Humeau, S., Shuster, K., Lachaux, M.-A., and Weston, J. (2019). Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *ArXiv preprint*, abs/1905.01969.

IFCN (2024). International fact-checking network (ifcn) — poynter.org/. https://www.poynter.org/ifcn/. [Accessed 03-02-2024].

Ioffe, S. (2010). Improved consistent sampling, weighted minhash and l1 sketching. In *2010 IEEE international conference on data mining*, pages 246–255. IEEE.

Iwendi, C., Mahboob, K., Khalid, Z., Javed, A. R., Rizwan, M., and Ghosh, U. (2021). Classification of covid-19 individuals using adaptive neuro-fuzzy inference system. *Multimedia Systems*, pages 1–15.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Jiang, M., Gao, Q., and Zhuang, J. (2021). Reciprocal spreading and debunking processes of online misinformation: A new rumor spreading–debunking model with a case study. *Physica A: Statistical Mechanics and its Applications*, 565:125572.

Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with gpus. *ArXiv preprint*, abs/1702.08734.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Kazemi, A., Garimella, K., Gaffney, D., and Hale, S. (2021). Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

Kazemi, A., Li, Z., Pérez-Rosas, V., Hale, S. A., and Mihalcea, R. (2022). Matching tweets with applicable fact-checks across languages. *ArXiv preprint*, abs/2202.07094.

Klein, D. O. and Wueller, J. R. (2018). Fake news: A legal perspective. *Australasian Policing*, 10(2).

Köhler, J., Shahi, G. K., Struß, J. M., Wiegand, M., Siegel, M., Mandl, T., and Schütz, M. (2022). Overview of the clef-2022 checkthat! lab: Task 3 on fake news detection. In *CLEF (Working Notes)*, pages 404–421.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., et al. (2023). Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

La Gatta, V., Wei, C., Luceri, L., Pierri, F., and Ferrara, E. (2023). Retrieving false claims on twitter during the russia-ukraine conflict. *arXiv preprint arXiv:2303.10121*.

Lange-Ionatamishvili, E., Svetoka, S., and Geers, K. (2015). Strategic communications and social media in the Russia Ukraine conflict. *Cyber war in perspective: Russian aggression against Ukraine*, pages 103–111.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Leadstories (2020). Fact check: Alkaline diet does not prevent you from getting coronavirus | lead stories.

Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, C., Yates, A., MacAvaney, S., He, B., and Sun, Y. (2020a). Parade: Passage representation aggregation for document reranking. *ArXiv preprint*, abs/2008.09093.

Li, X., Liu, Y., Mao, J., He, Z., Zhang, M., and Ma, S. (2018). Understanding reading attention distribution during relevance judgement. In Cuzzocrea, A., Allan, J., Paton, N. W., Srivastava, D., Agrawal, R., Broder, A. Z., Zaki, M. J., Candan, K. S., Labrinidis, A., Schuster, A., and Wang, H., editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 733–742. ACM.

Li, Y., Jiang, B., Shu, K., and Liu, H. (2020b). Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *ArXiv preprint*, abs/2011.04088.

Limaye, R. J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., and Labrique, A. (2020). Building trust while influencing online covid-19 content in the social media world. *The Lancet Digital Health*, 2(6):e277–e278.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ma, J., Korotkov, I., Yang, Y., Hall, K., and McDonald, R. (2021). Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

Ma, X., Zhang, X., Pradeep, R., and Lin, J. (2023). Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

MacAvaney, S., Cohan, A., and Goharian, N. (2020). Sledge: A simple yet effective zero-shot baseline for coronavirus scientific knowledge search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4171–4179.

Maldita (2020). El bulo de que una dieta alcalina previene el contagio de coronavirus · maldita.es - periodismo para que no te la cuelen.

Mansour, W., Elsayed, T., and Al-Ali, A. (2022). Did i see it before? detecting previously-checked claims over twitter. In *European Conference on Information Retrieval*, pages 367–381. Springer.

Mansour, W., Elsayed, T., and Al-Ali, A. (2023). This is not new! spotting previously-verified claims over twitter. *Information Processing & Management*, 60(4):103414.

Martin-Valdivia, D.-G. M.-T. (2020). Sinai at mlia covid-19. *MLIA COVID-19*.

McGlynn, J., Baryshevtsev, M., and Dayton, Z. A. (2020). Misinformation more likely to use non-specific authority references: Twitter analysis of two covid-19 myths. *Harvard Kennedy School Misinformation Review*, 1(3).

Mejias, U. A. and Vokuev, N. E. (2017). Disinformation and the media: the case of Russia and Ukraine. *Media, culture & society*, 39(7):1027–1042.

Mihaylova, S., Borisova, I., Chemishanov, D., Hadzhitsanev, P., Hardalov, M., and Nakov, P. (2021). Dips at checkthat! 2021: Verified claim retrieval. In *CLEF (Working Notes)*, pages 558–571.

Miller, C., Inskip, C., Marsh, O., Arcostanzo, F., and Weir, D. (2022). Message-based Community Detection on Twitter.

Mu, Y., Jin, M., Grimshaw, C., Scarton, C., Bontcheva, K., and Song, X. (2023). Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Nakov, P. (2020). Can we spot the" fake news" before it was even written? *ArXiv preprint*, abs/2008.04374.

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., et al. (2022a). Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pages 368–392. CEUR Workshop Proceedings (CEUR-WS. org).

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J. M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., et al. (2022b). The CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *European Conference on Information Retrieval*, pages 416–428. Springer.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., and Martino, G. D. S. (2021a). Automated fact-checking for assisting human fact-checkers. *ArXiv preprint*, abs/2103.07769.

Nakov, P. and Da San Martino, G. (2021). Fake news, disinformation, propaganda, and media bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4862–4865.

Nakov, P., Da San Martino, G., Alam, F., Shaar, S., Mubarak, H., and Babulkov, N. (2022c). Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims. *Conference and Labs of the Evaluation Forum*.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., et al. (2021b). The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *ECIR (2)*.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. In Besold, T. R., Bordes, A., d'Avila Garcez, A. S., and Wayne, G., editors, *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. (2022). Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

Nielsen, D. S. and McConville, R. (2022). Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert. *ArXiv preprint*, abs/1901.04085.

Nogueira, R., Lin, J., and Epistemic, A. (2019a). From doc2query to docttttquery. *Online preprint*.

Nogueira, R., Yang, W., Lin, J., and Cho, K. (2019b). Document expansion by query prediction. *ArXiv preprint*, abs/1904.08375.

Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Nyhan, B. and Reifler, J. (2015). Estimating fact-checking's effects. *Arlington, VA: American Press Institute*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Panchendrarajan, R. and Zubiaga, A. (2024). Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *arXiv preprint arXiv:2401.11969*.

Park, C. Y., Mendelsohn, J., Field, A., and Tsvetkov, Y. (2022). VoynaSlov: A Data Set of Russian Social Media Activity during the 2022 Ukraine-Russia War. *arXiv preprint arXiv:2205.12382*.

Park, S., Park, J. Y., Chin, H., Kang, J.-h., and Cha, M. (2021). An experimental study to understand user experience and perception bias occurred by fact-checking messages. In *Proceedings of the Web Conference 2021*, pages 2769–2780.

Pavleska, T., Školkay, A., Zankova, B., Ribeiro, N., and Bechmann, A. (2018). Performance analysis of fact-checking organizations and initiatives in europe: a critical overview of online platforms fighting fake news. *Social media and convergence*, 29:1–28.

Pennycook, G. and Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Peters, C. (2000). Cross-Language Information Retrieval and Evaluation Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 21–22, 2000 Revised Papers. In *Conference proceedings CLEF*, page 132. Springer.

Pikuliak, M., Srba, I., Moro, R., Hromadka, T., Smolen, T., Melisek, M., Vykopal, I., Simko, J., Podrouzek, J., and Bielikova, M. (2023). Multilingual previously fact-checked claim retrieval. *arXiv preprint arXiv:2305.07991*.

Pillai, R. M. and Fazio, L. K. (2021). The effects of repeating false and misleading information on belief. *Wiley Interdisciplinary Reviews: Cognitive Science*, page e1573.

PolitiFact (2024). PolitiFact — politifact.com. https://www.politifact.com/. [Accessed 03-02-2024].

Ponte, J. M. and Croft, W. B. (2017). A language modeling approach to information retrieval. In *ACM SIGIR Forum*, volume 51, pages 202–208. ACM New York, NY, USA.

Posetti, J. and Bontcheva, K. (2020). Policy brief 1, disinfodemic: Deciphering covid-19 disinformation. Technical report, United Nation Educational, Scientific and Cultural Organization.

Posetti, J., Bontcheva, K., et al. (2020). Disinfodemic: dissecting responses to covid-19 disinformation.

Pradeep, R., Nogueira, R., and Lin, J. (2021). The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *ArXiv preprint*, abs/2101.05667.

Procter, R., Catania, M. A., He, Y., Liakata, M., Zubiaga, A., Kochkina, E., and Zhao, R. (2023). Some observations on fact-checking work with implications for computational support. *arXiv preprint arXiv:2305.02224*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020a). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020b). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Recuero, R., Soares, F. B., Vinhas, O., Volcan, T., Hüttner, L. R. G., and Silva, V. (2022). Bolsonaro and the Far Right: How Disinformation About COVID-19 Circulates on Facebook in Brazil. *International Journal of Communication*, 16:24.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reis, J., Melo, P. d. F., Garimella, K., and Benevenuto, F. (2020). Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *ArXiv preprint*, abs/2006.02471.

Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L. L., and Hersh, W. R. (2020). Trec-covid: Rationale and structure of an information retrieval shared task for covid-19. *Journal of the American Medical Informatics Association*.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., and Van Der Linden, S. (2020). Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Saleh, S., Khojasteh, H. A., Sellat, H., and Pecina, P. (2020). Cuni-mtir at covid-19 mlia@ eval task 2. *MLIA COVID-19*.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2024). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Schuetz, S. W., Sykes, T. A., and Venkatesh, V. (2021). Combating covid-19 fake news on social media through fact checking: antecedents and consequences. *European Journal of Information Systems*, 30(4):376–388.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Shaar, S., Alam, F., Da San Martino, G., and Nakov, P. (2022). The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Shaar, S., Alam, F., Martino, G. D. S., and Nakov, P. (2021). Assisting the human fact-checkers: detecting all previously fact-checked claims in a document. *ArXiv preprint*, abs/2109.07410.

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020a). That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeno, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., et al. (2020b). Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In *CLEF (Working Notes)*.

Shahi, G. K. and Nandini, D. (2020). Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *ArXiv preprint*, abs/2006.11343.

Shang, X., Chen, Y., Fang, Y., Liu, Y., and Vincent, S. (2023). Amica: Alleviating misinformation for chinese americans. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3145–3149.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Shearer, E. (2021). More than eight-in-ten americans get news from digital devices. *Pew Research Center*.

Sheng, Q., Cao, J., Bernard, H. R., Shu, K., Li, J., and Liu, H. (2022). Characterizing multi-domain false news and underlying user effects on chinese weibo. *Information Processing & Management*, 59(4):102959.

Sheng, Q., Cao, J., Zhang, X., Li, X., and Zhong, L. (2021). Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5468–5481, Online. Association for Computational Linguistics.

Shi, P. and Lin, J. (2019). Cross-lingual relevance transfer for document retrieval. *ArXiv preprint*, abs/1911.02989.

Shliselberg, S.-H. M. and Dori-Hacohen, S. (2022). RIET Lab at CheckThat! 2022: Improving decoder based re-ranking for claim matching. *Working Notes of CLEF*.

Shu, K. and Liu, H. (2022). *Detecting fake news on social media*. Springer Nature.

Shu, R., Nakayama, H., and Cho, K. (2019). Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48.

Singh, I., Bontcheva, K., and Scarton, C. (2021a). The false covid-19 narratives that keep being debunked: A spatiotemporal analysis. *ArXiv preprint*, abs/2107.12303.

Singh, I., Bontcheva, K., Song, X., and Scarton, C. (2022). Comparative analysis of engagement, themes, and causality of ukraine-related debunks and disinformation. In *International Conference on Social Informatics*, pages 128–143. Springer.

Singh, I., Deepak, P., and Anoop, K. (2020). On the coherence of fake news articles. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 591–607. Springer.

Singh, I., Scarton, C., and Bontcheva, K. (2021b). Multistage bicross encoder for multilingual access to covid-19 health information. *PloS one*, 16(9):e0256874.

Singh, I., Scarton, C., and Bontcheva, K. (2023). Utdrm: unsupervised method for training debunked-narrative retrieval models. *EPJ Data Science*, 12(1):59.

Siwakoti, S., Yadav, K., Bariletto, N., Zanotti, L., Erdogdu, U., and Shapiro, J. N. (2021). How covid drove the evolution of fact-checking. *Harvard Kennedy School Misinformation Review*.

Snopes (2024). Snopes.com — snopes.com. https://www.snopes.com/. [Accessed 03-02-2024].

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., and Bontcheva, K. (2021). Classification aware neural topic model for covid-19 disinformation categorisation. *PloS one*, 16(2):e0247086.

Spenkuch, J. L. and Toniatti, D. (2016). Political advertising and election outcomes. *Kilts Center for Marketing at Chicago Booth–Nielsen Dataset Paper Series*, pages 1–046.

Stănescu, G. (2024). Informational war: Analyzing false news in the israel conflict. *SOCIAL SCIENCESANDEDUCATIONRESEARCHREVIEW*.

Swire, B., Berinsky, A. J., Lewandowsky, S., and Ecker, U. K. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society open science*, 4(3):160802.

Tasnim, S., Hossain, M. M., and Mazumder, H. (2020). Impact of rumors and misinformation on covid-19 in social media. *Journal of preventive medicine and public health*, 53(3):171–174.

Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., and Todorov, K. (2019). Claimskg: A knowledge graph of fact-checked claims. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*, pages 309–324. Springer.

Teyit (2020). Yeni koronavirüsün alkali beslenerek yok edilebileceği iddiası · teyit.org.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Thorne, J. and Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

TOI (2024). Covid-19: 'panic due to fake news' led to migrant exodus, no record of number of deaths, govt says.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv preprint*, abs/2302.13971.

Tsonis, A. A., Deyle, E. R., Ye, H., and Sugihara, G. (2018). Convergent cross mapping: theory and an example. *Advances in nonlinear geosciences*, pages 587–600.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Vo, N. and Lee, K. (2020). Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W. R., Lo, K., Roberts, K., Soboroff, I., and Wang, L. L. (2021). Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, K., Reimers, N., and Gurevych, I. (2021). TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang, K., Thakur, N., Reimers, N., and Gurevych, I. (2022). GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 27.

Watts, C. (2017). Extremist content and russian disinformation online: Working with tech to find solutions. *Foreign Policy Research Institute*.

Wikipedia (2024). Wikipedia — wikipedia.org/. `https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest`. [Accessed 03-02-2024].

Wu, J., Liu, Q., Xu, W., and Wu, S. (2022). Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P. N., Ahmed, J., and Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yablokov, I. (2022). Russian disinformation finds fertile ground in the West. *Nature Human Behaviour*, pages 1–2.

Yang, E., Nair, S., Chandradevan, R., Iglesias-Flores, R., and Oard, D. W. (2022). C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2512.

Yang, W., Zhang, H., and Lin, J. (2019). Simple applications of bert for ad hoc document retrieval. *ArXiv preprint*, abs/1903.10972.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., and Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 87–94, Online. Association for Computational Linguistics.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., and Sharif, S. (2021). An analysis of covid-19 vaccine sentiments and opinions on twitter. *International Journal of Infectious Diseases*, 108:256–262.

Zeng, X., Abumansour, A. S., and Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Zhang, D., Vakili Tahami, A., Abualsaud, M., and Smucker, M. D. (2022a). Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2099–2104.

Zhang, E., Gupta, N., Tang, R., Han, X., Pradeep, R., Lu, K., Zhang, Y., Nogueira, R., Cho, K., Fang, H., and Lin, J. (2020a). Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 31–41, Online. Association for Computational Linguistics.

Zhang, H., Cormack, G. V., Grossman, M. R., and Smucker, M. D. (2020b). Evaluating sentence-level relevance feedback for high-recall information retrieval. *Information Retrieval Journal*, 23(1):1–26.

Zhang, Y., Guo, B., Ding, Y., Liu, J., Qiu, C., Liu, S., and Yu, Z. (2022b). Investigation of the determinants for misinformation correction effectiveness on social media during COVID-19 pandemic. *Information Processing & Management*, 59(3):102935.

Zheng, J., Baheti, A., Naous, T., Xu, W., and Ritter, A. (2022). Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 2132–2151.

Zhou, J., Han, X., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2019). GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.