Human-Centered Natural Language Processing for Countering Misinformation

by

Ashkan Kazemi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2024

Doctoral Committee:

Professor Rada Mihalcea, Co-Chair
Assistant Research Scientist Verónica Pérez-Rosas, Co-Chair
Associate Professor Ceren Budak
Associate Professor Scott A. Hale, University of Oxford
Associate Professor Lu Wang

Ashkan Kazemi

ashkank@umich.edu

ORCID iD: 0000-0002-2475-1007

Dedication

I dedicate this thesis to my parents Akram Zareeyan and Abolfazl Kazemi, and my brother Amir Mohammad Kazemi, whose love, support, and sacrifices have provided me the opportunity to grow and reach my ambitious dreams of pursuing a doctoral degree at the University of Michigan.

Acknowledgments

Thank you to my parents who dedicated their life to raising me, and made sacrifices so that I can have the privilege of pursuing my dreams of getting a PhD at the University of Michigan. My mother's deep passion for learning taught me to be curious from a young age, I learned from my father to persevere through difficult times, and my brother was my supportive best friend growing up. Your loving support throughout my life and especially these past several years gave me the strength to power through a pandemic alone and cope with not seeing you in four and a half years. I am eternally grateful for everything you have given me.

Thank you to my advisors Rada and Veronica. You gave me an opportunity to be part of the LIT lab and supported me through difficult times in my PhD journey. Thank you for all your guidance and feedback during this time, and having an open mind and being flexible to exploring my curious research ideas. I have learned many invaluable lessons from working together and I hope I can pay them forward.

Thank you to my committee members: doctors Rada Mihalcea, Veronica Perez-Rosas, Scott Hale, Ceren Budak, and Lu Wang; your insightful feedback on my research added depth and quality to this dissertation.

I am grateful for all the mentorship I have received throughout my career and education. Thank you to Scott Hale for your kind and valuable mentorship throughout my PhD. My experience during the research internships I completed at Meedan would not have been the same without your help and supervision. Your guidance gave me confidence that it is possible to make impactful and equitable contributions using natural language processing to global issues such as misinformation.

Thank you to my undergraduate professors who educated me and provided support for my ambitious academic goals: doctors Fatemeh Ghassemi (my undergraduate thesis advisor who gave me the opportunity to teach software engineering and compilers as her teaching assistant), Ramtin Khosravi (my undergraduate professor who taught me all I know about software engineering and provided me with teaching opportunities at his software engineering lab), Azadeh Shakery (who made me passionate about working with and teaching data storage and processing), Hamid Mahini (my manager at TAPSI and my first mentor in machine learning research), Jalil Rashed Mohassel (a University of Michigan alumni whose deep knowlede and skillful teaching gave me confidence in advanced math and electrical circuits and inspired me to pursue a PhD at the Ann Arbor campus), and Bashir Sadjad (my algorithms professor who empowered us into believing that we can all be great computer scientists one day.)

Thank you to my partner Stella Han who lovingly supported me to finish working on my dissertation. Thank you to my good friends Nima Mirzaei, Omid Oliyan, Sarah Emeritz, Ehsan Oliyan, Beheshteh Makari, Mohsen Heidari Khouzani, Alican Büyükçakır, Cristian Paul Bara, Caleb Belth, Do June Min, and Santiago Castro. You brought joy and happiness to the cold and dark winters of Ann Arbor.

I would like to extend a big thank you to my fellow LIT lab members and co-authors, it was a great experience working together and learning from you: Laura Biester, Steve Wilson, Santiago Castro, Oana Ignat, Do June Min, Siqi Shen, Kiran Garimella, Devin Gaffney, Zehua Li, Qinyue Tan, Zhijing Jin, Naihao Deng, and Artem Abzaliev.

Last but not least, thank you to those who funded this thesis: The John Templeton Foundation, the Robert Wood Johnson Foundation, the Galler Fellowship at the University of Michigan, and Meta Platforms Inc. My dissertation would not have been possible without your support.

I would also like to acknowledge that this doctoral thesis was conducted entirely on the traditional lands of the Anishinabae people of the Three Fires Confederacy – the Ojibwe, Ottawa & Potawatomi– and the Wyandot, and pay my respect to elders both past and present.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

As curbing the spread of online misinformation has proven to be challenging, we look to artificial intelligence (AI) and natural language technology for helping individuals and society counter and limit it. Despite current advances, state-of-the-art natural language processing (NLP) and AI still struggle to automatically identify and understand misinformation. Humans exposed to harmful content may experience lasting negative consequences in real life, and it is often difficult to change one's mind once they form wrong beliefs. Addressing these interwoven technical and social challenges requires research and understanding into the core mechanisms that drive the phenomena of misinformation.

This thesis introduces human-centered NLP tasks and methods that can help prioritize human welfare in countering misinformation. We present findings on the differences in how people of different backgrounds perceive misinformation, and how misinformation unfolds in different conditions such as end-to-end encrypted social media in India. We build on this understanding to create models and datasets for identifying misinformation at scale that put humans in the decision making seat, through claim matching, matching claims with fact-check reports, and query rewriting that scale the efforts of fact-checkers. Our work highlights the global impact of misinformation, and contributes to advancing the equitability of available language technologies through models and datasets in a variety of high and low resources and languages.

We also make fundamental contributions to data, algorithms, and models through: multilingual and low-resource embeddings and retrieval for better claim matching, reinforcement learning for reformulating queries for better misinformation discovery, unsupervised and graph-based focused content extraction through introducing the Biased TextRank algorithm, and explanation generation through extractive (Biased TextRank) and abstractive (GPT-2) summarization.

Through this thesis, we aim to promote individual and social wellbeing by creating language technologies built on a deeper understanding of misinformation, and provide tools to help journalists as well as internet users to identify and navigate around it.

CHAPTER 1

Introduction

Online misinformation is a complex multifaceted phenomena, as it involves various digital interactions among humans and their computers such as engaging with highly personalized social media. Humans struggle to identify whether a post contains misleading information, and even when misinformation is successfully taken down, the damages to exposed users could have already triggered negative behavior change. Social media– often feed and ad-based, has mechanisms that enable cascading and complex changes to society [1].

It is helpful to think of misinformation as "information pollution." Pollution is the introduction of harmful materials into the environment, either naturally such as volcanic ash or introduced by human activity such as runoff produced by factories[1]. Pollutants damage our air quality, water and the environment at large. We can apply the famous "duck test[2]" to pollution and misinformation: if misinformation looks like pollution and acts like pollution, then it is *information* pollution. Similar to toxic water pollutants that poison city drinking waters, the spread of information pollution into our online social lives causes damages to individuals and societies that are sometimes beyond repair. For instance, online misinformation has ignited catastrophic social distress in recent years such as a genocide in Myanmar [2] and worsening global public health during the COVID-19 pandemic [3].

Similar to other negative externalities of social media, misinformation's complexity rises from the continuous interactions among humans and computers. Social media algorithms constantly change and adapt to a user's personalized content recommendations learned from historical user engagement data. What is recommended to users may end up shaping user interests rather than being based on them. This is because the foremost priority of ad-based social media is to increase user engagement– measured in minutes, not user welfare. Such complex cycles of engagement at population scale have the potential to change humanity [1] and are often difficult to reverse [4, 5].

[1] https://education.nationalgeographic.org/resource/pollution/

[2] https://en.wikipedia.org/wiki/Duck_test

1.1 The Social and Economic Dimensions of Misinformation

Misinformation has a real human cost. Take the hypothetical example of "Alice" who has recently engaged with COVID-19 misinformation, and who might see similar misleading or extremist content in her feed since recommender systems infer that suggesting such content to Alice will likely increase her time on their platform based on regressing from her past activity. "Bob" who knows Alice from college, reads Alice's COVID-19 conspiracy theories on his social media, and in light of Bob's past experiences and biases and trust for Alice, the misleading content resonates deeply enough for him to repost, further cascading misinformation onto his like-minded online friends in a polluted information cycle. What continues to be "recommended" to Alice and Bob through their social media feeds- which is not necessarily what they would be interested in, is controlled by a recommender system which has access to massive computation and curated information about Alice, Bob and many others similar to them and can make inferences about users' behaviors and interests based on similar users. Thus, it is necessary that we address misinformation with a human-centered approach that prioritizes user welfare and harm reduction. To that end, we first have to identify the individuals and groups involved, and gain a deeper understanding of the social and economic aspects of misinformation, as well its adverse impacts on public health.

1.1.1 The Stakeholders of Misinformation

A human-centered approach to addressing misinformation requires a detailed understanding of the actors involved and their incentives. We identify the following individuals and collectives to be central stakeholders in the online misinformation ecosystem:

- **Online media.** The owners of internet platforms stand to profit more from misinformation and controversy than from regular content. [6] They also experience political backlash because of online misinformation, which in effect imposes public relations related costs on platforms. Since online media control the apparatus of information circulation, they are the most influential of all stakeholders.

- **Advertisers.** The vast majority of revenue of online media is through the sale of user attention to advertisers. This gives advertisers noticeable power over the platforms, and therefore they can exert force on internet companies using their spending as leverage. In the United States' traditional broadcasting and news media, the advertisers' lobbying power has led the Federal Communications Commission (FCC)'s to regulate "obscenity, indecency, and profanity" [7], and fine media companies for publishing such content. Advertisers may choose to follow similar paths in response to online misinformation tainting their

brands. According to Media Matters' report [8], Twitter lost half of its top 100 advertisers, who purchased nearly $2 billion worth of ads since 2020, only a month after Elon Musk acquired the company. The advertisers included large corporations such as American Express, Citigroup, Chipotle, Nestle, Black Rock, and Chanel, and some publicly cited controversies and concerns around looser content moderation post Musk take over.

- **Journalists and fact-checkers.** At the forefront of reporting on world events, journalists and fact-checking organizations often have to spend extra time and energy to go against misinformation, by avoiding them in reporting and doing extra work debunking falsehoods. Journalism has become a more challenging and risky profession in recent years in part as a side effect of excessive information pollution. It is worth noting that the rivalry among traditional news media and online media for user attention is a confounding factor that sometimes interferes with journalism's impartiality in promoting the best course of action against misinformation. Nevertheless, journalists and fact-checkers remain our best source of professional and expert advocacy against misinformation, since they have an economic interest in protecting their work from falsehoods.

- **Civil society.** Misinformation can have far-reaching and harmful effects on civil society. When people are exposed to false or misleading information, they may form incorrect opinions and beliefs that can harm the social fabric of a community. For example, misinformation can create divisions among people by fueling prejudice, mistrust, and hate. It can also erode public trust in institutions, such as the government, media, and scientific community, which are critical for maintaining a healthy democracy. Misinformation can also lead to the spread of harmful practices and ideologies. For example, false information about vaccine safety can discourage people from getting vaccinated, leading to the spread of preventable diseases. Similarly, misinformation about climate change can undermine efforts to address this pressing global issue. In addition, misinformation can also have serious consequences for individual and collective decision-making. People who rely on false information to make decisions may end up taking actions that are not in their best interest or the interest of society as a whole. Therefore civil society and grassroots organizations are an important lobby against misinformation.

While at first glance it seems that all parties must want to combat misinformation, they collectively have struggled to do so over the years. Social media owners prioritize short-term profits over the potential long-term risks of misinformation to their business. The architecture of online advertisement affords the advertisers to turn a blind eye and collect profits, since ads are targeted towards demographics and are not attached to content. The former two stakeholders possess the

power, but lack the will to make structural changes to control misinformation. The other stakeholders, mainly civil society and journalists are in the reverse position, as they do not possess as much control over the flow of information, but have demonstrated interest and will to counter misinformation.

1.1.2 Social Cost of Misinformation

A growing literature in economic and social research has been shedding light on the costs incurred by misinformation on society. In a 2019 report [9], a group of economists and cybersecurity analysts placed a $78 billion price tag on the global damages of fake news, citing the most affected sectors as *stock market losses and volatility* ($39 billion), *financial misinformation* ($17 billion in US alone), and *reputation management* and *public health* (US only) costing an annual $9 billion each, all in a single year. According to research from the Economic Policy Institute published in 2017, retirement savers lose an annual $17 billion from acting on misleading advice from financial advisors with conflicts of interest. [10, 11]

Additionally, activists and non-profit organizations have mobilized in recent years to study the finances driving online misinformation, and at times have successfully demonetized the interests behind pushing misleading narratives. Such efforts include "Sleeping Giants," an activist organization comprised of mostly anonymous members with active chapters in the US, Australia, Brazil, Canada, France, and Germany who since their inception in late 2016, have successfully demonetized extremist and fake news websites as famous as Breitbart News, causing 820 corporations including AT&T, BMW, and Visa to stop advertising with the far right outlet; [12] and Global Disinformation Index (GDI),[3] a not-for-profit organization that publishes open research on news markets around the world by providing dynamic exclusion lists of global news organizations rated high risk for misinformation to adtech companies, effectively providing a mechanism for systematically defunding misinformation.

In late 2022 the families of the victims of the Sandy Hook elementary school massacre which occurred ten years prior, won two defamation cases against Alex Jones, the infamous conspiracy theorist who circulated baseless lies about the victims being hired actors by the government in a conspiracy to take away Americans' guns. Jones has been ordered to pay $1.49 billion in damages in two Sandy Hook defamation cases, and awaits a third trial pending investigation in Texas. [13] The ruling is a first of its kind in the United States, setting precedent in punishment for defamation through deploying misinformation.

[3] `https://www.disinformationindex.org`

1.1.3 Risks and Challenges of Misinformation for Public Health

Misinformation can have serious public health implications, as it can spread false or misleading information about health issues, treatments, and interventions. Some of the key public health implications of misinformation include:

- **Discouraging vaccination.** False information about vaccine safety can discourage people from getting vaccinated, leading to the spread of preventable diseases. This can have serious consequences for public health, especially in the context of outbreaks and pandemics. There is a growing body of research studying the impact of misinformation on vaccination and public health. [14, 15, 16, 17]
- **Promoting risky treatments.** Misinformation about health treatments can lead people to seek out dangerous or ineffective remedies, which can be harmful to their health. For example, false information about the dangers of conventional medical treatments can lead people to rely on unproven alternative therapies. [18, 19, 20]
- **Undermining public trust in science.** Misinformation about health issues can erode public trust in science and scientific institutions.[21, 22, 23] This can make it difficult for public health authorities to effectively communicate important health information and promote evidence-based practices. [24, 25]
- **Delaying treatment.** False information about symptoms and treatments can lead people to delay seeking medical help, which can have serious consequences for their health. For example, false information about the causes of cancer can discourage people from seeking early detection and treatment, which can reduce the chances of successful treatment.

Such implications even at small scales pose huge risks to community and public health as social changes such as anti-vaccination movements can cause exponentially worse public health outcomes, leading the US surgeon general to declare misinformation as a public health emergency. [24] Many parts of the healthcare industry including doctors, nurses, and medical staff are impacted by misinformation in a variety of ways such as:

- **Healthcare resource allocation.** Misinformation can lead to an overuse or misuse of healthcare resources. For example, people may seek unnecessary medical treatments or tests based on false information, which can strain healthcare systems and divert resources away from those who need it most.
- **Health system costs.** The unnecessary overuse of healthcare resources caused by exposure to information pollution can increase healthcare costs, which can negatively impact patients, hospitals, and governments.

- **Burnout of hospital staff.** Hospital staff, including doctors, nurses, and support staff, may experience stress and burnout as they work to manage the consequences of misinformation. For example, they may have to spend extra time educating patients about accurate health information or addressing the fallout from misinformation-driven health decisions. [25]

1.2 Human-Centered NLP to Counter Misinformation

In recent years amid the high volume of information pollution on the internet, we have observed a trend of NLP and AI tools aiming to counter misinformation. The solution to fake news may seem trivial to computational linguists: build a predictive model that can classify news into fake and real, and use that to filter the information flow. Seminal papers in recent years have proposed tasks such as fake news identification [26, 27, 28], deepfake [29] and neural fake news detection [30], and automated fact-checking [31] to similar ends.

However, as the interdisciplinary community studying misinformation grew, NLP researchers learned what fact-checkers and journalists want are often NLP and AI tools to augment and complement their procedures, and not fully automatic fake news detection [32]. Social media platforms who also aspire to automate away monotonous and often expensive content moderation tasks, use AI to increase fact-checker efficiency, and use signals from users to automatically flag fake news. Acknowledging the stakeholders' needs leads to the development of more effective antidotes for misinformation.

My goal for this thesis is to utilize the recent advances in NLP and AI to help and protect individuals like Alice and Bob and their communities from misinformation. We aim to prioritize the needs of individuals and communities to create language technology that is human-centered. To achieve this goal, NLP can help us **identify** misinformation at scale, **understand** the intricacies of misinformation, and build tools and guardrails so users can **safely navigate** the online information landscape.

To create human-centered language technology that reliably identifies misinformation or helps users safely navigate the internet, we must first understand how misinformation affects the individuals we want to help. While most of the work in NLP has considered a "one size fits all" approach while solving user-centric tasks, such as misinformation detection, recent work has started to challenge this assumption and develop discrete [33] or continuous [34] user representations that encode information about the user background, or construct demographic-aware word representations [35, 36, 37]. Similarly, media consumption and individuals' relationship with media vary across the socioeconomic spectrum, as well as across different cultures and platforms. Taking these differences into account often makes or breaks efforts against misinformation.

There is a lack of trust in automatic fake news detection within the NLP community as automatic

fake news detection systems perform far below production-ready levels. To establish reliability and trust in such solutions, we need to find ways to keep humans in the decision making loop and rely on them for critical decisions, while leveraging AI automation at scale for detecting misinformation. Involving humans in identifying fake news is currently the most reliable way to ensure trust in its decisions. X (formerly Twitter)'s Community Notes project is an example of a practical social computing system that utilizes crowd intelligence for identifying and contextualizing misinformation. Prior research on digital juries [38] suggests that a civics-oriented digital jury for decision-making in content moderation is more procedurally just compared to existing content moderation pipelines. In this work, we create language technology that enables journalists, fact-checkers, and moderators to scale their efforts in identifying misinformation.

In addition to identifying misinformation, humans also need help navigating polluted information on the internet. For instance, non-experts consuming health and medical information on the internet can be misled or confused by technical terms or medical jargon, and therefore can benefit from simple NLP-powered interventions that provide explainers to their questions. Or journalists researching misleading claims on social media may have no luck in finding relevant information through social media search engines, and suggestions to improve their queries have the potential to expand their reach of knowledge. Given their human-centered form, these NLP-based interventions have a higher chance of success if they focus on the specific needs of individuals and communities they wish to serve.

Creating human-centered language models, resources, and algorithms for countering misinformation requires taking individual differences into account, enabling and prioritizing the needs of humans, and being equitable in doing so.

In this thesis we aim to achieve these goals through addressing the following research questions:

1.2.1 How to use language technology to gain a human-centered understanding of misinformation?

The space of possible interventions using NLP for countering misinformation is large, and finding the most effective solution for prioritizing the wellbeing of humans and their communities requires a better understanding of how we are all affected by misinformation.

1. While encrypted platforms like WhatsApp are much less studied in prior work, partly due to a lack of public data, it is worth emphasizing the impact of WhatsApp: the platform has about five fold more users than X (formerly Twitter) as of 2023[4]. In chapter 2 we look at *tiplines*, a promising direction for uncovering and contextualizing fake news in end-to-end

[4] https://en.wikipedia.org/wiki/List_of_social_platforms_with_at_least_100_million_active_users

encrypted social media such as WhatsApp. We conduct a case study of the 2019 Indian general election on WhatsApp, and understand through the lens of tiplines about misinformation on encrypted platforms and in non-conventional study settings (multilingual data, on a platform that is understudied.) We believe that studies such as ours presented in chapter 2 are not only necessary to understand the dynamics of misinformation in non-US settings, but also help us envision a more accurate view of misinformation that encompasses more diverse observations about the phenomenon and leads to more equitable outcomes.

2. In chapter 3 we study how individuals from different backgrounds are affected by misinformation. We model individuals' and groups perceptions of misinformation by fine-tuning a text-to-text pretrained language model and the Misinfo Reaction Frames dataset. Our study helps us learn more about the differences that may exist across several demographic groups such as women, the uneducated, or non-white individuals, and informs the design of more effective language technology against misinformation by taking into account those differences.

1.2.2 How to utilize language technology to identify misinformation at scale while keeping humans in the loop?

The state-of-the-art fully automated fake news detection is far from being practical, as the decisions of the systems are low-accuracy and unreliable. To increase the complexity another fold, misinformation is a global issue, and therefore to be equitable we should address it in various languages and platforms. Although research is active in the fully automatic front, based on a better understanding of what humans need to counter misinformation [32], a more achievable strategy for identifying online misinformation at scale is a human-centered approach, in which humans use AI and NLP powered systems to identify and fact-check misinformation. Such human-AI collaborations are naturally more trustworthy, as the critical decisions of the pipeline are still overseen by humans, while retaining the scalability benefits of AI and NLP.

1. In chapter 4 we enable fact-checkers to scale up their efforts by matching similar claims that leads to streamlined triage and prioritization of prevalent claims by fact-checkers. We create two datasets (one for claim detection and another for claim matching) in five high and low resource languages, and use the data to evaluate our sentence embedding model trained using knowledge distillation and multilingual parallel corpora, as well as state of the art multilingual embedding models for claim matching. Claim matching ensures fact-checkers investigate repetitive claims only once, and our work achieves this goal equitably for journalists around the world by making NLP models and data that work for both high and low resource languages.

2. In chapter 5 we focus on finding fact-checks for social media posts across different languages. We create a multilingual dataset of tweet and fact-check pairs in single and cross language settings to automatically find fact-checks for the claims made in the input post. We use these datasets to evaluate state-of-the-art multilingual models for our task across different languages as well as in cross-lingual settings. Such systems are useful for fact-checkers and internet users to spot online misinformation, and it can save expensive fact-checking labor for previously fact-checked claims, especially in cross lingual settings, e.g., a fact-check made in English could apply to similar claims it discusses in Hindi, and can expand the reach of the fact-checks beyond the original language.

3. Discovering and stopping misinformation early is among the best ways to counter misinformation. However, formulating an effective query that discovers misleading claims is not arbitrary, as the initial tip received by fact-checkers often is not the best query across different platforms. In Chapter 6 we present an adaptable offline reinforcement learning agent that transforms claims into better queries by learning to edit them using human-interpretable actions to improve misinformation discovery for arbitrary search endpoints. We train a decision transformer on replay trajectories of the FEVER [31] dataset to suggest edits (e.g. remove, or replace tokens) for turning the claim into a more effective search query. We use retrieval performance measures such as precision or recall as weak signals of reward, to compensate for the scarcity of access to social media search APIs. Our query rewriting approach can transform claims into queries that are more effective in discovering misleading claims on social media.

1.2.3 How can NLP help users safely navigate around misinformation in online environments?

As we build language technology to identify and understand misinformation better, users still interact with online falsehoods. So it is important to empower users with NLP tools that provide explanations and context around potentially misleading content on the internet, and help them navigate the online information landscape safely. For instance, non-expert readers following health and medical news might find themselves in need of more explanations about certain claims in the articles. Communities and civil society wanting to develop policy to limit online misinformation can also benefit from the measurements and transparency that NLP can provide them.

1. Chapter 7 discusses algorithms for contextualizing news on the internet for users by generating extractive and abstractive explanations of news articles or long fact-check reports. We investigate the effectiveness of Biased TextRank (extractive) and a fine-tuned GPT-2 model

(abstractive) on LIAR-PLUS and the novel Health News Reviews datasets. These explanations can provide context to internet users consuming potentially misleading information and protect them from harm.

2. It is important to empower internet users to protect themselves from online misinformation. Prior work on digital juries [38] suggests that a civics-oriented digital jury for decision-making in content moderation is more procedurally just compared to existing content moderation pipelines. In WhatsApp and other end-to-end encrypted social media, centralized content moderation by the platforms is not possible as the encrypted content is only available to the users involved in the conversation. We study the effectiveness of a potential opt-in solution called a "tipline" in Chapter 2 through the case study of the 2019 Indian general elections on WhatsApp. We apply multilingual embeddings as well as image embeddings on the texts and images shared on public WhatsApp groups during the election months to analyze the effectiveness of tiplines in uncovering misinformation for users who have opted in to use this service. Our work in this chapter enables users to seek help from journalists and fact-checkers in identifying misleading content on end-to-end encrypted platforms.

1.3 Thesis Organization

The thesis is organized as follows: In chapter 2, we discuss how a crowd-sourced "tipline" can uncover viral misinformation on end-to-end encrypted social media (WhatsApp, Signal, Telegram, and so on) in a timely manner by using image and multilingual sentence embeddings. In chapter 3, we dig deeper into the question of how users' background and demographics may effect their perception of misinformation, and how that effect varies across demographic groups. These chapters broadly address RQ1 (using NLP to understand misinformation.)
Chapters 4 through 3 address the third research question: we introduce claim matching in Chapter 4; the task of grouping similar claims that can be fact-checked together, in high and low resource languages and train a multilingual sentence embedding model using knowledge distillation that improves over existing state-of-the-art multilingual embedding models. Chapter 5 discusses finding existing applicable fact-checks for social media posts in single-language and cross-language settings in English, Portuguese, Spanish and Hindi. Along with the previous two, chapter 6 addresses RQ2 by leveraging reinforcement learning to help fact-checkers discover misleading claims by rewriting their initial claim into an effective query for arbitrary social media search endpoints. These three chapters help scale efforts in identifying and fact-checking online misinformation.
Chapter 7 addresses the problem of providing context for claims under investigation by using

abstractive (GPT-2) and extractive (Biased TextRank) summarization methods, and addresses the third research question alongside chapter 2 in empowering users to navigate around misinformation in their online interactions. We conclude the thesis in Chapter 8 by revisiting the research questions introduced in this Chapter (1) and highlight the contributions of the thesis in creating human-centered language technology for countering misinformation.

CHAPTER 2

Uncovering Misinformation on End-to-End Encrypted Social Media

There is currently no easy way to discover potentially problematic content on WhatsApp and other end-to-end encrypted platforms at scale. In this chapter, we analyze the usefulness of a crowd-sourced tipline through which users can submit content ("tips") that they want fact-checked. We compared the tips sent to a WhatsApp tipline run during the 2019 Indian general election with the messages circulating in large, public groups on WhatsApp and other social media platforms during the same period. We found that tiplines are a very useful lens into WhatsApp conversations: a significant fraction of messages and images sent to the tipline match with the content being shared on public WhatsApp groups and other social media. Our analysis also shows that tiplines cover the most popular content well, and a majority of such content is often shared to the tipline before appearing in large, public WhatsApp groups. Overall, our findings suggest tiplines can be an effective source for discovering potentially misleading content.

2.1 Introduction

Platforms such as WhatsApp that offer end-to-end encrypted messaging face challenges in applying existing content moderation methodologies. End-to-end encryption does not allow the platform owner to view content. Rather, only the sender and recipients have access to the content—unless it is flagged by a receiving user [39]. Even though WhatsApp is extremely popular, used by over 2 billion users all over the world, there is currently no large-scale way to understand and debunk misinformation spreading on the platform. Given the real-life consequences of misinformation [40] and the increasing number of end-to-end encrypted platforms, developing tools to understand and uncover misinformation on these platforms is a pressing concern.

One potential solution is to make use of misinformation "tiplines" to identify potentially misleading or otherwise problematic content [41]. A tipline is a dedicated service to which "tips" can be submitted by users. On WhatsApp, a tipline would be a phone number to which WhatsApp users can forward potential misinformation they see in order to have it fact-checked. We call the messages sent by users "tips."

Table 2.1: Examples of English text messages forwarded to the WhatsApp tipline to be fact-checked. Please note that grammar and spelling errors are in the originals. The content we analyzed includes messages in multiple languages and formats (e.g., text, images, and links).

UNESCO Declare India's "Jana Gana Mana" the World's Best National Anthem
When you reach poling booth and find that your name is not in voter list, just show your Aadhar card or voter ID and ask for "challenge vote" under section 49A and cast your vote. If you find that someone has already cast your vote, then ask for "tender vote" and cast your vote. If any polling booth records more than 14% tender votes, repolling will be conducted in such poling booth. Please share this very important message with maximum groups and friends as everyone should aware of their right to vote.
Happened today on 47 street (Diamond Market) New York $100,000 given away in ref to Modi victory .. see how this millionaire Indian is doing ..
Coal India is on the verge of ruin! 85,000 crore loss due to Modi!

In this chapter, we address two main research questions:

- How effective are tiplines for identifying potentially misleading content on encrypted social media platforms?
- What content is submitted to tiplines for fact-checking?

and make the following contributions:

- Using state-of-the-art text and image matching techniques, we compared content sent to the tipline to the content collected from a large-scale crawl of public WhatsApp groups (these are WhatsApp groups where the link to join is shared openly), ShareChat (a popular image sharing platform in India similar to Instagram), and fact checks published during the same time in order to understand the overlap between these sources.
- The tipline covers a significant portion of popular content: 67% of images and 23% of text messages shared more than 100 times in public WhatsApp groups appeared on the tipline.
- We found that a majority of the viral content spreading on WhatsApp public groups and on ShareChat was shared on the WhatsApp tipline before appearing in the public groups or on ShareChat.
- Compared to content by popular fact-checking organizations, the messages from tiplines cover a much higher proportion of WhatsApp public group messages. We suspect this is because fact-checking organizations typically fact-check content primarily based on

signals from open social media platforms like Facebook and Twitter, whereas the tipline is a crowdsourced collection of content native to WhatsApp.

2.2 Related Work

While this chapter is, to the best of our knowledge, the first peer-reviewed study on WhatsApp tiplines, tiplines are quite common in practice. WhatsApp, for instance, currently lists 54 fact-checking organizations with accounts on its platform[1]. Other efforts include the Comprova project[2] and FactsFirstPH[3], an initiative of over 100 organizations uniting around the 2022 Philippine presidential election. Tiplines are similar to features on platforms such as Twitter and Facebook that allow users to flag potential misinformation for review, but tiplines are operated by third parties and can provide instantaneous results for already fact-checked claims [42].
In this study, we used data from a WhatsApp tipline that ran during the 2019 Indian general election as part of the Checkpoint project[4]. Checkpoint was a research project led by PROTO[5] and Pop-Up Newsroom, technically assisted by WhatsApp[6]. The goal of this project was to study the misinformation phenomenon at scale—natively in WhatsApp—during the Indian general election. The tipline was advertised in the national and international press during the election[7]. There was an advertising campaign on Facebook, but no specific call to action was present in WhatsApp itself. Table 2.1 presents some examples of text messages submitted to the tipline. The goal of this article is to understand what content is submitted, analyze how effective tiplines can be for discovering content to fact-check, and shed light on the otherwise black-box nature of content spreading on WhatsApp.

2.3 Data

We used a wide range of data sources in this work including WhatsApp tipline data, social media data from WhatsApp public groups and ShareChat, and published fact checks. All the data used pertains to the four-month period between March 1, 2019, and June 30, 2019. This period includes the 2019 Indian general election, which took place over a period of six weeks in April and May 2019.

[1] IFCN Fact Checking Organizations on WhatsApp

[2] Comprova project website

[3] FactsFirstPH website

[4] Checkpoint project website

[5] PROTO is an Indian organization that describes itself as, "a social enterprise that is trying to achieve better outcomes in civic media through collaboration and research". website

[6] Pop-Up Newsroom is a joint project of Meedan and Fathm that designs and leads global election and event monitoring journalism efforts.

[7] Announcement article for WhatsApp tipline ahead of the India elections.

Tiplines. In 2019, PROTO led the Checkpoint project using Meedan's open-source software to operate a WhatsApp tipline. PROTO advertised their WhatsApp number asking users to forward any potentially misleading content related to the election. They advised that they would be able to check and reply to some of the content that they received. Over the course of four months, 157,995 messages were received. Of these, 82,676 were unique and consisted of 37,823 text messages, 10,198 links, and 34,655 images. We obtained a list of links, text messages, and images along with the timestamps of when they were submitted to the tipline. We have no information about the submitting users beyond anonymous ids.

WhatsApp public groups. There are currently over 400 million active WhatsApp users in India. With the availability of cheap Internet data and smartphones with WhatsApp pre-installed, the app has become ubiquitous. Aside from messaging friends and family, Indians use WhatsApp to participate in political discourse [43]. Political parties have taken this opportunity to create thousands of public groups to promote their political agendas. These groups have been shown to be quite prevalent, with over one in six Indian WhatsApp users belonging to at least one such group [44].
In addition to the image and text items submitted to the tipline, we have data from large "public" WhatsApp groups collected by Garimella and Eckles [45] during the same time period as the tipline ran. The dataset was collected by monitoring over 5,000 public WhatsApp groups discussing politics in India. For more information on the dataset, please refer to Garimella and Eckles [45].

Sharechat. ShareChat is an Indian social network that is used by over 100 million users.11 It has features similar to Instagram and is primarily multimedia focused [46]. Unlike WhatsApp, ShareChat provides global popularity metrics including likes and share count, which allowed us to construct a proxy for the popularity of the content on social media. ShareChat curates popular hashtags based on topics such as politics, entertainment, sports, etc. During the three months of data collection, every day, we obtained the popular hashtags related to politics and obtained all the posts containing those hashtags. This provides a large sample of images related to politics that were posted on ShareChat during the data collection period (March 1 to June 30, 2019).

Fact checks. We also collected fact checks and social media data from the time period in English and Hindi. We crawled popular fact-checking websites in India and obtained articles and any tweets linked within the articles following the approach of [47] and [48]. Overall, we found 18,174 fact-check articles in 49 languages from 136 fact checkers from all over the globe. To select fact checks concerning the Indian general election, we filtered the data to require that either the fact check be written in an Indian language or the fact-checking domain be within India's

Table 2.2: Datasets used in this work. The values shown in parentheses indicate the number of unique messages/images. We only collected image data from ShareChat.

Datasets	**#Text messages (unique)**	**#Images (unique)**
Public groups	668,829 (445,767)	1.3M (977K)
ShareChat	-	1.2M (401K)
Checkpoint	88,662 (37,823)	48,978 (34,655)
Fact-check articles	5,444 (5,444)	-
Fact-check tweets	811 (245)	-

country code top-level domain.
In total, we obtained 3,224 and 2,220 fact checks in English and Hindi respectively. The fact checks were of content from various social media platforms, including Twitter. Whenever available, we obtained the links to the original tweets that were fact-checked and downloaded these. We obtained 811 tweets in total, 653 (182 unique) in English and 158 (63 unique) in Hindi.
A summary of all the data collected is shown in Table 2.2.

2.4 Methods

Image similarity. To identify similar images, we used Facebook's PDQ hashing algorithm and Hamming distance. PDQ is a perceptual hashing algorithm that produces a 64-bit binary hash for any image. Small changes to images result in only small changes to the hashes and thus allow visually similar images to be grouped. This allows, for instance, the same image saved in different file formats to be identified. Images with a Hamming distance of less than 31 were considered to be similar. The same threshold was used previously by [49]. Similar images were clustered together using the DBSCAN [50] algorithm.
To construct the visual summary of the images shown in Figure 8, we first obtained a 1,000-dimensional embedding for each image using a pretrained ResNeXt model [51]. Next, we clustered these embeddings using a k-means clustering algorithm and chose k = 20 using the elbow method. For each cluster, we picked four randomly sampled images and created a mosaic of the 20 clusters.

Text similarity. To identify similar textual items, we used a multilingual sentence embedding model trained for English, Hindi, Bengali, Marathi, Malayalam, Tamil, and Telugu [42]. [42] evaluated this model for claim matching using similar data and found applying a cosine similarity threshold of 0.9 to pairs of messages resulted in the best performance, with an overall F1 score of 0.73. The model performs better on English and Hindi (which are 82% of our data), with an

average F1 score of 0.85. Throughout this chapter we used a cosine similarity threshold of 0.9 for matching text items.

Text clustering. We custered text items using online, single-link hierarchical clustering. Each new message arriving to the tipline was compared to all previous messages, and the best match found. If this match was above the similarity threshold, then we added the new message to the same cluster as the existing message. We applied the same process to the public group messages. To enable quick retrieval, we constructed a FAISS [52] index using our Indian XLM-R embeddings of all the public group messages. We then queried this index for each tipline message and recorded all matches with a cosine similarity score of at least 0.9. We remove any duplicate matches (i.e., cases where two tipline messages matched the same public group message) before analyzing the matches.

2.5 Results

2.5.1 Tiplines capture content quickly; popular content often appears in tiplines before appearing in public groups.

We examined the effectiveness of tiplines in three ways: speed (i.e., how quickly new content appears in tiplines), overlap with content in public groups, and volume. We began with speed and first examined how long it took for an item to be shared by someone to the tipline. The intuition behind this is that one facet of an effective solution is its ability to identify potential misleading content quickly before it spreads widely.

Figure 2.1 shows the time difference between an image being shared on a public group and the tipline. Negative values on the x-axis indicate that the content was shared in a public group first. We see that roughly 50% of all the content was shared in public groups first, with around 10% of content going back to over a month. However, if we focus on the subset of the top-10% most shared images within the public groups, the distribution looks very different. We clearly see that a majority of the content (around 80%) was shared on the tipline before being shared in the public groups, indicating that the tipline does a good job covering the most-shared content quickly. Similar trends exist for images on ShareChat (Figure 2.2). In fact, images sent to the tipline have significantly more shares (41 vs. 29) and likes (51 vs. 40) on ShareChat compared to images not sent to the tipline ($p ¡ 0.01$ for a t-test of means).

Comparing the text messages within the public groups to the tipline messages leads to similar results (Figure 2.3). To make this comparison, we first clustered all text messages in the public groups and, separately, in the tipline. This comparison only uses the text messages from the tipline within clusters having at least five unique messages that were annotated as having claims

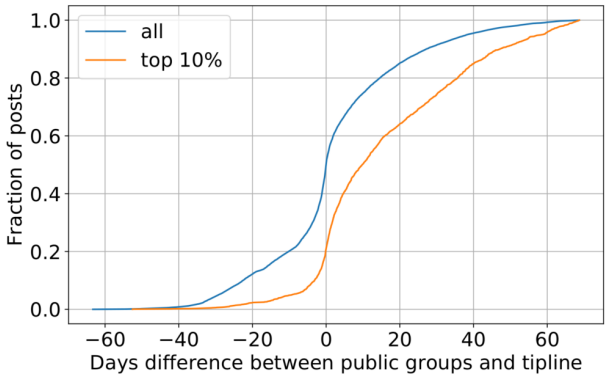

Figure 2.1: Time difference between the sharing of images on public groups and the tipline. Approximately 50% of the images were shared on public groups first. However, if we consider just the top 10% most shared images in the public groups, they were mostly shared first on the tipline. (Negative values on the x-axis represent items being shared in the public groups before being shared on the tipline.)

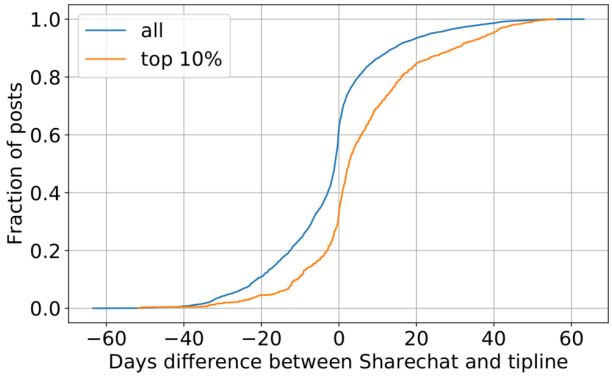

Figure 2.2: Time difference for images shared on sharechat and the tipline. The most popular content was more likely to be shared on the tipline first compared to all content.

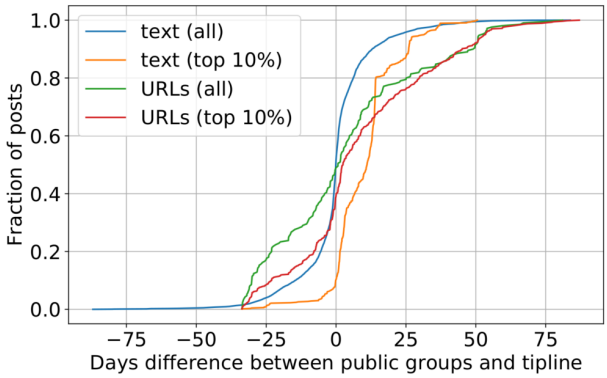

Figure 2.3: Time difference between the sharing of text messages and urls in the whatsapp tipline and public groups.

that could be fact-checked to avoid the risk of matching spam or less meaningful content. We again find that the most shared content was often shared to the tipline before spreading widely within the public groups. Similar trends also exist for URLs (Figure 2.3, green and red lines). These findings suggest that content submitted to the tipline may have been circulating person-to-person or in smaller, private groups not in our data before the content was submitted to the tipline or appeared in the large, public groups in our data. Popular content on non-encrypted social media platforms often spreads quickly through large broadcast events [53, 54]; such broadcast events may be rarer on WhatsApp, however, due to the limits on message forwarding and the size of groups.7

2.5.2 Tiplines capture a meaningful percentage of content shared in public groups.

A second facet of effectiveness is content overlap: for tiplines to be an effective source of content for fact-checking, we would want them to identify content spreading in other sources of data, including WhatsApp public groups, fact checks, and open social media platforms. We first examined the coverage and computed the number of shares for images in the public groups or on ShareChat and computed what percentage of the images with different numbers of shares appear in the tipline dataset. Figures 2.4 & 2.5 show the results. For both the public groups and ShareChat, we used logarithmic bucketing of the number of shares of items to estimate message popularity. The results show tiplines have good coverage of popular content: 67% of the images shared more than 100 times in the public groups were also submitted to the tipline. We repeated the analysis with text messages and found that 23% of text messages shared more than 100 times in the public groups were also submitted to the tipline (Figure 2.6). To put matters into perspective, we conducted a similar experiment matching all the fact-checked text claims and their corresponding social media posts from the same time period against WhatsApp public groups messages. Only 10% (12/119) of textual content from popular clusters in public groups (shared more than 100 times) matched with at least one text (claim or fact-checked tweet) from Indian fact-checks during this period.

Exact copies of about 10% of popular URLs (i.e., URLs shared over 1,000 times) on public groups were also submitted to the tipline. Because of shortened URLs, content takedowns, and the 2-year time difference between data collection and analysis, grouping URLs was very challenging. We therefore limited further analysis of URLs for this research question.

We found many text messages and images submitted to the tipline did not appear in the public groups, which suggests tiplines also capture content being distributed in WhatsApp in smaller-group or person-to-person settings. Out of the 23,597 unique clusters of images submitted to the tipline, only 5,811 clusters (25%) had at least one match with an image from the public groups. Next, we checked which text messages from the clusters with claims matched messages found in

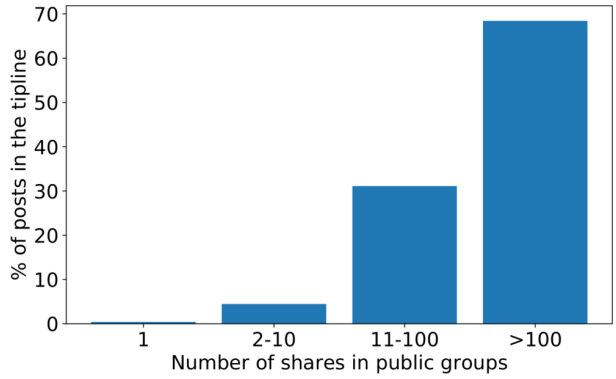

Figure 2.4: Coverage of Images: The x-axis shows the number of shares on the public groups and y-axis shows the percentage of images with x shares that match with an image submitted to the tipline. Images that are highly shared on the public groups are much more likely to be also shared to the tipline.

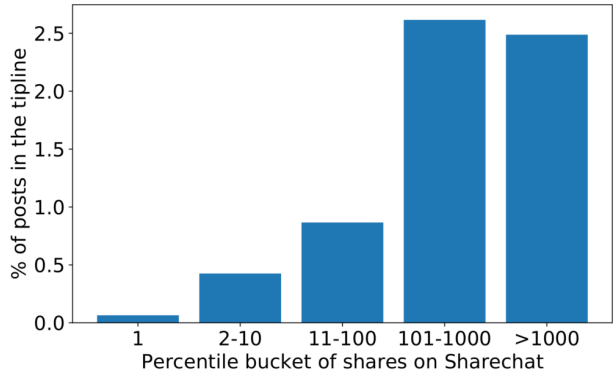

Figure 2.5: Coverage: Similar to Figure 2.4, images shared more often on ShareChat are more likely to appear in the tipline.

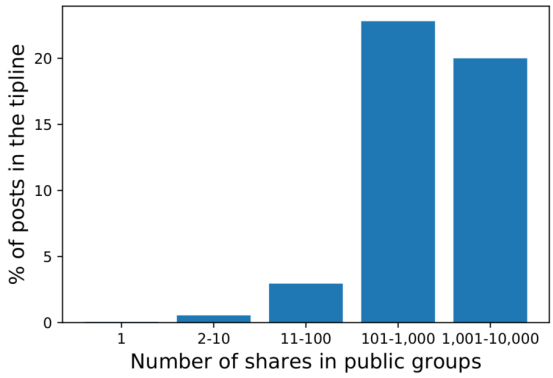

Figure 2.6: Coverage of text message: The x-axis shows the number of shares on the public groups and y-axis shows the percentage of text message with x shares that match with a text message submitted to the tipline. Text messages that are highly shared on the public groups are much more likely to be also shared to the tipline. Messages in the public groups are first clustered together to determine the number of shares of each message.

the public group data. We found that 93% of the 257 relevant clusters match at least one message in the WhatsApp public group dataset. Far from being a skewed result where only few large clusters match, we found a large number of messages across clusters of all sizes match at least one public group message. The per-cluster average of tipline messages matching to the public group data is 91%. This suggests that if we had included clusters with fewer than five unique messages, we may have seen additional matches. We did not include these, as we only wanted to include messages we knew had fact-checkable claims (and we only annotated clusters with at least five unique messages). Additional annotation would likely yield more relevant messages and matches.

Seven percent of the text clusters with fact-checkable claims from the tipline did not match any public group messages. This implies that collecting messages from public groups and using tiplines can be complementary even though neither is a full sample of what is circulating on WhatsApp.

Finally, we measured the potential impact tiplines could have on preventing the spread of misinformation. For this, we looked at items that were shared on both the tipline and in the public groups. We identified the timestamp when an item was first shared on the tipline and counted the number of shares of the item on the public groups before and after this timestamp. The intuition here is that if an item was shared on the tipline, it is in the pipeline to be fact-checked. We found that 38.9% of the image shares and 32% of the text message shares in public groups were after the items were submitted to the tipline.

2.5.3 Tiplines capture diverse content, and a large percentage of this content contains claims that can be fact-checked.

To investigate the third research question, we took an in-depth look into images, text messages, and links sent to the tipline, and here we present examples of the most popular submissions.

2.5.3.1 Images

The tipline received 34,655 unique images, which clustered into 23,597 groups. Figure 2.7 shows the three most submitted images to the tipline. Each of these three images was submitted by at least 60 unique users. All three of these images were fact-checked and found to be false. Figure 7a shows a 'leaked' government circular alleging a terrorist plot during the elections. This was in fact an old circular taken out of context. Figure 7b falsely alleges that Pakistani flags were raised during a political rally, and Figure 7c shows doctored screenshots of a TV news program.

We constructed a visual summary of all the unique images sent to the tipline, as shown in Figure 2.8. The mosaic shows various categories of images sent to the tipline at a high level. As we

(a) (b) (c)

Figure 2.7: Most shared images on the tipline.

move from the top left to the bottom right, we can see a lot of images on the top left of Figure 8 containing pictures of newspapers, and in general images with text. As we go to the bottom left, we see memes and pictures containing quotes of politicians, and on the bottom right, images of people/politicians. Pictures of newspapers or images with text on them are the most dominant type, constituting over 40% of the content, followed by memes which make up roughly 35% of the content.

2.5.3.2 Text Messages

Of the 88,662 text messages sent to the tipline, 37,823 are unique (not exact duplicates). We further organized the messages by clustering them using the Indian XLM-R model [42] and a threshold of 0.9, which resulted in 20,856 clusters (or groups) of near duplicate messages. Each cluster represents a group of text messages with nearly the same meaning. There were 559 clusters with five or more unique messages. We hired an Indian journalist with fact-checking experience during the 2019 Indian general election to annotate each of these clusters for the quality of the clustering and to identify clusters with claims that could be fact-checked as defined by Konstantinovskiy et al. [55], which excludes several statement categories such as personal experience and spam. The annotation interface presented three examples from each cluster: one with the lowest average distance from all other messages in the cluster, one with the highest average distance from all other messages in the cluster, and one message chosen randomly. We found 257 clusters (out of the 559, 46%) comprising 2,536 unique messages were claims that could be fact-checked. Overall, 173 clusters (1,945 unique messages, 7,131 total messages) were related to the election, and 84 clusters (591 unique messages, 2,473 total messages) were claims unrelated to the election.

The clusters were generally all high-quality: in 98% of the clusters all three messages made the same claim. In 2% of the clusters (11 clusters, 159 unique messages) the three items annotated

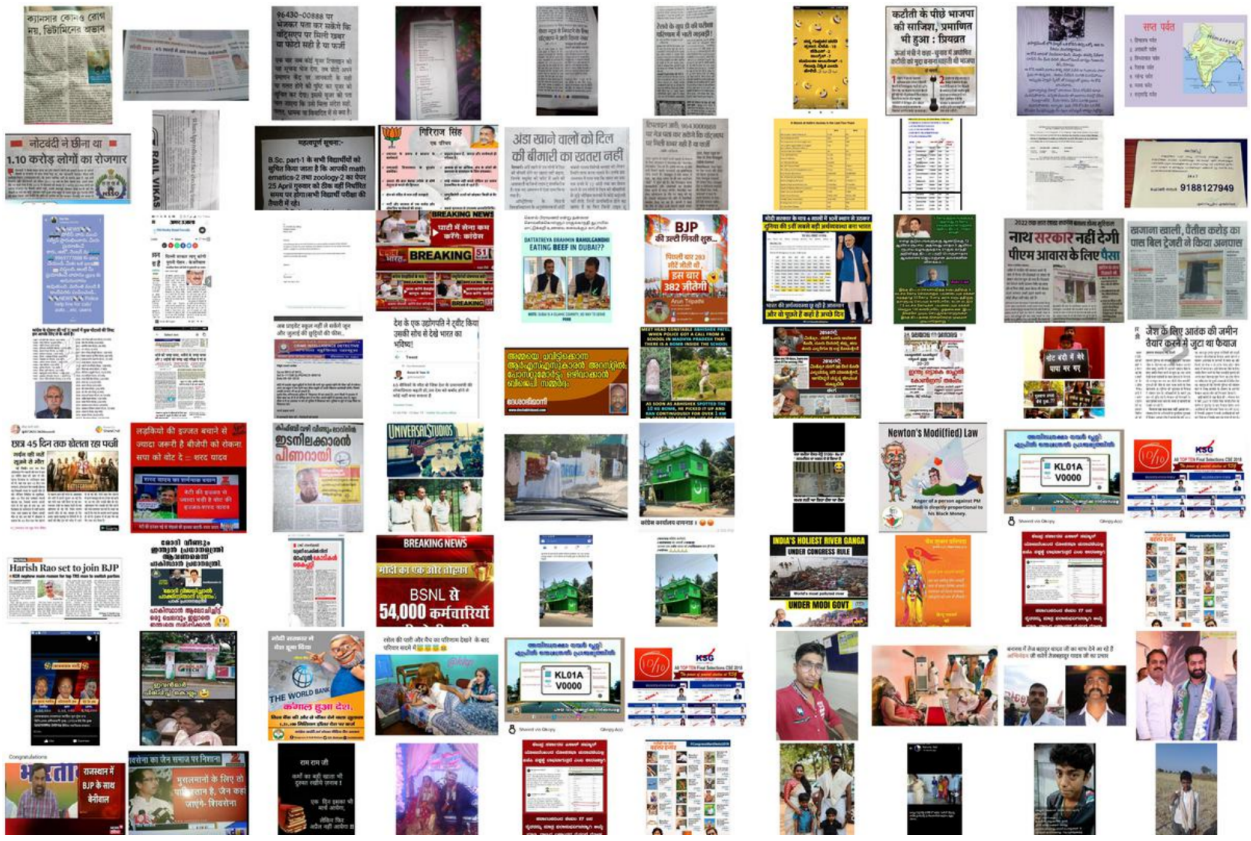

Figure 2.8: A visual summary of the images submitted to the tipline. The mosaic is a collection of 20 clusters obtained from the 34K images submitted to the tipline. Each cluster is represented as 2x2 grid of images randomly sampled from the cluster.

should not have been clustered together.

There were also 231 clusters that did not have fact-checkable claims. These were usually advertising/spam (114 clusters, 1,245 unique messages) or messages specific to the tipline (177 clusters, 2,957 unique messages). The tipline-specific messages include messages following up on submitted pieces of content, requests for more information about the tipline, and requests for fact checks in additional languages.

We took the 257 clusters that were annotated as containing claims and found that 203 contained messages in only one language (usually Hindi) while the other clusters contained between two and six languages. Languages were detected via CLD3 and were selected when a known language was detected and that detection was reported as reliable by CLD3[8]. Within the clusters with election-related claims, the largest cluster was misinformation advising voters to ask for a "challenge vote" or "tender vote" if they find they are either not on the voter list or have been marked as already voting[9]. There were 213 unique messages totaling 2,121 submissions to the

[8] https://github.com/google/cld3

[9] https://archive.is/BWsqR

tipline with this claim across five languages. Other prominent themes within the election-related clusters included messages attacking BJP leader Narendra Modi, pro-BJP messages, and messages criticizing Indian National Congress Party leader Rahul Gandhi.
The largest cluster with a non-election claim was misinformation about the tick marks on WhatsApp. It claims that three blue tick marks indicate the government had observed the message[10]. There were two clusters with different variants of this claim totaling 78 unique messages and 1,000 submissions across Malayalam and English.
Of the 2,536 messages in the clusters containing claims, Hindi (47%), English (35%), and Malayalam (6%) were the most common languages. Marathi, Telugu, and Tamil each accounted for roughly 2% of the messages. This likely reflects both the socio-linguistic characteristics of India as well as the fact that the tipline was most heavily advertised in Hindi and English.
In total, there were 9,604 submissions to the tipline comprised of 2,536 unique messages annotated as containing fact-checkable claims (i.e., 7,068 submissions within the set are exact duplicates). It took an average of 5 hours (SD = 1.4) for half of the total number of submissions in each of the clusters with claims to arrive to the tipline. 90% of the submissions in each of these clusters arrived within an average of 128 hours (SD = 17). This suggests slightly slower dynamics than those that have been seen with the signing of petitions [56] and the sharing of news stories on non-encrypted social media [57].

2.5.3.3 URLs

Another common content type in WhatsApp groups and tiplines is URLs. The tipline received 28,370 URLs (12,674 unique URLs), which contained URLs from 2,781 unique domains. A list of most frequent domains is presented in Table 2.3. The most prevalent websites submitted to the tipline were social media (YouTube, Facebook, Twitter, and Blogger), news outlets (IndiaTimes and DailyHunt), and URL shortening services (Bitly and TinyURL).

2.6 Discussion and Conclusion

Our results show the effectiveness of tiplines in content discovery for fact-checking on encrypted platforms. We show that:
A majority of the viral content spreading on WhatsApp public groups and on ShareChat was shared on the WhatsApp tipline first, which is important as early identification of misinformation is an essential element of an effective fact-checking pipeline given how quickly rumors can spread [6]. The tipline covers a significant portion of popular content: 67% of images and 23% of text messages shared more than 100 times in public WhatsApp groups appeared on the tipline.

[10] `https://archive.is/BWsqR`

Table 2.3: Top 10 domains most shared with the WhatsApp tipline around the Indian general election period.

Domain	Total URLs
YouTube	2,350
Blogger	2,107
Bitly	1,636
Google	1,471
Facebook	1,192
RechargeLoot	724
IndiaTimes	587
DailyHunt	574
Twitter	515
TinyURL	465

Compared to content from popular fact-checking organizations, the messages sent to tiplines cover a much higher proportion of WhatsApp public group messages. While misinformation often flows between platforms [58], this suggests that tiplines can capture unique content within WhatsApp that is not surfaced by fact-checking efforts relying on platforms without end-to-end encryption. These insights demonstrate tiplines can be an effective privacy-preserving, opt-in solution to identify potentially misleading information for fact-checking on WhatsApp and other end-to-end encrypted platforms. At the same time, there is the possibility of malicious uses and attacks on tiplines that may negatively affect fact checkers, share personal information from others, or poison the dataset. As we discuss in the findings, it is necessary to filter spam and other low-quality submissions. We analyzed submissions qualitatively to identify those with a claim that could be fact-checked, but there are several machine-learning approaches in development for this task [59, 60]. Tiplines, like systems for content moderation, must prioritize fact checkers' mental health [61]. The Meedan software used in the Checkpoint project, for instance, now uses Google's SafeSearch API to place a content screen over potentially explicit images. Similar systems, however, are needed to protect fact checkers from vicarious trauma as well as personal attacks in audio, video, and text in the myriad languages in which fact checkers operate. We can further reduce harm and malicious activity by designing friction into tiplines such as menu systems and limits on the number of requests per user to prevent denial of service attacks. We are currently investigating the data governance and safeguards needed to share tipline data more widely with academics for research [62].

In addition to the general public, we see three main stakeholders who could benefit from this research: academics, fact-checking organizations, and social media companies. Researchers or journalists trying to use data from encrypted social media apps like WhatsApp could make use of

data from such tiplines to study WhatsApp. The current model for identifying and fact-checking viral content on WhatsApp is to monitor conversations in a convenience sample of public WhatsApp groups [45, 63]. However, this requires technical skill and is resource intensive to manage. To our knowledge, monitoring of public groups has occurred only in academic settings.

Another solution that fact-checking organizations follow is to monitor non-encrypted social media platforms such as Facebook or Twitter and assume that content viral on one of these platforms likely overlaps with viral content on other platforms. Our work shows that there are far more matches between tipline content and public group messages on WhatsApp than between public group messages and either published fact checks or open social media content. This notable difference in the coverage of WhatsApp public groups stresses the opportunity tiplines provide for identifying misinformation on encrypted platforms. Although the volume of messages sent to the tipline is only 10% the volume of messages in the public groups, our analysis shows that tiplines can effectively help discover the most viral content being shared in the public groups. As end-to-end encryption prevents other forms of monitoring, identifying the most popular content on an end-to-end encrypted platform is useful to fact checkers, even if only a subset of that content is actual misinformation. The data we have for analysis does not include the fact-checks for the content submitted to the tipline, but our analysis shows that the majority of content submitted to the tipline contains claims that can be fact-checked.

Further research is needed to determine the best way fact checkers can prioritize content submitted to tiplines, filter spam and low-quality materials, combine signals from other platforms (e.g., from CrowdTangle and/or Twitter), and study the impact of fact-checks distributed via tiplines. Some methods, such as claim extraction [59, 60] and claim matching [42, 64], are directly applicable to tiplines, while other aspects require further work. Our analysis shows content is often submitted to tiplines before spreading in larger groups; however, this is only one step of the fact-checking progress. To be effective, we need systems that help fact checkers prioritize content for fact-checking, respond to that content, and disseminate fact-checks before the problematic content spreads widely. Nakov et al. [32] provide an overview of many ways in which further research and tool development could assist human fact checkers, and nearly all of these are applicable to tiplines as well. Our data predates the introduction of the "frequently-forwarded" flag on WhatsApp, but a report from Spanish fact-checking organization Maldita.es suggests this flag can be very useful for prioritizing content from WhatsApp tiplines [65].

Our analysis also found that most users sending content to the tipline were motivated to have the content they sent fact-checked: users would often follow up on content they submitted if it had not yet been fact-checked. We are unaware of any successful tiplines run solely as research projects, which suggests that fact-checking organizations and academics will need to partner together to scale tiplines and create meaningful tipline experiences for users. This will involve

setup costs and take time to foster dedicated contributors who are willing to forward potentially misleading content to a tipline.

It's worth noting that the tipline, public group, and fact-check content we studied were drawn from a specific period of time around a large political event (the 2019 Indian general election). It is unclear how the dynamics would differ for a less eventful time period. Several always-on WhatsApp misinformation tiplines were launched in December 2019, and the number has grown since. We encourage researchers to support civil society organizations running these tiplines, as they represent a valuable way to better understand the dynamics of misinformation on such end-to-end encrypted platforms.

Tiplines can also be used to collect hashes of popular misleading or hateful content. Hashes are small 'signatures' or 'fingerprints' that do not contain the original content but can be used to identify very similar content. Hashes can thus be used to develop on-device solutions that work in encrypted settings. For instance, Reis et al. [49] examine images and propose an on-device approach to alerting users to content that has been fact-checked on WhatsApp. Their solution focuses on PDQ hashes for images and requires a list of hashes for known pieces of misinformation. Our analysis in this chapter suggests that tiplines could be a successful way to populate such a list. The most popular images are likely to be submitted to a tipline, and, even better, they are very likely to be submitted to the tipline before they are widely shared within public groups. Thus, if a list was populated based on images sent to tiplines, it might identify many these shares.

Using advances in the state-of-the-art techniques to find similar image and text messages, an on-device fact-checking solution could identify up to 40% of the shares of potential misinformation in public WhatsApp groups while preserving end-to-end encryption if content can be prioritized appropriately and responded to quickly. Such a solution could operate similar to personal antivirus software where individuals can choose from a variety of vendors and fully control what happens when a potential match is identified.

In this first chapter, we unfolded some of the dynamics of misinformation on end-to-end encrypted social media, through the case study of tiplines, a crowdsourced opt-in tool for uncovering misinformation in closed social media, during the coverage of the 2019 Indian general election on public WhatsApp groups. In the following chapter, also contributing to the first research question (NLP for human-centered understanding of misinformation), we discuss how user background and demographics determine their belief in misinformation, and how the link varies across different demographic groups.

CHAPTER 3

Toward Understanding the Role of Demographics in Misinformation Perception

How are individuals different in believing news, and what do users' background and demographics tell us about their patterns of misinformation perception? In this chapter, we investigate the role of demographics in the perception of misinformation by measuring the effect that demographic information has on predicting the users' belief of news. We develop models of users' perception of news through fine-tuning pretrained text-to-text language models on frames of user reactions to headlines, generated from a disaggregated, user-centric version of the Misinfo Reaction Frames (MRF) dataset. We find that incorporating knowledge about users' demographics helps to model users more accurately, especially when making predictions about the misinformation class, with absolute improvements in F scores of up to 5.18% across demographics. We further present analyses on models that consider the group to which individuals perceiving misinformation belong, and find that a similar trend holds across demographic subgroups but with varying degrees of performance improvement in the presence of user demographics. We observe a linear relationship between in-group homogeneity and the effect of demographics on belief in news, further suggesting that users belonging to similar-minded groups are more predictable based on their demographic and background information. Our study provides insights as well as a framework, for understanding the connection between demographics and the perception of misinformation.

3.1 Introduction

Misinformation affects various pockets of society differently; for instance the socioeconomically disadvantaged and people experiencing significant inequalities are shown to be more vulnerable to conspiracy theories and misinformation [66], and there is a strong link between helplessness (e.g., due to poverty or inequality [67, 68]) and *pareidolia* [69] – the tendency to see patterns where they do not necessarily exist. Upward mobility and sound socioeconomic decisions for individuals in said groups can be challenging if they fall down the rabbit hole of conspiracies

Figure 3.1: Even though Alice and Bob are reading the same news, they have different reactions to it, and may perceive its veracity differently as portrayed in this example from the disaggregated MRF dataset.

[70]. To address such problems, it is important to learn more about the complex relationship between demographics and misinformation vulnerability.

Additionally, individuals of various backgrounds often exhibit diverse patterns of belief or skepticism when it comes to different types of news. This phenomenon is influenced by numerous factors, including one's cultural, social, and political affiliations [71, 72]. Individuals may be more inclined to believe news that aligns with their pre-existing beliefs or reinforces their worldviews. The credibility of news sources plays a crucial role; individuals from diverse backgrounds might place varying levels of trust in different media outlets based on their perceived bias or reliability. Socioeconomic factors can also impact news consumption, as individuals with differing levels of education, income, and access to information may evaluate news through distinct lenses. Ultimately, understanding these variations in belief or skepticism is essential for promoting media literacy and fostering constructive dialogue among individuals of diverse backgrounds.

In this paper we explore the link between four demographic attributes: age, gender, education, and race, and misinformation perception. To the best of our knowledge, this is the first linguistics-inspired study of this phenomenon. Since individuals are much more complex than a description of their demographics and background [73, 74], in this study we take these attributes as proxies that can capture some of the latent factors affecting news perception. Our study makes the following assumption that if modeling a users perception of misinformation is improved by taking demographics into account, then it follows that the user demographics partly determines users news belief patterns.

We study the impact of demographics on modeling misinformation perception using a disaggregated and user-centric version of the **Misinfo Reaction Frames (MRF)** dataset [75]. We define misinformation perception as a binary classification task where a text-to-text language model is prompted with a news headline, and examples of past news headlines perceptions by the same

user, and is asked how they would perceive a given headline i.e., whether they think it is fake or real. We fine-tune the small FLAN-T5 [76] pretrained LM to build a predictive model of users perception of misinformation. What differentiates this task from misinformation detection is that the same headline might be perceived differently and inaccurately by various users, so in order to predict a user's perception of any given headline, it is required for models of misinformation perception to incorporate both knowledge about the query headline, as well as the user that is perceiving it as demonstrated in 3.1.

We show that in zero-shot settings, users are better modeled when the model is trained on users with known demographics, compared to a model that is trained on users with no demographic information. The improvements observed by training on users with demographics are up to 33.74% in zero-shot, a relative jump of about 150%. When modeling existing users (hot start settings), we also find varying degrees of improvement in misinformation perception when taking demographics into account across demographics subgroups, of up to 5.18% F1 (fake) score. We also find that the variance in the improvements can be explained by a measure of in-group homogeneity (i.e. Krippendorff's alpha), and that groups that agree more often on what is misinformation are also more predictable based on what groups they identify with.

Our contributions are as follows:

- A framework based on pretrained text-to-text language models for modeling individuals' perception of fake and real news; we conduct experiments based on this framework in zero-shot (i.e., new users) and hot start (i.e., existing users) settings. We use our framework to study the relationship between demographics and misinformation perception, providing experimental results and analyses for a deeper understanding of how individuals of different backgrounds perceive the veracity of news headlines.

- Our experiments suggest the perception of misinformation from users of all backgrounds' is better explained when their demographic information is taken into account by the models; F1 scores for the misinformation class as well as accuracy increased consistently in different settings and across groups when demographic information about the user was provided to the misinformation perception models.

- The degree to which demographics affect the belief in news varies across demographic groups; for instance the effect of demographics on misinformation perception is twice more likely among younger users compared to older ones. We present experimental evidence that in-group homogeneity might explain the variation on the effect of demographics; the more like-minded the group-members, the larger is the effect of demographics on F1 (fake) and accuracy scores.

3.2 Related Work

3.2.1 Demographics and Misinformation

Prior research shows a strong link between helplessness and *pareidolia* [69] – the tendency to see patterns where they do not necessarily exist, which consequentially may translate to increases in susceptibility to misinformation. Poverty and inequality have been shown to increase vulnerability to conspiracy theories and misinformation [66] for particular demographic groups.
Several studies in health communication have shown that cultural factors, such as spoken language, location, ethnic group or age, can affect individuals susceptibility to misinformation [77, 78, 79] thus leading to disparities. Speakers from different cultural backgrounds are often exposed to misinformation targeting their identities and beliefs, for instance, during the current COVID-19 pandemic in the US, health and vaccine related misinformation and conspiracy theories were found to be more pervasive in the Spanish speaking community, while politicized health misinformation was more pervasive in the English speaking community [80].

3.2.2 Demographic-Aware NLP

While most of the work in natural language processing has considered a "one size fits all" approach while solving user-centric tasks, such as misinformation detection, recent work has started to challenge this assumption and develop discrete [33] or continuous [34] user representations that encode information about the user background. Work has been done to construct demographic-aware word representations Bamman et al. [35], Garimella et al. [36], Welch et al. [37], where separate word representations are created for each demographic group being considered. A recent study by King and Cook [81] compared how to improve a language model with user-specific data using priming and interpolation, depending on the amount of data available, learning a new model for each user. Welch et al. [82, 83] explored predicting response time, common messages, and speaker relationships from personal conversation data. Zhang et al. [84] conditioned dialog systems on artificially constructed personas and Madotto et al. [85] used meta-learning to improve this process.

3.3 Data

We use a disaggregated version of the Misinfo Reaction Frames (MRF) dataset [75] that consists of annotations made by individual crowdworkers, as well as the annotators' demographic attributes such as age bracket, education, race, gender, and media diet that were optionally collected. Similar to the MRF dataset, instances of the disaggregated dataset include a headline, and

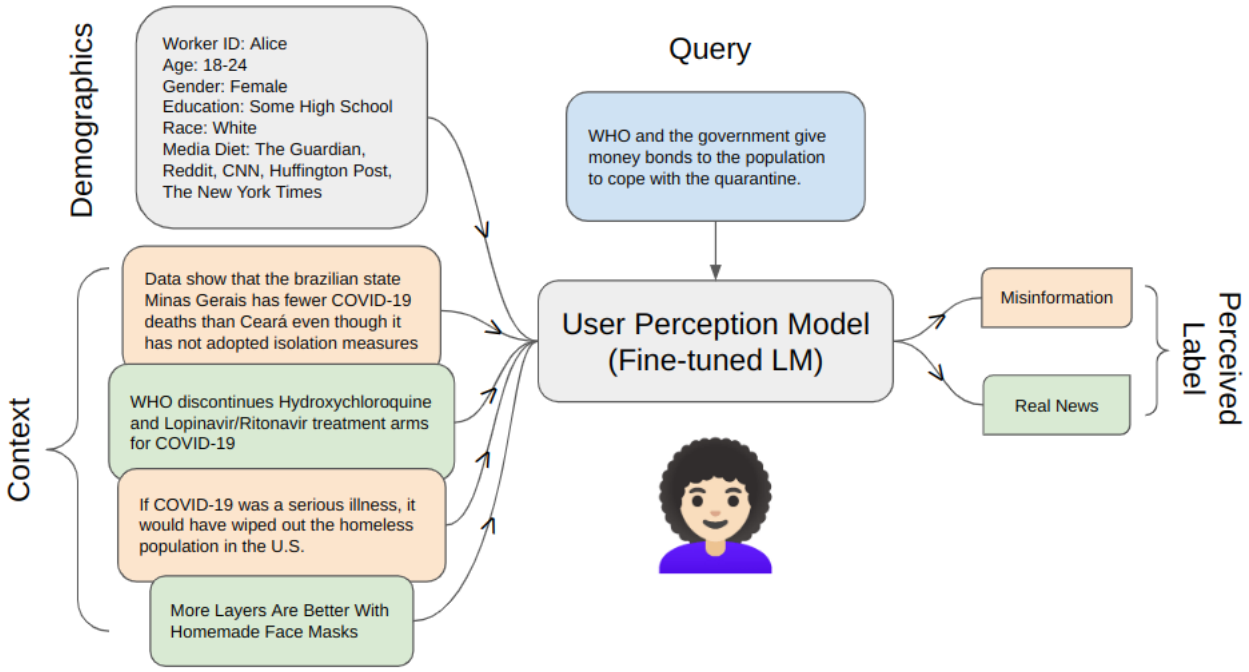

Figure 3.2: An instance of one data point generated from the MRF dataset as trajectories of user perceptions of headlines. The trajectories include a header that includes demographic information about the user, four to eight headlines, sorted temporally in the order they were annotated by the worker, and a query about user's perception of the following news headline. Alice's perception of this headline in this case is "real news."

how that headline is perceived by the annotator through reactions, and an indicator of their belief in the headline.

The main differences between the disaggregated dataset and the one released in Gabriel et al. 75 are: (i) data points in the MRF dataset map onto multiple data points in the disaggregated MRF, with potentially various annotations, (ii) the worker ID and demographic attributes (whenever available) are attached to the annotations. This additional information facilitates a deeper analysis of what drives perception of misinformation, and whether demographic attributes play a substantial role in that.

3.3.1 Generating User-Centered Frames of News Perception

We create a user-centered misinformation perception frame that includes a worker ID, worker demographic header (set to unknown when demographic information is missing), and a history of past reactions to headlines (between four to eight, chosen randomly at generation.) To produce these frames, we first group the annotations by worker IDs, drop workers with fewer than 100 annotations, and sort the annotated frames in temporal order for each worker. We run a sliding window over each worker's annotations, selecting four to eight samples of user misinformation

Group Assignment	A vs. A	A vs. B	B vs. B
A: Young, B: Old	63.11%	59.91%	62.70%
A: Men, B: Women	50.40%	65.16%	68.02%
A: College, B: < College	59.62%	60.72%	95.57%
A: White, B: Non-white	70.62%	64.13%	41.18%

Table 3.1: The in-group and inter-group agreement using Krippendorff's alpha for four demographic groups.

perception as reaction history, and the last item as the query headline. An example of a datapoint generated using this approach is presented in figure 3.2.

3.3.2 Dataset Statistics

The disagreggated MRF dataset includes 81 workers, 19 of whom have responded to at least one demographic question. Out of the 19 users with demographics there are: 10 women versus 7 men[1], 11 white versus 6 non-white (i.e. asian, black, latino, and mixed races), 10 college educated (having a bachelor's degree or higher) versus 7 below college (education under bachelor's level), and 5 under 35 years old versus 13 over 35 years. On average, each crowd worker has ˜50 user-centered misinfo perception frames (˜300 misinfo reaction frames). The most active worker with demographics has 159 datapoints, while the worker with the least frames has only 17 annotations. There is a total of 4,378 datapoints from users without demographic information, which make about 82% of the disaggregated MRF dataset.

We measure inter-annotator agreement among four demographic groups within the MRF dataset which are presented in table 3.1, as it can serve as a measure of in and out-group convergence in perceiving veracity of news. We calculate Krippendorff's alpha for in-group and inter-group agreement among crowdworker subgroups. The second and last columns in table 3.1 represents the agreement among in-group participants (e.g. Men vs Men and Women vs Women), while the third column reports inter-group agreement (e.g. Men vs Women.)

In each demographic group, we observe that inter-group agreement is lower than or equal one or both of the demographic subgroups: agreement between young-old groups is lower than young-young and old-old subgroup agreements, and in case of college educated vs below college subgroups- college vs non-college agreement is lower than or equal college-college and non-college-non-college agreements.

When comparing men-men with men-women, as well as non-white in-group agreement versus white-non-white, we observe higher agreement among subgroups than within the subgroup. This

[1]The option to choose a non-binary gender was provided to the annotators. However the recorded participation of non-binary crowdworkers is very low for the MRF dataset, which we deem as an important gap that requires further studies.

indicates that men agree more often with women than men in perceiving the authenticity of news headlines, and a similar statement is true about the non-white subgroup. In such cases, we can interpret the low in-group agreement among men and non-white subgroups as in-group non-homogeneity that may be caused because of reasons beyond gender or race [86]. For instance, the non-white subgroup comprises of black, asian, and latino annotators which makes the non-white group less homogenous compared to the white group.
We provide further analyses on the differences observed among demographic subgroups in 3.5.3, and we refer to the original MRF paper for more details about the dataset.

3.4 Methods

Misinformation perception can be a multifaceted phenomena to study, as there is a broad array of cognitive processes involved in individuals perception of news. Users may question the authenticity and veracity of the claims made in a headline, have certain sentiments evoked in them, or interpret the writer's intentions in various ways. Much of these aspects are encoded in aggregate in the *Misinfo Reaction Frames* [75] approach to studying perception of misinformation.
To focus on the effect of demographics on perception of misinformation on individuals, we use a disaggregated version of the MRF dataset, in which individuals share their perception of potentially misleading news. Additionally, we only take the perceived veracity label (which may differ from actual veracity label) of headlines into account when modeling misinformation perception. This enables us to go in depth in investigating the link between demographics and perception of misinformation in individuals.
We take advantage of the following observation in our study design: if modeling a user's perception of misinformation can be made more accurate when exposed to user's demographic metadata, then we can say that the user's demographics partly explain their misinformation belief patterns.

3.4.1 Problem Statement

Alice is presented with a potentially misleading news headline H with no information about the author. We are interested in predicting the probability that Alice perceives the veracity of the claims in H as real or fake news, $P(H, u_A)$. Similar to Alice, Bob may also come across H, but may perceive the headline differently, $P(H, u_B)$. Therefore modeling Alice or Bob's perception of H requires going beyond learning about real or fake news, as we also need to gain knowledge about Alice and Bob's behaviour and tendencies in perceiving news.
We assume that we only know (i) Alice and Bob's demographics, $D(u_A)$ and $D(u_B)$ and (ii) a few examples (e.g. 4 to 8) of past perception of news headlines, $P(H_i, u_A)$ and $P(H_i, u_B), i =$

Model	Acc.	F1 (fake)	F1 (real)
Users without Demographics	69.18%	22.50%	**79.18%**
+ Demographic Users	**71.46%**	**56.24%**	76.67%

Table 3.2: Zero-shot misinformation perception results: evaluating models pretrained on users with no demographic information (first row) and users with demographics (second row.)

Demographic Groups	- without demographics			+ with demographics			% of users
	Acc.	F1 (fake)	F1 (real)	Acc.	F1 (fake)	F1 (real)	w/ ↑ F1 (fake)
Age $<$ 35 *(N=5)*	74.49%	29.27%	**70.97%**	**74.60%**	**31.99%**	69.99%	**80%**
Age $>=$ 35 *(N=13)*	71.08%	25.54%	**67.25%**	**71.31%**	**27.77%**	66.67%	38%
Men *(N=7)*	73.45%	28.47%	**66.55%**	**73.56%**	**30.79%**	66.09%	**57%**
Women *(N=10)*	69.44%	27.90%	**67.02%**	**69.71%**	**30.54%**	66.09%	50%
College Educated *(N=10)*	**72.41%**	28.27%	67.62%	72.20%	**28.90%**	**67.73%**	50%
Bellow College *(N=7)*	69.21%	27.95%	**65.69%**	**70.01%**	**33.13%**	63.74%	**57%**
White *(N=11)*	68.50%	26.68%	**65.34%**	**68.94%**	**30.21%**	64.05%	45%
Non-White *(N=6)*	**75.84%**	30.82%	69.56%	75.62%	**31.44%**	**69.83%**	**67%**

Table 3.3: Hot-shot misinformation perception classification results for different demographic groups. *% of users w/ ↑ F1 (fake)* is the percentage of users in each group that experienced an improvement in performance in the presence of demographic information about them.

$1, 2, ..., N$. Given all of these as observations for Alice and Bob, we want to predict if they believe the news headline H by learning an estimate of Alice and Bob's model of news perception, $\hat{P}(u_A)$ and $\hat{P}(u_B)$.

3.4.2 Modeling Perception of Misinformation Using Pretrained Language Models

Motivated by the modeling described in 3.4.1, we build classifiers of misinformation perception for users through fine-tuning a text-to-text pretrained language model (PLM). Our few-shot fine-tuning task (as in figure 3.2) is designed to follow our conceptual problem statement: a user-centered frame of news perception includes four to eight instances of reaction to headlines, any demographics available about the user (or "unknown" if not), and a query news headline. Our goal is to predict how the query headline is perceived by the user (i.e. misinformation or real). Our formulation is flexible enough to accommodate both zero-shot perception prediction (i.e. for modeling new users), as well as supervised perception prediction (i.e. for modeling existing users.) This is due to providing our fine-tuned PLMs with past context in input, so they can make a reasonable guess about a new user's perception of headlines, as well as old ones.

3.5 Experiments

Using the user-centric and transformed MRF dataset, we design and run experiments that allow us to probe the role of demographics in misinformation perception.
Our experiments are two fold: zero-shot and hot start. In the zero-shot setting we know nothing about the users being modeled, while in the hot start portion of experiments we focus on developing models of individuals based on their data.

3.5.1 Experimental Setting

We conduct our experiments using the small Flan-T5 [76] pretrained language model, PyTorch Lightning [87], and the transformers library [88] on three Nvidia 1080Ti GPUs.
Evaluations are cross-validated (either 10 fold per user or 19 fold, one fold per user), and we report accuracy and per-class F1 scores, denoted as *F1 fake* and *F1 real*. Since our task formulation follows a text-to-text structure, we count any generated output that does not follow our desired output format, i.e., generating anything other than *Perceived Label: misinformation or real news*, as the wrong prediction.

3.5.2 Zero-Shot Misinformation Perception Modeling

To learn about the relationship that demographics play in an individual's perception of misinformation, we fine-tune a FLAN-T5 model on users with and without demographic metadata, as reported in Table 3.2. The model trained on users without demographics (first row in table 3.2 is trained on more than 80% of all of our data, while the model that also includes users with demographics is trained in a *leave-one-user-out* strategy, meaning that for each user, the model from the first row is further fine-tuned on data from the other 18 users. Using leave one out testing in both settings, we report average zero-shot performance on all users for whom demographics are available.
The model trained on users with no demographic information underperforms in predicting unseen users for whom some demographic knowledge is known in test time. Once the same model is fine-tuned on users with accompanying demographic data, misinformation perception performance on unseen users increases a relative 150%, as the F1 (fake) metric jumps from 22.50% to 56.24%. Similarly, a 2.28% absolute increase is observed in accuracy, while the "no demographics" model performs better in the F1 (real) score by 2.51%.
The noticeable jump in the F1 (fake) score indicates that even though we are testing on unseen users, it makes a difference for our models to learn about users for whom demographics are available. Consequently, we can infer that models that observe demographics in training are

better at predicting the behaviour of new users with demographics. At a high level, these findings hint at how individuals' backgrounds shape their belief in news.

3.5.3 Perceiving Misinformation in Various Demographic Groups

To learn more about how people of different demographics perceive misinformation, we construct misinformation perception models for individuals based on a subset of their data. This setting is in contrast to the zero-shot experiments, where the fine-tuned PLM is never exposed to information about the user.

We group users with similar demographics (similar to the grouping in 3.3.2) and build models of individuals' perception of misinformation on top of the pretrained models from the previous section. We use 10 fold cross validation (8 training folds, 1 validation fold, and 1 test fold), and report average results of individuals within each demographic group in table 3.3. The last column of the table refers to the percentage of users in each group with performance increase in the presence of their demographic information.

The consistent average improvement of F1 (fake) score and accuracy when user demographics are provided to models suggest similar trends as observed in table 3.2 about the zero-shot results. Across all four groups of users, F1 (fake) scores– which evaluate models' quality in making predictions about perceiving misinformation, improve when models know more about *who* the users are, which implies various users are susceptible to misinformation because of their background and demographics.

F1 (real) score is moderately higher when demographics are absent, which is also similar to the zero-shot setting. However this difference is smaller compared to improvement of F1 (fake) scores: 0.73% average difference versus 2.48% average difference. The largest improvement in the presence of user demographics is also in the F1 (fake) score, where the below college group has the largest performance improvement (5.18%) of all among F1 (fake) scores.

There are also noticeable discrepancies among demographic subgroups, most visibly in the % of users with improved F1 (fake) score, between younger and older, and white and non-white subgroups. 80% of users under 35 years old were better modeled per F1 (fake) score if the models used their demographic information, in comparison to only 38% of those aged over 35 years. There is also a 22% difference when comparing users who identified as white versus those who identified as other races (67% versus 45%).

3.5.4 In-Group Homogeneity and The Effect of Demographics

Following the results discussed in 3.5.3, the observed effect of demographics on misinformation perception varies across demographics: some groups such as those educated below bachelor's

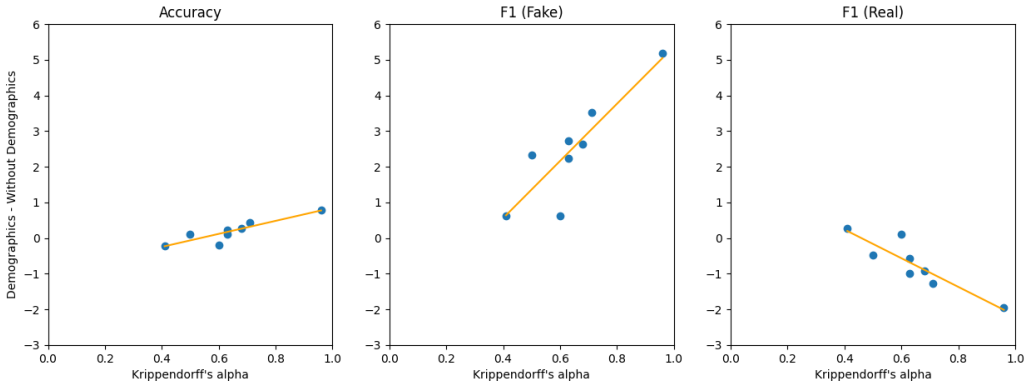

Figure 3.3: In-Group homogeneity (x-axis) graphed against the effect of demographics (y-axis) on accuracy, f1 (fake), and (real) scores.

level have a demographic effect of 5.18% in perceiving misinformation, while non-white participants only see a 0.62% difference on the same metric. Since the lowest difference in F1 (fake) scores is observed on the non-white group, which is a racially diverse and non-homogeneous group comprising of asian, latinx, black, and racially mixed users, the observation motivated us to investigate the relationship between in-group homogeneity and the effect of demographics in misinformation perception.

In figure 3.3 we graph the effect of demographics (i.e. the performance difference in the presence of user demographic information) in misinformation perception modeling across the three reported metrics (accuracy, f1 (fake), and (real) scores) against Krippendorff's alpha, which serves as a measure of group homogeneity for misinformation perception.

A clear trend that is observed across the three metrics is the linear relationship between in-group Krippendorff's alphas and the effect of demographics on misinformation perception modeling. The slope of the linear relationship is positive for accuracy and F1 (fake), meaning more in-group homogeneity is correlated with better predictability of misinformation perception in the presence of demographics. Compared with the misinformation class, the F1 (real) score has a smaller but negative slope.

3.6 Discussion and Future Work

Our experiments in this paper are based on a central assumption that if users are better modeled using their demographics, it means their demographics partly play a role in how they perceive misinformation. In both sets of experiments, we saw how demographic information about users can make models of misinformation perception better at identifying when users perceive headlines as misinformation, and being more accurate overall. In zero-shot misinformation perception

we modeled the misinformation class 2.5 times better when we fine-tuned the base model with data from users with demographics attached, indicating that **adding information about the background of users significantly improved zero-shot performance** in our task. Through modeling individual users with 10 fold cross validation, we also observe that **users of all demographic subgroups are better explained when the models know about their demographics**. These findings emphasize the role of demographics in how the users perceive misinformation. While there is no linear mapping of single demographic attributes to perception of misinformation in our results, we demonstrate that gender, education, race, and other relevant background information can determine how we perceive and believe news. Given how demographics makes a model of users perception of news more accurate, we can deduct that a complex link between users demographics and how they perceive news exists.

While the magnitude of the effect of demographics on perception of misinformation varies among demographic subgroups as portrayed in table 3.3, we see a clear correlation between in-group homogeneity and the effect of demographics on misinformation perception. **The more a group agrees within themselves on how to perceive certain headlines, the more their perception and belief of news follows their group identity**, which is captured in the F1 (fake) and accuracy differences in presence and absence of demographics in modeling individuals. On an intuitive level, if you belong to a like-minded group, it is easier to predict your belief in news based on the reactions of other members of your group. The relationship between group identity and belief in news and conspiracy theories has previously been confirmed [71, 72], and our results follow a similar trend.

In future work we plan on constructing larger scale datasets of misinformation perception so we can confirm and generalize our findings onto a broader scope, and include other demographic, socioeconomic, and behavioural attributes about the users under study. Additionally, we aim to study the different aspects of misinformation perception modeling, and to go beyond veracity labels into exploring the variations in reactions and inferences about misinformation and how those reactions might link with user demographics.

3.7 Conclusion

In this paper we modeled individuals' perception of misinformation, and studied how their backgrounds and demographics can help explain their beliefs about misinformation. Through several user-centered modeling experiments, we found that knowledge about a user's demographics such as age, gender, education, and race can improve models of misinformation perception. The improvements in accuracy and F1 (fake) scores happen both in zero-shot (up to 27.32%) and hot start (up to 5.18%) settings. These improvements indicate a link between susceptibility to

misinformation and demographic information about the user and the group they identify with. We presented in-depth analyses of results across demographic groups and found that there are discrepancies in how well demographics can help explain perception of misinformation. We found it is more than twice likely that demographics will help improve misinformation perception modeling in users under 35 years old than users over 35 years of age (80% versus 38%). Furthermore, we found that the more homogeneous a group becomes in perceiving misinformation, the user demographics play a larger role in determining their belief in misinformation, which is in line with prior findings on group identity and belief in conspiracies. We hope that our findings inspire future large scale studies in how people perceive online (mis)information.

In the past two chapters, we contributed to RQ1 by gaining a human-centered understanding of the phenomena of misinformation, through using NLP to learn about the dynamics of misinformation on WhatsApp (chapter 2), or investigating the relationship between users background and how they believe news (this chapter). In the next chapter, we begin addressing the second research question: building human-centered NLP to support fact-checkers in scaling up their efforts in countering online misinformation.

CHAPTER 4

Claim Matching

Manual fact-checking does not scale well to serve the needs of the internet. This issue is further compounded in non-English contexts. In this chapter, we discuss claim matching as a possible solution to scale fact-checking. We define claim matching as the task of identifying pairs of textual messages containing claims that can be served with one fact-check. We construct a novel dataset of WhatsApp tipline and public group messages alongside fact-checked claims that are first annotated for containing "claim-like statements" and then matched with potentially similar items and annotated for claim matching. Our dataset contains content in high-resource (English, Hindi) and lower-resource (Bengali, Malayalam, Tamil) languages. We train our own embedding model using knowledge distillation and a high-quality "teacher" model in order to address the imbalance in embedding quality between the low- and high-resource languages in our dataset. We provide evaluations on the performance of our solution and compare with baselines and existing state-of-the-art multilingual embedding models, namely LASER and LaBSE. We demonstrate that our performance exceeds LASER and LaBSE in all settings. We release our annotated datasets[1], codebooks, and trained embedding model[2] to allow for further research.

4.1 Introduction

Human fact-checking is high-quality but time-consuming. Given the effort that goes into fact-checking a piece of content, it is desirable that a fact-check be easily matched with any content to which it applies. It is also necessary for fact-checkers to prioritize content for fact-checking since there is not enough time to fact-check everything. In practice, there are many factors that affect whether a message is 'fact-check worthy' [89, 59], but one important factor is prevalence. Fact-checkers often want to check claims that currently have high viewership and avoid fact-checking 'fringe' claims as a fact-check could bring more attention to the claims—an understudied process known as amplification [90, 91]. While the number of exact duplicates and shares of a message

[1] https://doi.org/10.5281/zenodo.4890949
[2] https://huggingface.co/meedan/indian-xlm-r

Table 4.1: Example message pairs in our data annotated for claim similarity.

Item #1	Item #2	Label
पाकिस्तान में गनपॉइंट पर हुई एक डकैती को बताया जा रहा है मुंबई की घटना	कराची पाकिस्तान में घटित लूट को मुंबई का बताया जा रहा है।	Very Similar
பாகிஸ்தானில் உள்ள இந்திய தூதர் உடனடியாக டெல்லி திரும்ப மத்திய அரசு உத்தரவு	*செய்திகள்24/7* *FLASH* *பாகிஸ்தானில் உள்ள இந்திய தூதர் டெல்லி திரும்ப மத்திய அரசு உத்தரவு என தகவல்..*	Very Similar
Barber's salon poses the biggest risk factor for Corona! This threat is going to remain for a long duration. *At an average a barber's napkin touches 5 noses minimum* The US health dept chief J Anthony said that salons have been responsible for almost 50% deaths.	*The biggest danger is from the barbershop itself*. This danger will remain for a long time. *Barber rubs the nose of at least 4 to 5 people with a towel,* The head of the US Department of Health J. Anthony has said that 50 percent of the deaths in the US have happened in the same way that came in saloons.	Very Similar
ഇവിടുത്തെ മാമ മാധ്യമങ്ങൾ Live കാണിച്ചില്ലേലും ദേശീയ മാധ്യമങ്ങൾ ചെയ്തു കേട്ടാ	ഇവിടുത്തെ മാധ്യമങ്ങൾ കാണിച്ചില്ലേലും ദേശീയ മാധ്യമങ്ങൾ ചെയ്തു കേട്ടാ	Very Similar
Guys important msg:- There is the news of military bsf & cisf coming to Mumbai and having a seven days to 2 weeks curfew....	*Just received information* Entire Mumbai and pune will be under Military lockdown for 10 days starts from Saturday....	Somewhat Similar
Don't believe this FAKE picture of PM Modi; here's the truth	Don't believe this FAKE picture of Virat Kohli; here's the fact check	Very Dissimilar

can be used as a proxy for popularity, discovering and grouping together multiple messages making the same claims in different ways can give a more accurate view of prevalence. Such algorithms are also important for serving relevant fact-checks via 'misinformation tiplines' on WhatsApp and other platforms [92, 93, 94].

Identifying pairs of textual messages containing claims that can be served with one fact-check is a potential solution to these issues. The ability to group claim-matched textual content in different languages would enable fact-checking organizations around the globe to prioritize and scale up their efforts to combat misinformation. In this chapter, we make the following contributions: (i) we develop the task of claim matching, (ii) we train and release an Indian language XLM-R (I-XLM-R) sentence embedding model, (iii) we develop a multilingual annotated dataset across high- and lower-resource languages for evaluation, and (iv) we evaluate the ability of state-of-the-art sentence embedding models to perform claim matching at scale. We formally evaluate our methods within language but also show clusters found using our multilingual embedding model often have messages in different languages presenting the same claims.

We release two annotated datasets and our codebooks to enable further research. The first dataset consists of 5,066 messages in English, Hindi, Bengali, Malayalam, and Tamil that have been triple annotated for containing 'claim-like statements' following the definition proposed by fact-checkers in Konstantinovskiy et al. [89]. The second dataset consists of 2,343 pairs of social media messages and fact-checks in the same five languages as the first dataset annotated for claim similarity. Table 4.1 shows examples of annotated pairs of messages from the second dataset.

4.2 Related Work

4.2.1 Semantic Textual Similarity

Semantic textual similarity (STS) refers to the task of measuring the similarity in meaning of sentences, and there have been widely adopted evaluation benchmarks including the Semantic Textual Similarity Benchmark (STS-B) [2017, 2016, 2015, 2014, 2013, 2012] and the Microsoft Research Paraphrase Corpus (MRPC) [101]. The STS-B benchmark assigns discrete similarity scores of 0 to 5 to pairs of sentences, with sentence pairs scored zero being completely dissimilar and pairs scored five being equivalent in meaning. The MRPC benchmark assigns binary labels that indicate whether sentence pairs are paraphrases or not.
Semantic textual similarity is a problem still actively researched with a dynamic state of the art performance. In recent work from Raffel et al. [102], the authors achieved state-of-the-art performance on STS-B benchmark using the large 11B parameter T5 model. The ALBERT model [103] achieved an accuracy of 93.4% on the MRPC benchmark and is considered one of the top contenders on the MRPC leaderboard.
While semantic textual similarity is similar to claim matching, the nuances in the latter require special attention. Claim matching is the task of matching messages with claims that can be served with the same fact-check and that does not always translate to message pairs having the same meanings. Moreover, claim matching requires working with content of variable length. In practice, content from social media also has wide variation in lexical and grammatical quality.

4.2.2 Multilingual Embedding Models

Embedding models are essential for claim and semantic similarity search at scale, since classification methods require a quadratic number of comparisons. While we have seen an increasing number of transformer-based contextual embedding models in recent years [104, 105, 106], the progress has been asymmetric across languages.
The XLM-R model by Conneau et al. [107] with 100 languages is a transformer-based model with a 250K token vocabulary trained by multilingual masked language modeling (MLM) with monolingual data and gained significant improvements in cross-lingual and multilingual benchmarks. LASER [108] provided language-agnostic representation of text in 93 languages. The authors trained a BiLSTM architecture using parallel corpora and an objective function that maps similar sentences in the same vicinity in a high-dimensional space. Language-agnostic BERT sentence embeddings (LaBSE) by Feng et al. [109] improved over LASER in higher resource languages by MLM and translation language modeling (TLM) pretraining, followed by fine-tuning on a translation ranking task [110].

4.2.3 Claim Matching

Shaar et al. [64] discussed retrieval and ranking of fact-checked claims for an input claim to detect previously debunked misinformation. They introduced the task, as well as a dataset covering US politics in English, and two BM25 based architectures with SBERT and a BERT-based reranker on top. Vo and Lee [111] tackled a similar problem by finding relevant fact-check reports for multimodal social media posts. However these projects only focus on English data that mainly cover U.S. politics and at least one of the matching pairs is a claim from a fact-check report. Additionally, the data collection process used in Shaar et al. [64] might not necessarily capture all possible matches for a claim, since the dataset is constructed by including only the claims mentioned in one fact-check report and not all previous occurrences. This may skew results and increase the risk of the model having a high false negative ratio. Recently, the Check-That! Lab 2020 [112] has presented the same problem as a shared task. We improve on prior work by finding a solution that works for high- and low-resource languages and also for matching claims between pairs of social media content and pairs of fact-checks. We explicitly annotated claim pairs that might match, avoiding the aforementioned false negatives issue by design and providing more accurate models and evaluations.

4.3 Data Sources

The data used in this chapter comes from a variety of sources. We use a mixture of social media (e.g., WhatsApp) content alongside fact-checked claims, since it is essential for any claim-matching solution to be able to match content both among fact-checked claims and social media posts as well as within social media posts. Among the prevalent topics in our data sources are the COVID-19 pandemic, elections, and politics.

Tiplines. Meedan, a technology non-profit, has been assisting fact-checking organizations to setup and run misinformation tiplines on WhatsApp using their open-source software, Check. A *tipline* is a dedicated service to which 'tips' can be submitted by users. On WhatsApp, tiplines are phone numbers to which WhatsApp users can forward potential misinformation to check for existing fact-checks or request a new fact-check. The first tipline in our dataset ran during the 2019 Indian elections and received 37,823 unique text messages. Several additional always-on tiplines launched in December 2019 and ran throughout the 2020 calendar year. We obtained a list of the text of messages and the times at which they were submitted to these tiplines for March to May 2019 (Indian election tipline) and for February 2020 to August 2020 (all other tiplines). We have no information beyond the text of messages and the times at which they were submitted. In particular, we have no information about the submitting users.

WhatsApp Public Groups. In addition to the messages submitted to these tiplines, we have data from a large number "public" WhatsApp groups collected by Garimella and Eckles [45] during the same time period as the Indian election tipline. The dataset was collected by monitoring over 5,000 public WhatsApp groups discussing politics in India, totaling over 2 million unique posts. For more information on the dataset, please refer to Garimella and Eckles [45]. Such public WhatsApp groups, particularly those discussing politics have been shown to be widely used in India [44].

Fact-Check Reports. We aggregate roughly 150,000 fact-checks from a mixture of primary fact-checkers and fact-check aggregators. We employ aggregators such as Google Fact-check Explorer,[3] GESIS [113], and Data Commons, and include roughly a dozen fact-checking organizations certified by the International Fact-Checking Network with either global or geographically-relevant scope in our dataset. All fact-checks included at minimum a headline and a publish date, but typically also include a lead or the full text of the fact-check, as well as adjudication of the claim (e.g., truth or falsity), and sometimes include information of lesser value for our work such as author, categorization tags, or references to original content that necessitated the fact-check.

4.4 Data Sampling & Annotation

To construct a dataset for claim matching, we design a two-step sampling and annotation process. We first sample a subset of items with potential matches from all sources and then annotate and select the ones containing "claim-like statements."
In a second task, we annotate pairs of messages for claim similarity. One of the messages in each pair must have been annotated as containing a "claim-like statement" in the first annotation task. We sample possible matches in several ways in order to not unnecessarily waste annotator time. We describe these sampling strategies and other details of the process in the remainder of this section.

4.4.1 Task 1: Claim Detection

Task 1 presented annotators with a WhatsApp message or fact-check headline and asked whether it contained a "claim-like statement."
We first created a codebook by inductively examining the English-language data, translations of the other-language data, and discussing the task with two fact-checkers (one Hindi-speaking and one Malayalam-speaking). We began with the definition set out by practitioners [89] for a "claim-like statement" and created examples drawn from our data sources. Annotators were

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

Table 4.2: Claim-like statements. κ is Randolph's marginal-free kappa agreement on the collapsed data (Yes/Probably, No, Incorrect language). All languages were annotated by three annotators.

Language	Items	κ	Majority Yes
Bengali (bn)	1093	0.30	30%
English (en)	1000	0.60	54%
Hindi (hi)	1000	0.59	41%
Malayalam (ml)	1025	0.63	69%
Tamil (ta)	948	0.63	21%

asked whether the message had a claim-like statement and allowed to choose "Yes", "Probably", "No", or "N/A: The message is not in language X" (where X was the language being annotated). The instructions made clear "Probably" should be used sparingly and was intended for instances where an image, video, or other context was missing. The detailed instructions and an example of the interface are provided in the supplemental materials.

We recruited three native speakers for each of Hindi, Bengali, Tamil, and Malayalam through Indian student societies at different universities as well as independent journalists. All of our annotators had a Bachelor's degree and many were pursuing Masters or PhDs. We onboarded all annotators and discussed the risks of possibly politically charged, hateful, violent, and/or offensive content in the dataset. Our custom-built annotation interface provided the ability to skip any piece of content with one keystroke. We also encouraged annotators to take frequent breaks and calculated these breaks into our payments.

Our English-language data is a mix of Indian and global content. Two of our English annotators had previously completed the Hindi and Malayalam tasks while the third English annotator completed only the English-language task.

We calculate agreement using Randolph's marginal-free kappa [114]. This measure better estimates intercoder agreement in unbalanced datasets compared to fixed-marginal scores like Fleiss' kappa [115].

All participants annotated 100 items independently. We then discussed disagreements on these 100 items and updated the codebook if needed. The participants then annotated datasets of approximately 1,000 items in each language. Information about this final annotation dataset is presented in Table 4.2. Agreement between annotators for this task is lower than the next task but on par with annotation tasks for hate speech and other 'hard tasks' [116, 117] suggesting determining whether a message has a claim-like statement is harder than determining the similarity of the statements (Task 2).

Table 4.3: Task 2 dataset. κ is Randolph's marginal-free kappa agreement on the collapsed data (Very Similar, Not Very Similar, N/A). "V. Sim." is the percentage of cases where two or more annotators indicated the pairs were "Very Similar."

Lang.	Pairs	κ	Annotators	V. Sim.
bn	644	0.64–0.68	2–3	6%
en	398	0.69	4	15%
hi	399	0.90	3	21%
ml	604	0.91	3	7%
ta	298	0.85	2	11%

4.4.2 Task 2: Claim Similarity

The second task presented annotators with two messages and asked how similar the claim-like statements were in the messages. Annotators were given a four-point scale ("Very Similar", "Somewhat Similar", "Somewhat Dissimilar", and "Very Dissimilar"). We prepared a codebook with clear instructions for each response and examples in consultation with the two fact-checkers and discussed it with all annotators before annotation began. Annotators could also select "N/A: One or more of the messages is not in language X or does not contain a claim-like statement"). Our initial testing showed the largest source of disagreement was between "Somewhat Dissimilar" and "Very Dissimilar." We added guidance to the codebook but did not dwell on this aspect as we planned to collapse these categories together. We prioritize our evaluations on "Very Similar" or "Somewhat Similar" statements.

Although our goal is claim matching, this task asked annotators about the similarity of claim-like statements as the annotators were not all fact-checkers. We found asking the annotators to speculate about whether some hypothetical fact-check could cover both statements was unhelpful. Our codebook is constructed such that "Very Similar" pairs of messages could be served by one fact-check while "Somewhat Similar" messages would partially be served by the same fact-check. A link to the codebook is in the supplemental materials.

The same annotators from Task 1 completed Task 2 with a few exceptions. One Tamil annotator was unable to continue due to time restrictions, and one Bengali annotator only completed part of the annotations (we calculate agreement with and without this annotator in Table 4.3). We added a fourth English annotator in case there was another dropout but all English annotators completed. Table 4.3 shows a breakdown of the dataset by language. In general, agreement on this task, even among the same annotators as Task 1, was much higher than Task 1 suggesting claim similarity is an easier task than claim detection. The largest point of disagreement was around the use of the N/A label: discussing this with annotators we found it was again the disagreement about whether certain messages had claims leading to the disagreement.

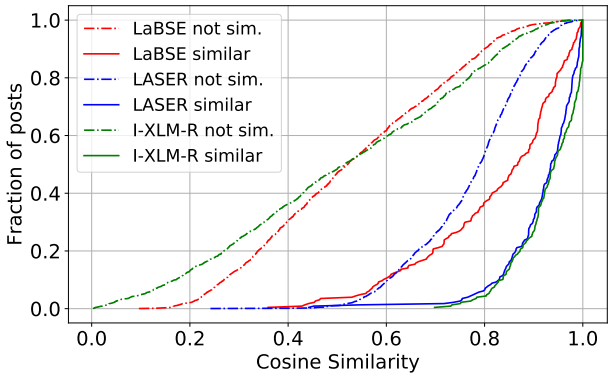

Figure 4.1: CDF of cosine similarities of all labeled data according to LASER, LaBSE, and I-XLM-R models. Legend: "similar" pairs were annotated by two or more annotators as being "Very Similar". "not sim." encompasses all other pairs, excluding "N/A" pairs.

4.4.3 Sampling

A purely random sample of pairs is very unlikely to find many pairs that match. We considered examining pairs with the highest cosine similarities only, but these pairs were likely to match in trivial and uninteresting ways. In the end, we used random stratified sampling to select pairs for annotation.

We first calculate all pairwise cosine similarities using multiple embedding models (described in Section 4.5). We then use stratified sampling to sample 100 pairs in proportion to a Gaussian distribution with mean 0.825 and standard deviation 0.1 for each model and language. We do this due to our strong prior that pairs close to zero as well as pairs close to one are usually 'uninteresting.' These represent pairs that either clearly do not match or (very often) clearly match. In practice, we still sample a wide range of values (Figure 4.1). We also include 100 random pairs for each language with the exception of Tamil due to annotator time limitations.

We used LASER, LaBSE, and our Indian XLM-R (I-XLM-R) model (details below) to sample pairs for all languages. Our Bengali and Malayalam annotators had additional capacity and annotated additional pairs drawn in a similar way.

4.5 Claim Matching Methods

4.5.1 Experimental Setup

We use a GPU-enabled server with one 1080 GPU to train our own embedding model and run the rest of our experiments on desktop computers with minimal runtime. We use the Elasticsearch implementation of the BM25 system and use the Sentence-Transformers (for I-XLM-R), PyTorch

(for LASER), and TensorFlow (for LaBSE)[4] to train and retrieve embeddings. We follow the approach of Reimers and Gurevych [118] for tuning the hyperparameters of our embedding model.

4.5.2 Training a Multilingual Embedding Model

We use the knowledge distillation approach presented in Reimers and Gurevych [118] to train a multilingual embedding model.[5] The approach adopts a student–teacher model in which a high quality teacher embedding model is used to align text representations of a student model by mapping embeddings of text in the student language to close proximity of the embeddings of the same text in the teacher language. Using this approach we train a model for English, Hindi, Malayalam, Tamil, and Bengali. We refer to this model as our Indian XLM-R model (I-XLM-R), and use it as one of the models we evaluate for claim matching.

Training Data. The knowledge distillation approach requires parallel text in both student and teacher languages for training embedding models. We find the OPUS parallel corpora [119] to be a useful and diverse resource for parallel data. We retrieve parallel data between English and the collection of our four Indian languages from OPUS and use it as training data.

Training Procedure. For a teacher model M_T and a student model M_S and a collection of (s_i, t_i) pairs of parallel text, we minimize the following MSE loss function for a given mini-batch B:

$$\frac{1}{|B|} \sum_{i \in B} [(M_T(s_i) - M_S(s_i))^2 + (M_T(s_i) - M_S(t_i))^2]$$

Intuitively, this loss function forces embeddings of the student model for both t_i and s_i to be in proximity of the teacher embeddings for s_i, therefore transferring embedding knowledge from the teacher to the student model. For training our Indian XLM-R model, we pick the English SBERT model as teacher [105] (for its high quality embeddings) and XLM-Roberta (XLM-R) as the student (for SOTA performance in NLP tasks and a universal vocabulary that includes tokens from 100 languages).

4.5.3 Model Architecture

We evaluate a retrieval-based claim matching solution built on top of the BM25 retrieval system [120] as well as an embeddings-only approach. In the first case, queries are fed into BM25 and

[4]We use `https://github.com/bojone/labse`.

[5]Trained models from Reimers and Gurevych do not include embeddings for Bengali, Tamil, and Malayalam, which motivated us to train the I-XLM-R model.

Table 4.4: MRR across different models and languages. Columns refer to reranking embedding models on top of BM25, with the exception of BM25 as the baseline.

Language	BM25	LASER	LaBSE	I-XLM-R
Bengali	0.4247	0.4170	0.4120	**0.5281**
English	**0.4286**	0.4247	0.4101	0.4221
Hindi	0.4524	0.4289	0.3675	**0.4849**
Malayalam	0.3903	0.3777	0.3651	**0.4023**
Tamil	**0.4747**	0.4050	0.4563	0.4634

the retrieved results are then sorted based on their embedding similarity to the input query. The top ranking results are then used as potential matches for the input claim. In the latter case, we classify pairs of items using features derived from the embedding models.

4.6 Results

For some applications, it is good enough to be able to rank the most similar claims and treat the problem of claim matching as an information retrieval problem. This is the case, for example, when fact-checkers are examining possible matches to determine if a new content item matches a previous fact-check. We discuss the performance of information retrieval approaches in Section 4.6.1.

In many other applications, however, we seek a system that can determine if the claims in two items match without human intervention. These applications demand a classification approach: i.e., to determine whether two items match. This allows similar items to be grouped and fact-checkers to identify the largest groups of items with claims that have not been fact-checked. We discuss the performance of simple classification approaches in Section 4.6.2.

4.6.1 Information Retrieval Approach

We find the mean reciprocal rank (MRR) metric to be a good IR-based performance measure for our system, since we only know of one match in the retrieved results by the system for our queries. We use the base BM25 system as a strong baseline to compare against. We also compare our system with other state-of-the-art multilingual embedding models used for reranking, namely LASER and LaBSE. Results are presented in Table 4.4.

The BM25 with I-XLM-R reranking outperforms other systems in all languages, with the exception of Tamil and English where the system performs comparably with the BM25 baseline. The largest lead in performance of the I-XLM-R based model is for Bengali, where the MRR score is more than 0.1 higher than the BM25 baseline.

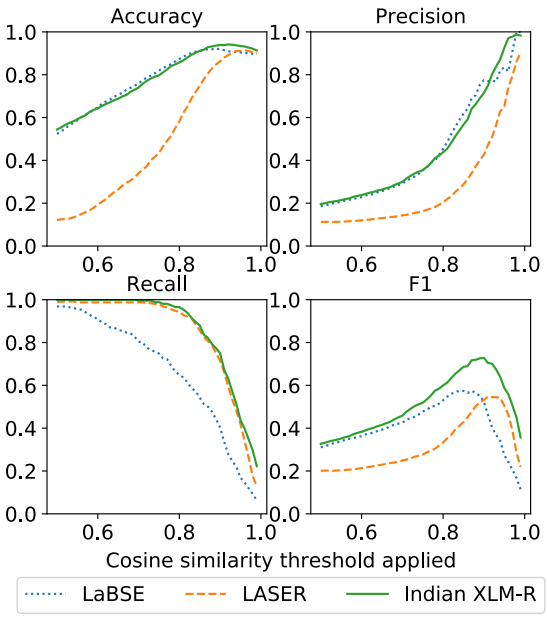

Figure 4.2: Accuracy, Precision, Recall, and F1 for simple thresholds on the cosine similarity scores.

Both LASER and LaBSE fall short on surpassing the baseline for any of the languages. LASER performs the worst on Tamil, where its MRR score is nearly 0.07 less than BM25. Similarly, LaBSE's largest difference with BM25 is in Hindi where it falls short by 0.085. Although there is room for improvement in some languages, the I-XLM-R seems the best choice if only one system is chosen.

After calculating MRR we also evaluated the systems on other metrics, namely "Mean First Relevant" (MFR, Fuhr [121]) and HasPositive@K [64]. Both measures did not demonstrate any meaningful patterns useful for selecting the best system. We do not include the details of these evaluations for brevity.

4.6.2 Classification Approaches

Responding to submitted content on a tipline, as well as grouping claims to understand their relative prevalence/popularity, requires more than presenting a ranked list as occurs in the information retrieval approaches in the previous subsection and in previous formulations of this problem [e.g., 64]. In this section we use the annotated pairs to evaluate how well simple classifiers perform with each model.

Threshold Classifier. The first 'classifier' we evaluate is a simple threshold applied to the cosine similarity of a pair of items. Items above the threshold are predicted to match while items with a similarity below the threshold are predicted to not match. In doing this, we seek to under-

Table 4.5: Maximum average F1 scores ± standard deviations achieved with 10 runs of 10-fold cross-validation and the corresponding thresholds (thres.) for each score. The 'classifiers' are simple thresholds on the cosine similarities.

	LASER	LaBSE	I-XLM-R
Language	F1 (thres.)	F1 (thres.)	F1 (thres.)
All	0.55±0.08 (0.91)	0.58±0.07 (0.84)	**0.73**±0.07 (0.90)
Bengali	0.68±0.21 (0.96)	0.58±0.23 (0.90)	**0.74**±0.19 (0.96)
English	0.85±0.09 (0.85)	0.77±0.15 (0.77)	**0.88**±0.10 (0.78)
Hindi	0.74±0.13 (0.88)	0.61±0.15 (0.87)	**0.82**±0.12 (0.87)
Malayalam	0.47±0.20 (0.92)	0.71±0.20 (0.85)	**0.79**±0.20 (0.89)
Tamil	0.26±0.21 (0.99)	0.50±0.20 (0.98)	**0.57**±0.15 (0.96)

stand the extent to which the embedding models can separate messages with matching claims from those with non-matching claims.

An ideal model would assign higher cosine similarity scores to every pair of messages with matching claims than to pairs of messages with non-matching claims. Table 4.5 shows the F1 scores averaged across 10 runs of 10-fold cross validation for binary classifiers applied to all languages and each language individually. In general, the Indian XLM-R model performs best at the task with F1 scores ranging from 0.57 to 0.88. As shown in Figure 4.2, our Indian XLM-R model outperforms LASER primarily in precision and outperforms LaBSE primarily in terms of recall. The numbers reported in Table 4.5's last column all come from I-XLM-R. The English-only SBERT model performs slightly better with a maximum F1 score of 0.90±0.09 at a threshold of 0.71 on English data, suggesting that the student model may have drifted from the teacher model for English during training. This drift is slight, however, and the cosine similarities across all English-language data for the two models are highly correlated with a Pearson's correlation coefficient of 0.93. The authors of SBERT released two additional multilingual models on that support English and Hindi, but do not support Bengali, Malayalam, or Tamil.[6] We find the models have comparable performance to I-XLM-R on English & Hindi while F1 scores for other languages are between 0.17 and 0.61.

Our dataset includes both social media messages (namely, WhatsApp messages) and fact-checks.

[6] `https://www.sbert.net/docs/pretrained_models.html` has 'xlm-r-distilroberta-base-paraphrase-v1' and 'xlm-r-bert-base-nli-stsb-mean-tokens'

Table 4.6: Claim matching classification results.

Model	Accuracy	F1 (+)	F1 (-)
LASER	0.805±0.064	0.789±0.087	0.814±0.039
LaBSE	0.797±0.059	0.791±0.067	0.800±0.055
I-XLM-R	**0.883**±0.036	**0.885**±0.036	**0.880**±0.037
All	0.868±0.036	0.868±0.036	0.866±0.039

Overall, performance is higher for matching fact-checks to one another than for matching social media messages to one another for all models. As an example, the best-performing model, Indian XLM-R, achieves a maximum F1 score of 0.76 with a threshold 0.87 for matching pairs of fact-checks, but only a maximum F1 score of 0.72 (threshold 0.90) for matching pairs of social media messages.

Claim Matching Classifier. We train an AdaBoost binary classifier that predicts if two textual claims match. The features are all precomputed or trivial to compute so that such a system could easily be run to refine a smaller number of candidate matches with minimal additional computation.

We use lengths of claims, the difference in lengths, embedding vectors of each item, and their cosine similarity as features. We build a balanced dataset by taking all the "Very Similar" pairs and matching every item with a randomly selected "Not Very Similar" (every other label) item from the same language. We do not differentiate between pairs in different languages as our per language data is limited and all features including the embedding vectors translate across languages as they are from mulitilingual embedding models.

Claim matching classification results are presented in Table 4.6. We evaluate models using 10-fold cross validation and report accuracy and F1 scores for each class averaged over 10 runs. Consistent with previous outcomes, it is clear that using the I-XLM-R cosine similarity and embeddings as input features results in better performance than other models, including the model with all features.

The positive class F1 scores for all models in Table 4.6 are notably higher than the threshold approaches (Table 4.5) suggesting information from the embeddings themselves and the lengths of the texts are useful in determining whether the claims in two messages match. The claim matching classifier is language-agnostic and is learning from only 522 datapoints, which underscores the quality of the I-XLM-R embeddings.

Error Analysis. We manually inspect the pairs classified in error using the "threshold classifier" and I-XLM-R. The pairs either have a similarity score above the matching threshold but

are “Not Similar” (false positives, 24/89) or are matches and have a score below threshold (false negatives, 65/89). 16 of the 24 false positives are labeled as “Somewhat Similar,” and manual inspection shows that these pairs all have overlapping claims (i.e., they share some claims but not others). There are no obvious patterns for the false negatives, but some of the errors are made in ambiguous cases.

We also examine the errors of one random fold of the AdaBoost classifier to further investigate where our model makes mistakes. There are a total of 10 wrong predictions (6 false negatives and 4 false positives). Of these, 2/6 and 1/4 are annotation errors. Within the false negatives, most other cases are pairs of text that are very similar but minimally ambiguous because of a lack of context, which annotators correctly resolved to being identical. An example of such a false negative is the pair of messages “Claim rare flower that blooms once in 400 years in the-himalayas-called-mahameru-pushpam” and “Images of Mahameru flower blooms once every 400 years in Himalayas.” False positives were all “Somewhat Similar” and “Somewhat Dissimilar” pairs that the classifier mistook for “Very Similar.” There were no significant discrepancies among languages in classification errors.

4.7 Discussion & Conclusions

Scaling human-led fact-checking efforts requires matching messages with the same claims. In this chapter, we trained a new model and created an evaluation dataset that moves beyond English and American politics. Our system is being used in practice to support fact-checking organizations.

We found that the embedding models can generally match messages with the same claims. Performance for matching fact-checks slightly exceeded that for matching social media items. This makes sense, given that fact-checks are written by professional journalists and generally exhibit less orthographical variation than social media items.

Too few examples of fact-checks correctly matched a social media item to evaluate performance in that setting. This is not a major limitation since nearly every fact-check starts from a social media item. So, in practice we only need to be able to match social media items to one another in order to locate other social media items having the same claims as the item that led to a fact-check.

We evaluated claim matching within each language, but the embedding models are all multilingual and could serve to match claims across languages. BM25 is not multilingual, but Elasticsearch can index embeddings directly. Previously de Britto Almeida and Santos [122] developed a Elasticsearch plugin to query embeddings by cosine distance, but since version 7.3 of Elasticsearch this functionality is now available natively in Elasticsearch [123], meaning a large set of

embeddings can be searched efficiently to find near matches across languages.

As a proof of concept, we took the 37,823 unique text messages sent to the Indian election tipline and clustered them using I-XLM-R and online, single-link hierarchical clustering with a threshold of 0.90. We found 1,305 clusters with 2 or more items; the largest cluster had 213 items. We hired an Indian journalist with experience fact-checking during the Indian 2019 elections to annotate each of the 559 clusters with five or more items by hand. The annotation interface presented three examples from each cluster: one with the lowest average distance to all other messages in the cluster, one with the highest distance, and one message chosen randomly. In 137 cases the examples shown for annotation were from multiple languages, and in 132 of those cases the journalist was able to identify the same claims across multiple languages. Although preliminary, this demonstrates the feasibility and importance of multilingual claim matching with these methods—an area we hope further work will tackle.

Our findings are supporting over 12 fact-checking organizations running misinformation tiplines. The deployed system uses I-XLM-R and automatically groups text messages with similarities over 0.95 and recommends possible matches from less-similar candidates that fact-checking organizations can confirm or reject. Matches can also be added manually. Initial feedback from the fact-checkers has been positive, and we are collecting data for further research and evaluation.

We prioritized the well-being of annotators and the privacy of WhatsApp users throughout this research. Our data release conforms to the FAIR principles [124]. We have no identifying information about WhatsApp users and any references to personally identifiable information in messages such as phone numbers, emails, addresses and license plate numbers are removed to preserve user privacy. We worked closely with our annotators preparing them for the risk of hateful content, encouraging frequent breaks, and paying well-above minimum wage. We took a compassionate response to COVID disruptions and other life stresses even when this meant less annotated data than was originally envisioned.

In this chapter we focused on grouping claims together to facilitate and scale human fact-checking. Our embeddings that beat state-of-the-art multilingual embeddings in a variety of high and low resource and non-English languages enable fact-checkers from around the world to address misinformation during critical news and event cycles such as elections in Brazil, the Philippines, and France. In the next chapter, we continue addressing RQ2 in the same trajectory, but through the lens of searching through existing fact-checks (long documents) for a social media post in different languages, as well as cross-lingual settings, and use NLP to further increase the reach of human fact-checking.

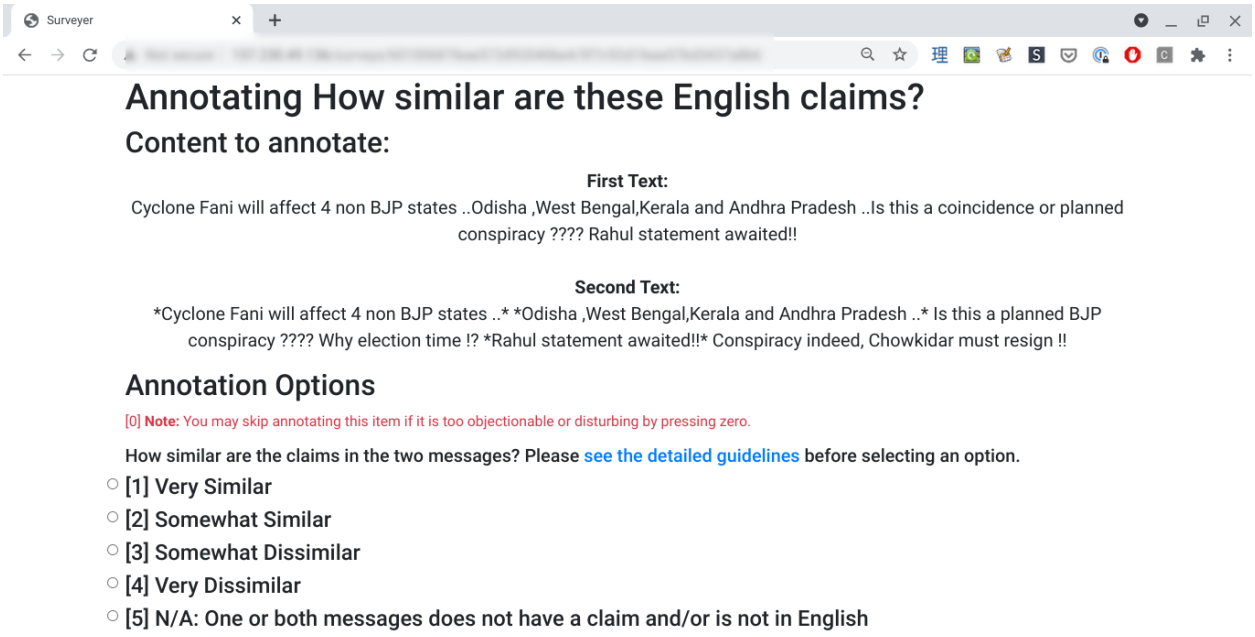

Figure 4.3: An example of the annotation interface

4.8 Supplemental Materials

4.8.1 Codebooks

Our codebooks are available openly. Due to the page limit for the supplemental materials, we provide hyperlinks to these codebooks:

- Claim detection codebook
- Claim similarity codebook

We coded a simple annotation interface, which is free and open-source: `https://github.com/meedan/surveyer/`. A screen capture of the annotation interface during the English-language claim-similarity task is shown in Figure 4.3

4.8.2 Per language results

Figure 4.4 shows the accuracy, precision, recall, and F1 scores for simple threshold classifiers. This is equivalent to Figure 4.2, but shows the plots for each language individually in addition to the overall values across all languages.
The figure also includes two additional embedding models from the SBERT website: *xlm-r-distilroberta-base-paraphrase-v1* and *xlm-r-bert-base-nli-stsb-mean-tokens*.[7] As discussed in the

[7] `https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models`

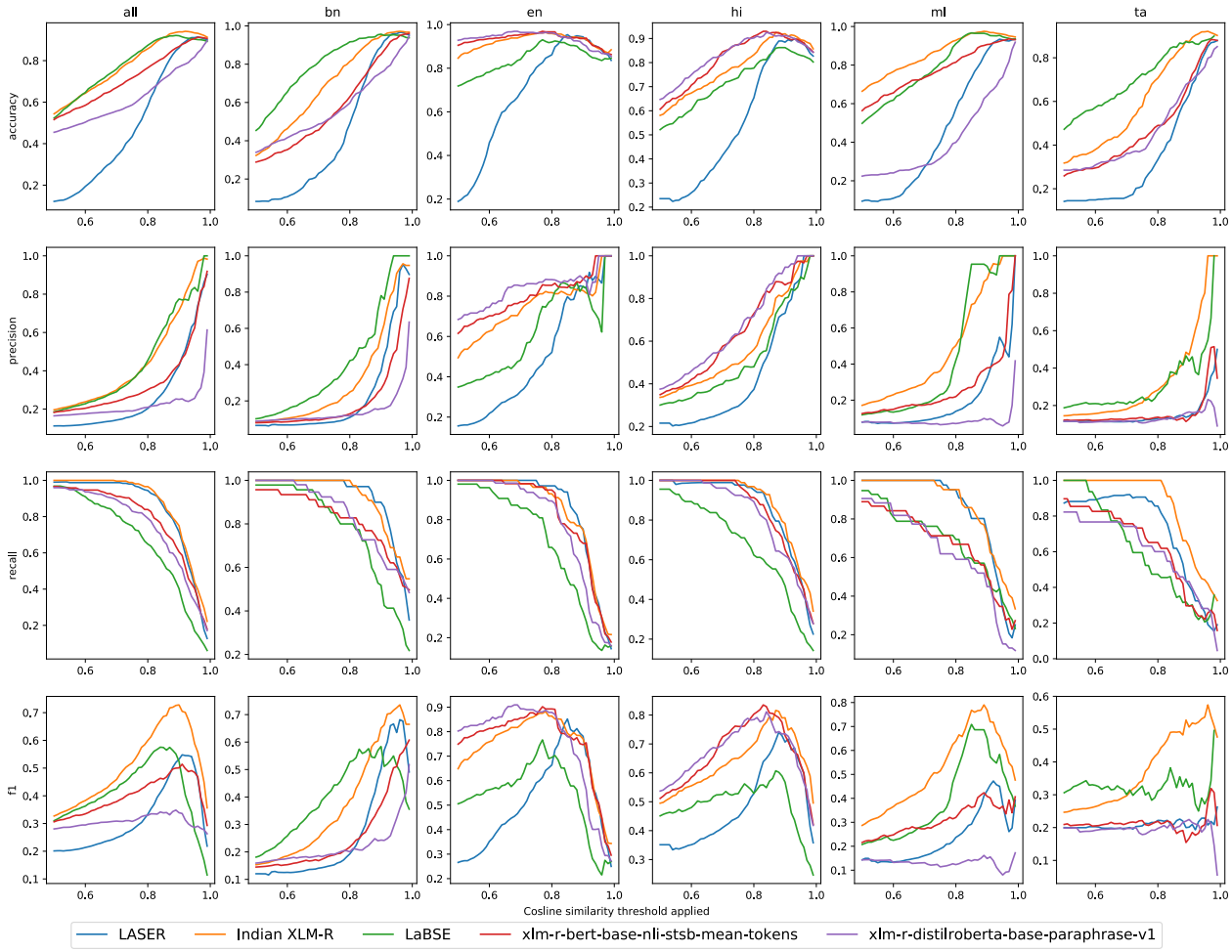

Figure 4.4: Accuracy, precision, recall, and F1 scores for each language individually. Positive class is "Very similar."

main text, we find our models far outperform these models for Bengali, Malayalam, and Tamil while performance for English and Hindi is similar.

4.8.3 Alternative definition of the positive class

The analysis in the chapter presents results for "Very Similar" compared to all other classes (N/A labels excluded). Here we show qualitatively similar results are obtained when the positive class is items for which a majority of annotators indicated "Very Similar" or "Somewhat Similar." As stated, somewhat similar matches are useful as a fact-check would partially address some of the claims in a somewhat similar match. Table 4.8 provides the distribution of labels for the claim matching dataset.

Table 4.7 presents F1 scores averaged across 10 runs of 10-fold cross validation using "Somewhat Similar" or "Very Similar" as the positive class. The results are similar to Table 4.5 in the

main text. F1 scores are generally higher, but our Indian XLM-R model still performs best. Surprisingly, LASER matches its performance in one language (Hindi).

Table 4.7: Maximum F1 scores (F1) and standard deviations achieved and the corresponding thresholds (thres.) for each score. The 'classifiers' are simple thresholds on the cosine similarities. Scores are the average of 10 rounds of 10-fold cross validation. The positive class is "Somewhat Similar" or "Very Similar."

	LASER	LaBSE	I-XLM-R
Language	F1 (thres.)	F1 (thres.)	F1 (thres.)
All	0.63±0.05 (0.88)	0.60±0.05 (0.82)	**0.76**±0.05 (0.82)
Bengali	0.63±0.09 (0.87)	0.65±0.11 (0.72)	**0.67**±0.12 (0.79)
English	0.90±0.09 (0.85)	0.81±0.12 (0.77)	**0.95**±0.08 (0.78)
Hindi	**0.82**±0.09 (0.88)	0.64±0.11 (0.77)	**0.82**±0.09 (0.82)
Malayalam	0.52±0.21 (0.92)	0.62±0.17 (0.85)	**0.76**±0.16 (0.85)
Tamil	0.42±0.16 (0.89)	0.54±0.18 (0.84)	**0.68**±0.13 (0.82)

Table 4.8: Label distribution for the claim matching dataset: VS is very similar, SS is somewhat similar, SD is somewhat dissimilar and VD is very dissimilar. NM refers to "no majority" meaning there wasn't consensus among annotators.

	VS		SS		SD		VD		NM	
Language	#	(%)	#	(%)	#	(%)	#	(%)	#	(%)
All	261	(11%)	121	(5%)	115	(5%)	1,417	(61%)	429	(18%)
Bengali	38	(6%)	62	(10%)	26	(4%)	225	(35%)	293	(45%)
English	64	(16%)	10	(3%)	21	(5%)	300	(75%)	3	(1%)
Hindi	84	(21%)	29	(7%)	10	(3%)	259	(65%)	17	(4%)
Malayalam	42	(7%)	9	(2%)	51	(8%)	474	(78%)	28	(5%)
Tamil	33	(11%)	11	(4%)	7	(2)	159	(53%)	88	(30%)

CHAPTER 5

Finding Fact-Checks for Social Media Posts

An important challenge for news fact-checking is the effective dissemination of existing fact-checks. This in turn brings the need for reliable methods to detect previously fact-checked claims. In this chapter, we focus on automatically finding existing fact-checks for claims made in social media posts (tweets). We conduct both classification and retrieval experiments, in monolingual (English only), multilingual (Spanish, Portuguese), and cross-lingual (Hindi-English) settings using multilingual transformer models such as XLM-RoBERTa and multilingual embeddings such as LaBSE and SBERT. We present promising results for "match" classification (86% average accuracy) in four language pairs. We also find that a BM25 baseline outperforms or is on par with state-of-the-art multilingual embedding models for the retrieval task during our monolingual experiments. We highlight and discuss NLP challenges while addressing this problem in different languages, and we introduce a novel curated dataset of fact-checks and corresponding tweets for future research.

5.1 Introduction

Fact-checking is an essential part of content moderation pipelines, since it provides ground truth for veracity judgements of a given claim. However, manual fact-checking is slow and expensive, as it requires human expertise.

This demand has been already identified by a recent survey study [32], showing that fact-checkers from 24 organizations in 50 countries expressed the need for reliable methods to detect previously fact-checked claims. Recent work in natural language processing (NLP) has focused mainly on the development of automatic systems to identify misinforming claims both in monolingual [64] and multilingual settings [42]. However, this research is mostly limited to short claims and does not directly assist the dissemination of existing fact-checks which are usually article-length documents. Moreover, existing work has addressed mainly claims in English with few exceptions such as work by Kazemi et al. [42] and the CheckThat! Lab [60], an evaluation lab that included a claim retrieval challenge in English and Arabic in their 2021 edition.

To contribute to this research direction, this chapter addresses the problem of matching and finding applicable fact-checks to social media posts. We approach this problem using two strategies (i) fact-check "matching", i.e., determining whether a social media post (tweet) and a fact-check pair match or not, and (ii) fact-check "retrieval", i.e., given a social media post (tweet), rank and return the most relevant fact-checks discussing the claims made in it. We address the "matching" task by building a binary classifier on top of XLM-RoBERTa (XLM-R), a large transformer-based multilingual language model [125]. For the "retrieval" task, we build an embedding similarity search system using sentence embeddings from LaBSE [109], SBERT [105] models and pairwise cosine similarity. We also analyze these tasks for languages other than English, i.e., Spanish and Portuguese. Further, we investigate a cross-lingual scenario where we seek to identify applicable English fact-checks to Hindi tweets.

5.2 Related Work

While fully automatic fake news detection and fact-checking systems [26, 126] remain an active research topic within the NLP community, there have been new research fronts in the fight against misinformation, including claim matching [64, 42], check-worthiness detection [127, 89], explanations [128, 129, 130], and detecting out of context misinformation [131, 132].
On the context of claim matching, Shaar et al. [64] introduced a retrieval-based version of the task where, for a given input claim, the goal is to rank similar check-worthy claims based on their relevance to the input claim. For this task, they focus on political related claims in English and a presented a rank model that relied on BERT [104] and BM25 based architectures. More recently, Kazemi et al. [42] focused on matching claims that can be served with one fact-check in five low and high-resource languages. Similarly Vo and Lee [111] conducted claim matching in a multimodal setting where they find previously debunked texts and images. In addition, The CheckThat! Lab 2021 evaluation presented claim matching as a shared task for English and Arabic.
Although works such as Shaar et al. [64] have matched English tweets and fact-checks, most of prior work has mainly focused on matching claims with other similar claims that are usually short in length. In this chapter, we seek to match claims with applicable fact-check reports that are significantly longer and potentially express the claim in different ways. Additionally, we approach the claim matching problem in multilingual and cross-lingual settings and experiment with recent neural models in multilingual NLP.
Among them, we use XLM-RoBERTa [125], a powerful multilingual transformer-based language model that have achieved competitive performance on cross-lingual and multilingual benchmarks. The model is trained on more than 2TBs CommonCrawl data and supports one hundred lan-

Table 5.1: An example tweet and a matching fact-check (both in English) from our dataset. The fact-checking article is redacted and can be found at this URL.

Tweet #1: Heartbreaking to see a barrage of fake WhatsApp forwards, kooky safety instructions, hysteria that **dams are breaking**, transformers are submerged and **electricity is being cut off**, hindering rescue efforts in Kerala.
Tweet #2: Heard a news that a **shutter of Cheerakuzhy dam, Thrissur broke.** Can someone pls confirm this news. Yet to find any reference in MSM. If true, pls inform authorities immediately. #KeralaFloods2018
Fact-Check Report: The Kerala government already on the back foot trying to battle a massive crisis due to relentless rain and flooding over the past week now have one more big worry - fake news led by incorrect reporting and rumours. The floods have already resulted in the deaths of 324 people in the past 17 days with thousands of people stranded across the state on rooftops and relief camps. ... [redacted] Worst of all is an audio clip in which a person is heard saying that the Mullaperiyar **dam has developed cracks** and in the next three hours, the downstream districts of Idukki, Ernakulam, Thrissur and Allapuzha will be washed away. He urges people to take it seriously as the government is hiding the information about the leak and that he got to know of it from a friend who works in Modi's office. (We have not uploaded the audio clip in the story to prevent further panic). ... [redacted] Yet another audio message was going around claiming that the **shutters of Cheerakuzhi dam** built across Gayatri river (also know as Bharathpuzha) in Thrissur **are damaged.** However, regional media Manorama News and Deshabhimani clarified that it is not true and this exaggerated message is meant to create a scare. ... [redacted] Another message which has created quite a scare is that the **whole state will have no electricity** tomorrow as the Kerala State Electricity Board (KSEB) will shut down its operations. [redacted] However, KSEB and Kerala Police were quick to respond and call the message fake. KSEB clarified through a Facebook post that KSEB employees are engaged in relentless efforts to restore electricity in the areas facing power cuts. To avoid danger during floods, power supply and production in certain parts have been temporarily discontinued. However, as the water recedes the power supply will be restored. The electricity board also appealed to people to not spread these rumours. ... [redacted]

Table 5.2: Per language statistics of our (tweet, fact-check) dataset.

Tweet Language	Article Language	# of Pairs
English	English	4,850
Hindi	English	664
Spanish	Spanish	617
Portuguese	Portuguese	402

guages. We also rely on recent language agnostic embedding models such as LaBSE (Language-agnostic BERT Sentence Embedding) [109], a sentence embedding model that can produce embeddings in 109 languages. This model was built using a combined pretraining method of masked and translation language modeling trained on 17 billion monolingual sentences from CommonCrawl and 6 billion translated pairs of sentences. Sentence-BERT (SBERT) [105] use twin and triplet networks on top of language models for producing sentence embeddings. In their follow up work to SBERT [118], they also propose an approach to convert monolingual embeddings into multilingual ones. We also use Elasticsearch's implementation of the BM25 retrieval system [120], which provides fast and scalable text search.

5.3 Data

Our data is derived from 150,000 fact-checks obtained from several sources, including (i) fact-checking organizations certified by the International Fact-Checking Network (IFCN) and (ii) fact-checking aggregators such as Google Fact-check Explorer,[1] GESIS [113], and Data Commons.[2] The collected fact-checks cover several languages, including English, Spanish, Portuguese, and Hindi. Each fact-check includes a claim and usually a justification article for the claim verdict, and metadata such as publication date, claim veracity and references to the original content that needed the fact-check.

Similar to Shahi [47] and Shahi et al. [48], we use social media links included in the fact-checks and their original news sources (whenever available) to build a dataset consisting of (tweet, fact-check) pairs. Given that the fact-checks include several languages, we obtain monolingual pairs in English, Spanish, and Portuguese and also cross-lingual pairs consisting of Hindi tweets and English fact-checks. In cases where the tweet contains a link (usually to a news article), we also append the preview text from the link to the tweet text, to capture more of the tweet's context. Since we match tweets and fact-checks automatically through references in the text we conducted an additional verification step to make sure that the identified pairs are indeed related. We thus

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

[2] https://datacommons.org/factcheck/faq

annotate a random sample of 100 English (tweet, fact-check) pairs to verify whether each fact-check is applicable to its matched tweet. The annotation was conducted independently by two annotators, reaching an 87% agreement between annotator responses. We find that 89% of the tweets in our sample matched their corresponding fact-checks and in most cases the pairs include at least one fact-check worthy claim. This finding suggests that while there is some degree of noise in the pairing process, most of the pairs are correct matches. Table 5.2 shows a summary of the final set of (fact-check,tweet) pairs per language. Sample (fact-check,tweet) pairs are shown in Table 5.1. Note that multiple tweets can be matched to the same fact-check.

5.4 Models & Baselines

5.4.1 Matching (tweet, fact-check) Pairs

We address the task of matching (tweet, fact-check) pairs as a binary classification problem using "match" or "not match" as possible labels.

Our dataset consists of only positive labels since we only collected matching (tweet, fact-check) pairs, and training a binary classifier also requires negative examples, so we explored several strategies to obtain negative samples. Initially, we selected negative examples by randomly pairing non-matching tweets and fact-checks. We then built a binary classifier using an XLM-R model fine-tuned on the resulting dataset. However, preliminary evaluations showed that the resulting classifier was not able to generalize well. We believe this is due to the classifier's lack of exposure to challenging negative samples, since most of random pairings are easily distinguished from matching (tweet, fact-check) pairs.

In order to get more challenging negative samples, we opted for finding non-matching (tweet, fact-check) pairs based on their pairwise similarity. We start by calculating the pair-wise cosine similarity across all possible (tweet, fact-check) pairs in the dataset, within the same multi/cross-lingual setting. Then, we use LaBSE embeddings [109] of tweets and fact-check articles and rank non-matching pairs by decreasing cosine similarities. Next, we pick the top negative samples from this set, i.e., pairs with similarities lower than 0.7, to reduce the number of false negatives. We train our XLM-R classifier with the resulting data and find an 15% absolute improvement of classification accuracy as compared to training on randomly selected pairs.

Since our dataset contains multiple languages, we conduct an additional set of experiments where we train separate classifiers for each language pair e.g., English, Spanish, Portuguese, Hindi-English, as well as a classifier that uses pairs in all languages. Results are presented in Table 5.3.

Table 5.3: Results from matching (tweet, fact-check) pairs as a binary classification problem. F1+ and F1- refer to the F1 score for the "match" and "not match" classes.

	Trained Separately			Trained Altogether		
Lang. Pairs	Acc.	F1+	F1-	Acc.	F1+	F1-
En-En	88.46%	88.72%	88.17%	88.61%	88.66%	88.54%
Hi-En	80.27%	80.53%	79.71%	80.50%	81.60%	78.90%
Es-Es	85.82%	86.07%	85.09%	88.57%	88.93%	88.06%
Pt-Pt	84.08%	83.67%	83.59%	87.44%	87.65%	87.25%

5.4.2 Finding Applicable Fact-Checks for Tweets

A different perspective on the problem of matching fact-checks with tweets is to retrieve and rank fact-checks based on their relevance to an input tweet. As opposed to binary classification, this approach provides a ranked list of options to choose from and requires human intervention to select the most appropriate fact-check. This strategy makes the search process more scalable since finding applicable fact-checks does not require the quadratic number of computationally expensive comparisons that make the binary classification approach computationally intractable for retrieval.

During our experiments we use BM25 as our baseline retrieval method. We use the implementation provided in Elasticsearch [120] . BM25 is inherently language agnostic since it relies on token matching. However, this makes it unable to handle cross-lingual text, which is the case of our set of (Hindi tweets, English fact-checks). To address this issue, we translate the Hindi tweets into English using Google translate before using BM25. Our preliminary experiments show that the use of translated tweets leads to a stronger baseline as compared to just applying BM25 to the original Hindi tweets. The translation is only to accommodate for the lack of cross-lingual operability of BM25 and is necessary for keeping consistent with our comparison methodology.

Additionally, we experiment with multilingual sentence embeddings, namely LaBSE and (multilingual) MPNet-SBERT. Since these embedding models only support inputs up to 512 tokens and fact-check articles are usually longer, we embed article paragraphs instead of whole articles. Thus, we compare an input tweet with paragraphs from the fact-check reports and not with full-length articles. Note that unlike the embedding-based models, BM25 is able to handle text in arbitrary length, so in order to carry out a fair comparison of the baseline and embeddings, we additionally provide a BM25 baseline using article paragraphs only.

The results for these experiments are presented in tables 5.4, 5.5 and 5.6. We discuss them in detail in the following sections.

5.4.3 Experimental Setup

We use HuggingFace's *transformers* [88] and the SBERT library to implement our models. We run our code on a GPU-enabled server. For the English retrieval experiments, we use LaBSE and the *paraphrase-mpnet-base-v2* model which we call "MPNet-SBERT", an SBERT embedding model trained on top of MPNet [133]. We use the multilingual version of the same model (*paraphrase-multilingual-mpnet-base-v2*) for Spanish, Portuguese and Hindi-English pairs.
To evaluate classification tasks, we use accuracy and F1 score as our main metrics. The classification experiments are conducted using 5-fold cross validation. For our retrieval experiments, we use "mean reciprocal rank" (MRR) and "mean average precision" (MAP@K) for different values of K.

5.5 Experiments in English

Results in Table 5.3 show a promising performance from our XLM-R models in matching tweet-fact check pairs in English, with accuracies of up to 89%. As observed, there is a slight performance increase when training the model with all languages as compared to using English only. While the increase in performance when training altogether is more significant for other languages, it is worth noting that the English performance remains robust to noise as using training data from other languages can introduce noise for a model applied on English only. Although there is a slight performance decrease in the "match" class, the performance gain for the "not match" class when using the training altogether model is large enough to improve the overall accuracy, which suggests potential benefits from using data in other languages.
Table 5.4 presents results for the retrieval-based evaluation. The full-length BM25 baseline achieves 65% MAP@1 and 72% MRR scores as the best performing model. The gap between MAP numbers mostly decreases as K increases which is an expected behavior for mean average precision. At first glance, it seems that feeding paragraphs from the article to the embedding models could account for the performance loss, since the full-length BM25 uses the whole document at once, therefore providing the upper bound performance for this task.While the paragraph BM25 system has a decrease in performance relative to full-length BM25, not all of the performance gap between embedding models and BM25 can be explained by the inability of embedding models to process longer documents. Among the embedding based models, MPNet-SBERT is the best performing model achieving 54% MAP@1 and 62% MRR. The second best performing model, LaBSE, is behind MPNet-SBERT by a noticeable margin of more than 8 MAP@1 and MRR points. A potential explanation for this performance decrease is that LaBSE is a multilingual model and performance decrease with respect to single-language models is often observed when a model supports multiple languages (100+ in LaBSE's case) at once.

Even though we see promising classification results, our experiments show that state-of-the-art NLP algorithms are still unable to compete against the BM25 baseline in finding applicable fact-checks.

Table 5.4: Results from retrieval experiments in English.

	MAP@K					MRR
Model	K=1	K=5	K=10	K=20	K=50	
Full-Length BM25	64.85%	70.82%	71.18%	71.30%	71.39%	71.51%
Paragraph BM25	62.03%	66.68%	67.27%	67.46%	67.57%	67.61%
LaBSE	44.98%	51.59%	52.24%	52.64%	52.81%	53.00%
MPNet-SBERT	53.56%	60.58%	61.20%	61.48%	61.68%	61.84%

5.6 Experiments in Other Languages

Since our dataset also covers Spanish and Portuguese, we conduct an additional set of experiments to assess the performance of our models in languages other than English. During these experiments, we test the same models used with English, with the exception of English MPNet-SBERT that was replaced with the multilingual version.

The results in Table 5.3 indicate that training a single XLM-R model on data from all languages performs more accurately on average (86.28%) in comparison with training separate models per language (84.66%) for matching. Particularly, we see a performance increase for Spanish and Portuguese, with accuracies of up to 88.57% and 87.44% respectively. Training a single XLM-R model on all languages leads to a performance improvement up to 3.36% for Spanish and Portuguese as compared to the single-language models, implying the transfer of task expertise across languages for XLM-R. A potential explanation of the fact that a single model has the leverage of larger data. We believe this is particularly effective when the languages are similar and can learn from each other's data. Also note that classifying (tweet, fact-check) pairs in multiple languages with a single XLM-R model is preferred since it saves computational resources and is easier to use.

We observe mostly similar trends to the English experiements for fact-check retrieval as shown in Table 5.5, with two exceptions: (i) multilingual MPNet-SBERT slightly outperforming the paragraph BM25 model by 1.19 MAP@1 and 2.55 MRR points in Spanish and (ii) LaBSE outperforming multilingual MPNet-SBERT by 5 MAP@1 and 2 MRR@1 points for Portuguese. Note that during these experiments, the embedding models mostly underperformed in comparison to both BM25 baselines, with the full-length BM25 outperforming the best embedding model by 14 MAP@1 and 11 MRR points in Spanish in comparison with MPNet-SBERT and 10

MAP@1 and MRR points in Portuguese in comparison with LaBSE.

Table 5.5: Results from retrieval experiments in Spanish and Portuguese. ML in "ML MPNet-SBERT" is short for multilingual.

Model	**MAP@K**					**MRR**
Spanish	K=1	K=5	K=10	K=20	K=50	
Full-Length BM25	73.41%	78.56%	78.78%	78.84%	78.90%	78.54%
Paragraph BM25	58.33%	63.65%	64.38%	64.78%	64.88%	64.56%
LaBSE	57.14%	62.83%	63.68%	64.01%	64.24%	64.68%
ML MPNet-SBERT	59.52%	66.28%	66.58%	66.74%	66.90%	67.23%
Portuguese						
Full-Length BM25	69.62%	74.09%	74.73%	74.99%	75.04%	75.04%
Paragraph BM25	69.62%	72.51%	72.97%	73.23%	73.32%	73.32%
LaBSE	59.49%	62.95%	63.25%	63.53%	63.89%	63.89%
ML MPNet-SBERT	54.43%	60.06%	61.07%	61.29%	61.55%	61.55%

5.7 Cross-Language Experiments

The retrieval results are presented in Table 5.6. Unlike the monolingual experiments, models from the previous section outperform BM25 by noticeable margins in the retrieval setting and perform competitively with other language pairs in classification too. The only difference is that for the cross-lingual Hindi-English pairs in retrieval, the tweets are first translated into English. Also, the single XLM-R model trained on data from all language pairs classifies (Hindi tweet, English fact-check) pairs comparably with the monolingual models with 80.57% accuracy according to Table 5.3. Although there is a 5.8% accuracy decrease compared to the best mean accuracy (altogether), the Hindi-English XLM-R performance is still considered competitive for the more difficult task of cross-lingual matching.

Furthermore, we observe high cross-lingual performance from LaBSE, better than its performance on English and close with Portuguese and Spanish. LaBSE outperforms the best BM25 system (full-length BM25) by about 7.5 MAP@1 and 7 MRR points. However, there is a large performance gap between the embedding models (12% MAP@1, 9.5% MRR) as multilingual MPNet-SBERT has the worst performance of all systems but still performs not too far worse than the BM25 baselines. The improvement of LaBSE over ML MPNet-SBERT can be attributed to the fact that LaBSE was trained specifically for cross-lingual representation learning. We believe that BM25's underperformance can be attributed to translation errors. However, this is one of the few ways (other than translating the fact-checks) that BM25 can support cross-lingual queries,

ultimately making this a downside of using BM25. Overall, the use of XLM-R and LaBSE for cross-lingual matching and retrieval of fact-checks is a promising direction.

Table 5.6: Results from cross-lingual retrieval experiments with tweets in Hindi and fact-check articles in English. For BM25 systems, the tweet is first translated into English before being fed to BM25.

	MAP@K					MRR
Model	K=1	K=5	K=10	K=20	K=50	
Full-Length BM25	47.95%	52.59%	52.99%	53.11%	53.15%	52.80%
Paragraph BM25	45.89%	50.31%	50.78%	50.98%	51.02%	50.48%
LaBSE	55.48%	59.12%	59.63%	59.92%	60.14%	59.79%
ML MPNet-SBERT	43.15%	48.72%	49.17%	49.68%	49.87%	50.22%

5.8 Discussion and Future Work

Our experiments show promising performance from XLM-R in the matching classification of tweets and fact-check pairs, with the single XLM-R model trained on all data performing on average 86.28% accurately. While the binary XLM-R classifier performs reasonably well on full-length articles, we found that it does not perform as well in classifying (tweet, fact-check paragraph) pairs when we used it to refine retrieval results. Reranking classifiers have shown promising results in prior work [134], however they were not particularly applicable in our case since we do not have paragraph-level labels for the fact-check articles to train a classifier that can rerank paragraphs.

Both BM25 baselines outperform or perform competitively with state-of-the-art multilingual sentence embedding models in monolingual retrieval settings. However, there is a key difference between BM25 and the embedding models: BM25 can handle articles of arbitrary length, whereas both LaBSE and MPNet-SBERT can handle only up to 512 tokens of input. This is a source of performance loss for LaBSE and SBERT in our task. In future work, we plan to explore long document transformers such as Longformer [135] to address this problem. We believe that a long document multilingual embedding model can provide improvements not only to our research problem, but to many other areas such as news NLP and legal document processing in multilingual settings.

It is important to note that the input length limit does not explain all of the performance gap. We found that the paragraph BM25 system still outperforms the embedding-based systems by at least 10 MAP@1 and 9.5 MRR points in Portuguese experiments and performs similarly to LaBSE and MPNet SBERT in Spanish. While specialized embeddings like question answering embedding models exist for English through SBERT, they are neither necessarily applicable

in searching through fact-checks nor easy to come by in non-English, multilingual and cross-lingual capabilities. Building specialized embedding models for searching through applicable fact-checks is a promising next step in improving the embedding-based retrieval systems.
LaBSE provides impressive results on cross-lingual (Hindi tweet, English fact-check) pairs, outperforming BM25 baselines and MPNet-SBERT as a single multilingual embedding model with support for more than one hundred languages. This is an important problem to solve, since misinformation travels across borders and being able to search through fact-checks across different languages can save a great deal of manual fact-checking efforts. Therefore, cross-lingual search of applicable fact-checks for social media posts has a great potential for extending the reach of manual fact-checking. Furthermore, since LaBSE performs better or comparable with multilingual MPNet-SBERT overall, this makes it the better choice for embedding models when searching for relevant fact-checks on non-English social media.

5.9 Conclusion

In this chapter, we approached a new version of the "claim matching" problem in which we match applicable fact-checks with social media posts to increase the reach of manual fact-checking. We addressed this problem using classification and retrieval based strategies in multiple languages (English, Hindi, Portuguese and Spanish).
Our results showed promising performance as we are able to classify matching (tweet, fact-check) pairs with accuracies of up to 89% in four language pairs. From our retrieval experiments we found that monolingual pairs of (tweet, fact-check) are better retrieved by BM25, which meaningfully outperforms state-of-the-art multilingual embedding models in the retrieval task. Despite this, we observe promising performance in cross-lingual settings with LaBSE achieving more than 7.5 MAP@1 points improvement over the best BM25 baseline.
We identified the monolingual retrieval of applicable fact-checks as a challenging area for state-of-the-art NLP and highlighted the need for specialized and long document multilingual embeddings as an important direction for future work.
Our newly curated multi/cross-lingual dataset of (tweet, fact-check) pairs in English, Spanish, Portuguese and Hindi is publicly available at http://lit.eecs.umich.edu.
In the following chapter, we will continue our efforts in building human-centered language technology for fact-checkers, to aid them in coming up with more effective queries than enable misinformation discovery across social media platforms, further increasing fact-checker's efficiency in countering misinformation at scale.

CHAPTER 6

Query Rewriting for Effective Misinformation Discovery

We propose a novel system to help fact-checkers formulate search queries for known misinformation claims and effectively search across multiple social media platforms. We introduce an adaptable rewriting strategy, where editing actions for queries containing claims (e.g., swap a word with its synonym; change verb tense into present simple) are automatically learned through offline reinforcement learning. Our model uses a decision transformer to learn a sequence of editing actions that maximizes query retrieval metrics such as mean average precision. We conduct a series of experiments showing that our query rewriting system achieves a relative increase in the effectiveness of the queries of up to 42%, while producing editing action sequences that are human interpretable.

6.1 Introduction

With the wide spread of both human and automatically generated misinformation, there is an increasing need for tools that assist fact-checkers while retrieving relevant evidence to fact-check a claim. This process often involves searching for similar claims across social media using initial clues or keywords based on users' intuition. However, the available mechanisms for search on social media sites are often platform-specific, with restrictions on the allowed number of search queries and access to retrieved documents. This can be attributed, among others, to the dynamic nature of social media feeds, the differences among users' interactions, and the architectural differences in how platforms perform search on their data. As a result, optimizing for arbitrary black-box search end-points containing ever-changing and different document sets means that a generic claim rewriter operating across all search end-points has a high chance of being sub-optimal.

To address these challenges, we draw upon a direct collaboration among fact-checkers and NLP researchers, and introduce an adaptive claim rewriting system that can be used for effective misinformation discovery. We develop an interface in which users can edit individual tokens in the input claim using a predefined set of actions, and obtain updated queries leading to different lev-

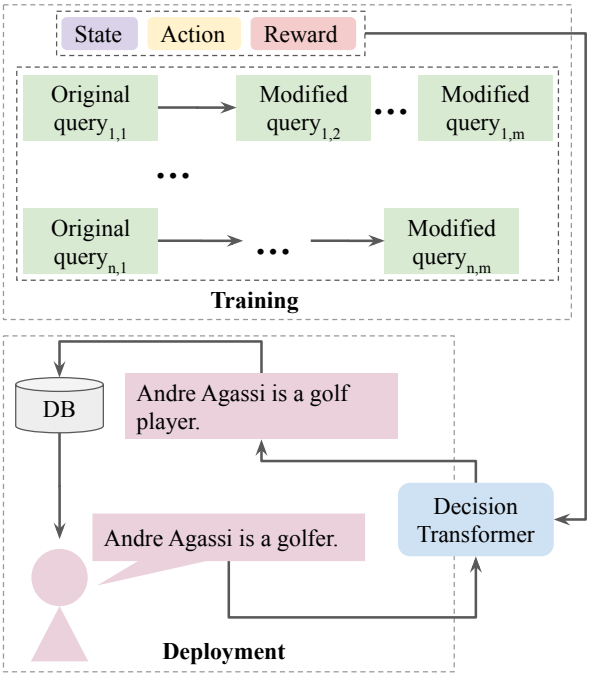

Figure 6.1: Overview of our proposed approach: we train a decision transformer with "state", "action" and "reward" sequences discovered by searching the space of potential query edits. During the deployment stage, the decision transformer predicts action(s) to rewrite the claim into a more effective query.

els of retrieval performance. Using this environment, we build a system that learns to rewrite input claims as effective queries by leveraging reinforcement learning (RL) to maximize desired retrieval metrics (e.g., average precision at K (AP@K)). An offline RL agent is then trained to learn the best editing sequences using a decision transformer model [136] as shown in Figure 6.1. Given the limited access to social media search APIs, we use off-the-shelf retrievers such as BM25 [120] and approximate K-nearest neighbours (kNN) [137] to simulate platform search endpoints. Our system is trained using a modified version of FEVER [31], a well known misinformation dataset containing a mix of true and false claims linked to Wikipedia evidence sentences. We transform FEVER claims into sequences of (claim, edit action, reward) triplets by using Breadth First Search (BFS) and heuristics such as constraining search space depth. These triplets are used to train a decision transformer model to autoregressively predict a sequence of editing actions leading to retrieval improvements.

Through several experiments, we show that our query rewriting approach leads to relative performance improvements of up to 42% when compared to using the original claim. We also find that a simplified version of this approach— i.e., fine-tuning a classifier to predict a single edit, leads to comparable performance while being more resource efficient during training and inference. We conduct ablation experiments to further evaluate the model performance across several settings,

including variations on the retriever type, the reward metric, and the presence of negative training examples.
To the best of our knowledge, our system is the first to leverage RL to learn to edit text from a set of human-readable actions only. From a practical perspective, it provides initial experimental evidence on the potential of interpretable systems in helping users, including fact-checkers, media writers, and platform trust and safety teams, to more effectively discover misinformation on the Internet.

6.2 Related Work

Our work is closely related to three previous research directions.

Finding Similar Claims. The problem of finding similar claims has been explored from the perspective of system building, and supports a key step in human-led fact-checking [32]. Shaar et al. [64] conducted retrieval and ranking of previously fact-checked claims given an input claim to detect debunked misinformation in English. Kazemi et al. [42] tackled a similar problem in non-English languages. Kazemi et al. [138] investigated systems and models for finding applicable fact-checks for tweets.
While most prior work on this area has focused on building retrieval systems to identify similar claims, our work focuses on query rewriting to assist fact-checkers in the discovery of misinformation. During this process we assume that the retrieval system is a black-box to which we only have search access.

Query Rewriting. Query reformulation methods such as relevance feedback and local or global query expansion have been well-studied within the information retrieval literature. Lavrenko and Croft [139] proposed the *relevance model*, an unsupervised local expansion method in which the probability of adding a term to the query is proportional to the probability of the term being generated from language models of the original query and the document the term appears in. Cao et al. [140] proposed a supervised pseudo relevance feedback in which expansion terms are selected by a classifier that determines their usefulness to the query performance. Li et al. [141] introduced REC-REQ, an iterative double-loop relevance feedback process in which a user provides relevance feedback to a classifier that is trained to identify relevant documents.
RL approaches have been previously applied to query rewriting. Nogueira and Cho [142] and Narasimhan et al. [143] used RL to learn to pick terms from pseudo-relevant documents that upon addition to the query improve retrieval performance metrics such as recall. In more recent work, Wu et al. [144] proposed CONQRR, a system that rewrites conversational queries into standalone questions. The authors first trained a T5 model to generate human rewritten queries for the

QReCC dataset [145] and then used them to generate candidate queries, which are selected based on maximizing search utility by an RL agent.
A key difference between our method and prior work is that we do not use information from the retrieved documents to reformulate queries as the queries themselves are the only input to the model.

Text Editing Models. Also related to our work is research done on "text-editing" models [146]. This line of research has gained traction in recent years as models such as EdiT5 and LEWIS [147, 148] promise hallucination-free and controlled text generation for tasks where the input and output texts are similar enough so that a model can learn to transform the input into the output by applying a limited number of editing actions. Stahlberg and Kumar [149] proposed Seq2Edits, a fast text-editing model for text generation tasks such as grammatical error correction and text simplification. Seq2Edits uses an edited transformer encoder and decoder to generate sequences of edits for the positions in the input text that need to be altered with suggested new tokens. Reid and Zhong [148] introduced a multi-span text editing algorithm that uses Levenstein edit operations for the tasks of sentiment and politeness transfer in text, based on the intuition that text style transfer usually can be done with a few edits on the input text. Overall, text-editing models are usually faster than other sequence generation models such as seq2seq, since they only predict actions on a few input tokens rather than regenerating the whole sequence.

6.3 Methods

6.3.1 Problem Definition

In this chapter, we focus on the task of query rewriting for discovering similar claims from an opaque search end-point. We have a collection of input claims $(C_1, C_2, ..., C_n)$ that contain at least one fact-checkable claim. For any given claim C_i in the collection, there exists one or more collections of similar claims $(SC_{i1}, SC_{i2}, ..., SC_{im})$, either supporting or refuting the claim in-part or as a whole. The RL agent operates on a fixed set of actions $\mathbf{A} = \{A_1, A_2, ..., A_k\}$ that can be applied to any of C_i's tokens $(T_{i1}, T_{i2}, ..., T_{iq})$, where k is the number of possible actions, q is the number of tokens in C_i. We rewrite the query by applying the sequence of actions $(A_{ij}, 1 \leq i \leq k, 1 \leq j \leq q)$ generated by the RL model to the original query. We can then use this improved query to retrieve related evidence statements.

6.3.2 Model Overview

Our system rewrites a query using concepts from RL and query expansion. We pass the query into a pre-trained language model and then use the pooled representation from the final layer as

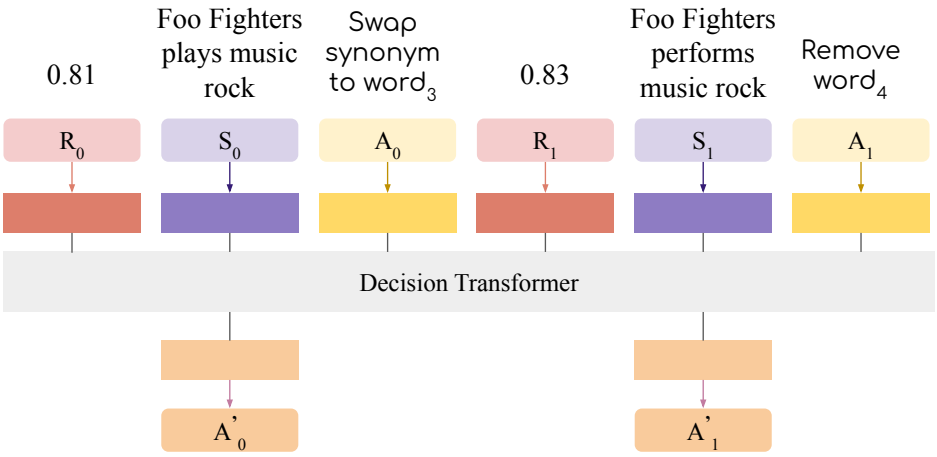

Figure 6.2: Model architecture. R, S, A represent reward, state and action, respectively. For instance, the state S_0 corresponds to a query, and the reward R_0 is the retrieval score such as AP@K. After we apply the action A_0 to the query S_0, the query becomes S_1. In inference time, the decision transformer predicts a series of actions $\{A'_0, A'_1, \cdots\}$ to apply to the original query.

the state representation. We use a decision transformer architecture, where states, actions, and rewards are provided to the model as a flattened sequence. The decision transformer uses a decoder-only GPT architecture [150] to learn the optimal policy during training time. During inference time, it autoregressively predicts actions for a given state. An overview of our model architecture is shown in Figure 6.2. Below, we describe important elements of the model architecture related to the query rewriting process.

Rewriting Actions. Queries are rewritten using the following set of actions.

(1) *Add synonym*: adds the synonym of a selected word to the query. Previous work by work by Riezler and Liu [151], Mandal et al. [152], showed that rewriting queries with synonyms can improve query performance by potentially resolving ambiguous query terms.

(2) *Swap with synonym*: replaces a specific word from the query with its synonym. This action has the same goal as *add synonym*. Note that it includes the removal of the original token *remove(original_token)* .

(3) *Change tense to present simple*: changes verb tense into present simple for selected verbs in the input. Changing verbs to their morphological variants has been previously found useful for query rewriting [153, 154].

(4) *Remove*: deletes selected words from the query. Previous work has found that deleting words in queries can lead to higher coverage of the search content [155].

We implement these actions using WordNet [156] and the spaCy's part-of-speech tagger. Note that only certain actions are permitted for each part of speech tag: verbs support all four actions, nouns, adjectives and adverbs support all actions except changing verb tense, and stop words and other parts of speech support only the remove action.

Action #	Edit, Position
action 0:	*swap with synonym, position 0*
action 1:	*swap with synonym, position 1*
...	...
action 32:	*add synonym, position 0*
action 33:	*add synonym, position 1*
...	...
action 126:	*remove, position 30*
action 127:	*remove, position 31*

Table 6.1: A 2D space of actions types and token indices mapped onto a linear action space.

State Representation. We use sentence embeddings of the input claim as its state representation. An input claim C_i is passed through a Sentence-BERT [105] network . The weights of the underlying pretrained language model (LM) are fine-tuned together with the decision transformer.
Action Representation. Our action space is two-dimensional: the first dimension represents the four action types (*add synonym*, *swap with synonym*, *change tense to present simple* and *remove*) and the second dimension represents the position of the token under edit, up to a maximum of 32 tokens. We pack these dimensions into a single dimension by taking their product, as shown in Table 6.1. Similar to the original implementation of the decision transformer, we pass the actions through a learned embedding layer to obtain an action vector representation.
Rewards. We use the retrieval score for the edited query as the system reward at time step t. Since the decision transformer uses returns-to-go to inform the model about future rewards, we use the sum of future rewards as a returns-to-go $R_t = \sum_{t'=t}^{T} r_{t'}$. We also experimented with a delayed reward strategy, where we set the returns-to-go for the last time step to be the maximum score for given claim seen during the data generation process, and zero for intermediate steps. During inference, we initialize returns-to-go to the maximum reward and decrease it by the achieved score after we apply an action.

6.3.3 Retriever

Since access to social media API search endpoints is limited, it is difficult to train an RL agent on top of them. Furthermore, the changing nature of misinformation on social media is another important factor to take into account, given that misinformative posts are periodically removed from social media platforms and are thus no longer available once fact-checked. These issues made us opt for a simulated search environment, with the added benefit of making our methods adaptable to arbitrary search endpoints. We experiment with two main systems:
BM25. A retriever frequently used in the literature as a retrieval baseline [120]. We use the Elasticsearch implementation of BM25 with the default parameters.

Approximate kNN. A kNN retriever implemented using Elasticsearch's dense vector retrieval. We encode our data using pre-trained Sentence-BERT [105] and use the embeddings to conduct an approximate kNN search using the Hierarchical Navigable Small Worlds (HNSW) algorithm [137].

6.4 Data

6.4.1 FEVER Dataset

The FEVER dataset [31] is a collection of manually written claims from Wikipedia that are connected with evidence sentences that either "support" or "refute" them. Since we are interested in claims linked to related evidence, we discard the claims in the dataset labeled as "NotEnough-Info." This leaves us with 102,292 claims in the training and 13,089 claims in the development sets. [157] identified issues caused by the construction processes of the original FEVER dataset such as uses of negation in claims being heavily correlated with the "refute" outcome, therefore causing a "claim only" fact verification system to performs as well as an evidence-aware fact verification system. However, since our work is not concerned with the fact verification application of FEVER, we do not find this to be an issue.

FEVER is a well-known dataset among the misinformation and fact-checking communities. Even if FEVER is not a social media dataset, it is nonetheless based on user-contributed data, and thus we believe that the findings obtained using this dataset can be generalized to claims on social media platforms with minor domain-specific revisions, especially since the linguistic structure of claims and discussions around them is similar to the claims in the FEVER dataset.

6.4.2 Generating RL-Friendly Training Data

To generate training data, we transform FEVER pairs (claim, evidence set) into sequences of editing actions that improve upon the original query. These transformations are obtained by exploring the state space of possible outcomes after applying different permutations of edits on the initial claims. We use a Breadth-First Search (BFS) strategy that applies editing actions to an input claim C_{i0} and finds the collection of the action sequences of $(C_{ij-1}, A_j, C_{ij}, R)$ that can improve the initial claim, where C_{ij} is the generated claim after applying the edit A_j to the claim C_{ij-1}, and R is the reward of $Q(C_{ij})$ (querying retriever with C_{ij}).

Although understanding the effects of different search algorithms on our model remains an interesting problem for future work, our experiments show that using simple heuristics on BFS search is effective while generating training data from the FEVER dataset. For instance, we find that limiting the depth of the breadth-first exploration to K levels is effective for improving the query results. Also, when conducting parallel runs on different sections of the dataset, even for K = 4,

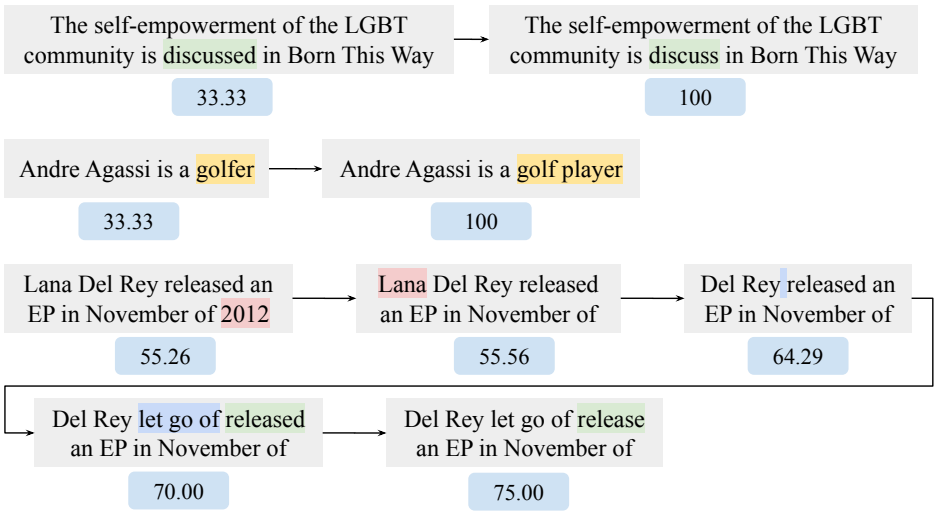

Figure 6.3: Sample sequence of claims generated by different actions: remove, change into present tense, swap synonym, add synonym highlight the token to remove, the corresponding tokens to change tense as well as to swap to its synonym, or the corresponding places to add synonym in red, green, yellow and blue, respectively. We report the corresponding AP@50 scores below each claim. Section 6.3.2 provides intuitions of why these actions lead to better scores.

the vanilla depth-limited BFS takes a half to 2 days to generate the training data. Additionally, we find that restricting the state-space search to include only improvement edits at every step reduces the size of the search space. We also prune search paths leading to minor improvements (i.e. less than 3%) or at random in 5% of instances. Since most edits do not lead to significant improvements, it is unlikely we skip meaningful paths during the search. Finally, we only include sequences with the highest gains through serial edits, e.g. picking the top 50 or 100 most beneficial editing sequences for each claim, in our training set. Overall, these heuristics improve the generation speed and quality of the training instances.

Moreover, our ability to learn good editing actions depends on how well we can generate training examples. By setting K, the maximum depth for search to 4, we are able to get improvements up to 41.21 AP@50 scores for 45,658 claims, on average, against the BM25 retriever. We also discard training examples with reward values already at maximum, since it is impossible to improve beyond the perfect score, and also edited claims leading to no improvement. Figure 6.3 shows examples of the sequences of claims generated by different actions. Table 6.2 reports the distribution of actions as well as the average improvement of AP@50 scores for each action when tested against the BM25 retriever.

	Remove	Swap_Syn	Add_Syn	Present
%	**76.64**	13.71	6.36	3.30
Δ	11.56	12.36	**12.84**	12.28

Table 6.2: Percentage (%) and mAP@50 (Δ) improvements per rewriting action against the BM25 retriever.

Model	mAP@50
Original Claim	26.83
Random Baseline	21.44
Decision Transformer$_{\text{sparse reward}}$	32.43
Decision Transformer$_{\text{dense reward}}$	**33.14**
Fine-Tuned One Action Classifier	31.95

Table 6.3: Experiment results with BM25 as retriever.

6.5 Experiments

We perform several experiments to determine the effectiveness of our adaptable query rewriting strategy. As a search environment, we use the BM25 and approximate kNN information retrieval methods described in Section 6.3.3.

6.5.1 Experiment Settings

During our experiments, we use the original decision transformer implementation.[1] We use a 6-layer decoder-only transformer with 8 heads, embedding dimension of 768. We set K (also called a block size) to be the maximum number of edits to the original query. We pad all sequences shorter than K. After flattening all the returns-to-go, states and actions, our sequence becomes of length $K * 3$. We use the *all-mpnet-base-v2* embedding model from the Huggingface's sentence transformers library.[2] We also experimented with the all-MiniLM-L12 model from the sentence transformers, but the results were worse, possibly because of all-MiniLM-L12 being a smaller model. Our intermediate state representations for an input claim is a vector of size 768. Our model is trained with cross entropy loss for 5 epochs performed on one Nvidia 2080Ti GPU.

6.5.2 Results

Results in Table 6.3 show that the decision transformer model with fine-tuned state embeddings and dense rewards outperforms all systems with BM25 as retriever and AP@50 as reward. The

[1] https://github.com/kzl/decision-transformer
[2] https://huggingface.co/sentence-transformers

same model trained with sparse rewards—hiding the intermediate rewards during training—does slightly worse than the dense reward setting, suggesting that providing more granular information about each action's reward during training brings performance advantages. Both models turn the input into a significantly more effective query with performance improvements of up to 23% (relatively) as compared to just searching for the original claim. According to Table 6.4 these gains are the highest for kNN as retriever and recall as reward. Table 6.3 also shows that performing a random sequence of edit actions negatively affects performance. This suggests that there is a "query improvement process" that needs to be learned and applying a random sequence of edits by itself does not bring any inherent advantages, i.e. our systems do well not because there is an inherent gain in how we transform the problem, since if that was true, applying random action sequences should have yielded improvements over the claim baseline, which it did not.

6.5.3 Analysis

Figure 6.4 shows the mean AP@50 (mAP@50) score changes for all the generated sequences for the experiment of decision transformer with sparse reward. We plot the mAP@50 scores for queries generated at each step, where the x-axis shows the number of edits, with 1 representing the original claim and 5 the final rewritten query. The size of the circle indicates the number of queries at each turn. If the claim achieves the perfect score, no further rewrites will be generated in the next turn and we stop early. We observe sequences with **improved mAP@50 scores** shrink along the turns. This indicates that some claims reach a perfect mAP@50 score after only one or two modifications. In contrast, for sequences with a **decreased mAP@50 scores**, the circle sizes remain the same while performance drops. This suggests that for such claims, the more the model modifies it, the worse its performance is. For the sequence of claims with **the same mAP@50 scores** at the beginning and the end, there is a slight up and down for the slopes for the lines in between. This suggests that there are some sequences where the modified query achieves a better score while later modifications hurt the performance or vice versa. However, such scenarios are rare. Of the 13,089 claims in the development set, 1541 claims have the same AP@50 scores at the beginning and the end. Among these, 1243 are constant along the entire sequence and 271 have minor score changes, as reflected in Figure 6.4 as the blue line ($\text{mAP@50}_e = \text{mAP@50}_b$). Figure 6.5 shows the distribution of actions in the model output corresponding to increased performance, no changes in performance, and decreased performance. We can see that most of the actions lead to no performance change. The *remove* and *swap with synonym* actions result more often in increases in performance than decreases. In contrast, *add synonym* and *change tense to present simple* more often result in performance reduction. Figure 6.6 shows the average change per action. In this plot we observe that the net performance changes for *remove* and *swap with synonym* are positive, with an average of 1.83 and 0.88, respectively. The net performance

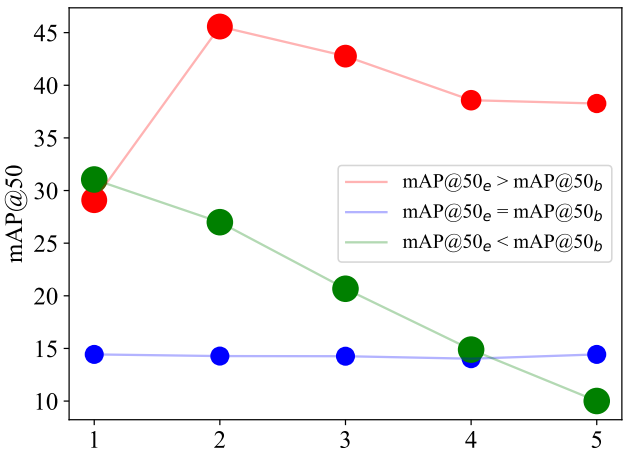

Figure 6.4: mAP@50 scores for all rewritten queries in the development set run against BM25. The x-axis indicates the claim rewriting sequence. The size of each circle represents the number of queries at each turn.The subscripts "e" and "b" correspond to "end" and "beginning" of the claim rewriting sequence, respectively.

Retriever(Query)	↑ rewards only			↑ + ↓ rewards		
	mAP@50	Recall	RR	mAP	Recall	RR
BM25(**RL**[Claim])	**32.43**	**35.8**	**30.23**	31.50	32.82	29.80
BM25(Claim)	26.82	29.68	22.30	26.82	29.68	22.30
kNN(**RL**[Claim])	**36.69**	**36.95**	29.17	34.49	35.06	**29.79**
kNN(Claim)	28.40	25.93	21.27	28.40	25.93	21.27

Table 6.4: Ablation experiments, RR refers to reciprocal rank.

changes for *add synonym* and *change tense to present simple* are negative, with an average of -0.34 and -0.98, respectively. We hypothesize that the model does not learn *add synonym* and *change tense to present simple* actions well due to the sparsity of such examples in the data as shown in Table 6.2. We further discuss the importance of these actions in Section 6.6.

6.5.4 Ablations

We conduct ablation experiments to evaluate the ability of our system in adapting to arbitrary endpoints and different performance metrics. Although the space of possible ablations is far larger than what we present here, we pick three dimensions of ablations that could be useful for practitioners and future researchers: (i) retriever type (BM25 or kNN), (ii) reward metric (average precision, recall, reciprocal rank) and (iii) presence of negative training examples.
Table 6.4 shows the results on each ablation when compared against a baseline of just using the initial claim. Across different metrics and retrievers we observe improvements in query performance: our system improves the original claim of up to 11% absolute recall points (42% relative

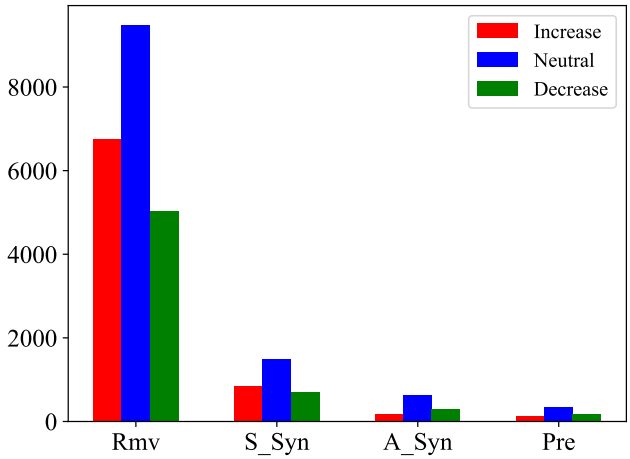

Figure 6.5: Distribution of predicted actions (*remove*, *swap with synonym*, *add synonym* and *change to present tense*) with AP@50 reward and BM25 retriever.

improvement) and works on both BM25 and kNN retrievers. We also observe that the inclusion of training sequences with query performance decrease (negative training examples), consistently leads to performance decreases on all metrics and retrievers as compared to just training on positive edit sequences –with the only exception of querying kNN with RR as reward. We posit that this performance gap is due to the difference in data quality, i.e, providing our models with noiseless training signals leads to more effective queries. However, even in cases where we include negative training examples our models still meaningfully improve over the original claim.

6.6 Discussion

Do we need to use (offline) RL for claim rewriting? It can be argued that a computationally expensive RL agent for query rewriting could be replaced by more economic design choices such as a sequence labeling model by fine-tuning a pretrained language model. In fact, as we discussed in the prior work section (6.2), researchers have indeed taken several different approaches for training neural text-editing models. In order to dig deeper into this question, we chose AP@50 as reward and trained a classifier on only the first edit in the training instances as 128-way classification (4 actions * 32 tokens), and the resulting classifier performed slightly worse than the RL agent trained on the whole edit sequence. However, we also observe from Figure 6.4 that when using the BM25 retriever and AP@50 as reward, the first action in training data is four times more effective than the following three actions on average, which means that the comparison between the classifier and the RL agent might not be a fair one. However, we also interpret the strong performance of the classifier as a more *efficient* alternative to training expensive reinforcement learning models. We leave a deeper comparison of the capabilities of sequence classification modeling and offline reinforcement learning for future work.

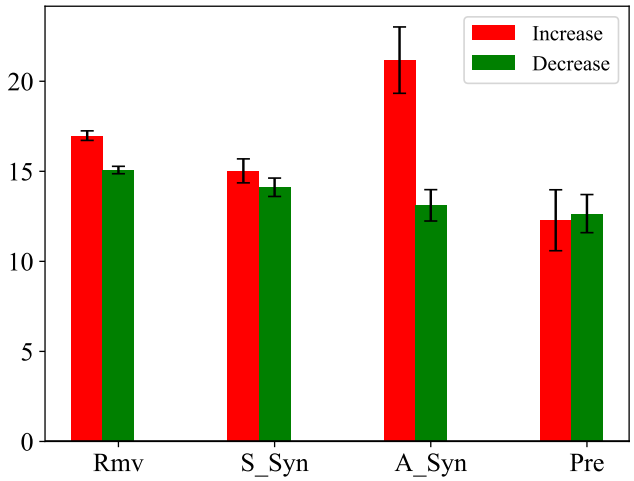

Figure 6.6: Average change in AP@50 scores of the predicted actions (*remove* (Rmv), *swap with synonym* (S_ Syn), *add synonym* (A_ Syn) and *change tense to present simple* (Pre)) against BM25. Statistics for actions with no changes in AP@50 are excluded as this results in 0 scores.

Are pretrained sentence embeddings good candidates for state representation? In our initial set of experiments we used frozen pretrained Sentence-BERT embeddings as state representation, and we did not see significant improvements over the initial claim. We observed a significant performance jump (5 mAP@50 points) once the sentence embeddings were also trained alongside the RL agent. This improvement highlights **the importance of state representation and shows that task-specific embeddings perform better than general-purpose embeddings**. This finding also indicates that the presence of Wikipedia data in the training data of LLMs does not simplify our task. Furthermore, there is significant prior work emphasizing the role and difficulty of the combinatorial and compositional nature of the state space in language tasks for reinforcement learning. [158], which also makes text-based RL agents a good choice for advancing our understanding of natural language.

What is the relation between query rewriting with sequence action learning and keyword extraction? We find that some of our models predict the *remove* action the vast majority of times, upwards of 80% in the case of using BM25 as retriever and AP@50 as reward. This brings up a natural question around how our method compares with keyword extraction methods, since the prevalence of remove edits during inference suggests that our approach works similar to keyword extraction. Our initial experiments with KeyBERT [159] show that this is not the case as **keyword extraction does not perform comparably with the claim baseline on BM25 and AP@50 as reward**. Although further analysis is required to make firm conclusions, it could be implied that including actions other than *remove* for rewriting queries can bring in significant gains.

6.7 Conclusion

In this chapter, we presented our findings on using an offline RL agent that learns editing strategies for query rewriting, so that fact-checkers can discover misinformation across social media platforms more effectively. Using a decision transformer, we showed that we can learn to rewrite misinformation claims by applying a series of interpretable actions such as adding synonyms or removing specific words. These actions can transform the claims into more effective queries, leading to a relative performance increase of up to 42% over a simpler kNN retriever baseline. Additionally, we conducted further analyses and ablation studies to develop a better understanding of our system, which showed that its adaptable to a variety of metrics and search engines. Our findings are an initial step towards building AI-assisted technologies to help fact-checkers discover online misinformation more effectively.

Future Work. While our work lays the grounds on using RL for building effective misinformation discovery tools, the practical application of our model requires further work to account for the limited access to social network APIs. This means additional constraints such as: (1) learning to rewrite claims under a fixed budget of training queries, and (2) learning without supervision. While there are already several solutions available for (2) [64, 42], we believe (1) is an exciting area for further exploration. Additionally, we posit our approach to be applicable on languages other than English since the RL agent we train is mainly language-agnostic.
In the recent three chapters, we presented datasets and models for different languages in helping fact-checkers scale their efforts through human-centered language technology (RQ2). In this chapter we contributed to this goal by helping fact-checkers find the best query for their search of similar claims across platforms using offline reinforcement learning. In chapters 4 and 5, we helped fact-checkers group similar claims together, and apply existing fact-checks to social media posts, through contributing novel datasets and models that outperformed state of the art, and enabled cross-lingual matching of tweets with fact-checks. In the next chapter, we address RQ3: building human-centered language technology to help internet users safely navigate around online misinformation. In particular, chapter 7 contributes to the third research question by helping users contextualize news through generating abstractive and extracting explanations of news.

CHAPTER 7

Contextualization by Explanation Generation

In this chapter, we explore the construction of natural language explanations for news claims, with the goal of assisting fact-checking and news evaluation applications. We experiment with two methods: (1) an extractive method based on Biased TextRank – a resource-effective unsupervised graph-based algorithm for content extraction; and (2) an abstractive method based on the GPT-2 language model. We perform comparative evaluations on two misinformation datasets in the political and health news domains, and find that the extractive method shows the most promise.

7.1 Introduction

Navigating the media landscape is becoming increasingly challenging given the abundance of misinformation, which reinforces the importance of keeping our news consumption focused and informed. While fake news and misinformation have been a recent focus of research studies [26, 126, 160], the majority of this work aims to categorize claims, rather than generate explanations that support or deny them. This is a challenging problem that has been mainly tackled by expert journalists who manually verify the information surrounding a given claim and provide a detailed verdict based on supporting or refuting evidence. More recently, there has been a growing interest in creating computational tools able to assist during this process by providing supporting explanations for a given claim based on the news content and context [129, 161]. While a true or false veracity label does not provide enough information and a detailed fact-checking report or news article might take long to read, bite-sized explanations can bridge this gap and improve the transparency of automated news evaluation systems.

To contribute to this line of work, this chapter explores two approaches to generate supporting explanations to assist with the evaluation of news. First, we investigate how an extractive method based on Biased TextRank [162] can be used to generate explanations. Second, we explore an abstractive method based on GPT-2, a large generative language model [163].

Our methods take as input a news article and a claim and generate a claim-focused explanation

by extracting or generating relevant information to the original article in relation to the claim. We evaluate our proposed methods on the health care and political domains, where misinformation is abundant. As current news on the COVID-19 pandemic and the elections are overloading social media outlets, we find these domains to be of timely importance. Through comparative experiments, we find that both methods are effective at generating explanations for news claims, with the extractive approach showing the most promise for this task.

Claim: nearly half of Oregon's children are poor.
Fact-Check Report: ...Jim Francesconi...said..."Nearly half of Oregon's children are poor." He said the information came from a 2012 report...According to that report, "nearly 50% of children are either poor or low-income." Francesconi almost immediately realized his mistake. "In retrospect, I wish I would have said poor or low income."...there is a distinction between poor and low income as far as the U.S. government is concerned." ... If you check the...Census information, you'll find that...23 percent of children in Oregon live in...below...poverty level while another 21 percent live in low-income families. As far as the U.S. government is concerned, about a quarter of the state's children are poor, not half... (redacted)
Supporting Explanation (Ground Truth): So where does this leave us? Francesconi said in an opinion piece that "nearly half of Oregon's children are poor." In fact, if you use federal definitions for poverty, about a quarter are poor and another quarter are low-income. But experts tell us that families that are described as low-income still struggle to meet their basic needs and, for all intents and purposes, qualify as poor. Be that as it may, Francesconi was referencing a report that used the federal definitions.
Biased TextRank (Extractive): "Nearly half of Oregon's children are poor." According to that report, "nearly 50% of children are either poor or low-income." Low income refers to families between 100 and 200 percent of the federal poverty level. As far as the U.S. government is concerned, about a quarter of the state's children are poor, not half.
GPT-2 Based (Abstractive):" That's still below the federal poverty level. But that's not half. About 47 percent of Oregon's children are not poor, according to the Census data. So the percentage of children in the state who are poor is not half yet. It's actually closer to half.

Table 7.1: An example data point from the LIAR-PLUS dataset, with ground truth explanations, and explanations generated by our methods.

7.2 Related Work

While explainability in AI has been a central subject of research in recent years [164, 165, 166], the generation of natural language explanations is still relatively understudied. Camburu et al. [167] propose e-SNLI, a natural language (NL) inference dataset augmented with human-annotated NL explanations. In their paper, Camburu et al. generated NL explanations for premise and hypothesis pairs for an inference task using the InferSent [168] architecture. Kumar and Talukdar [169] propose the task of generating "faithful" (i.e., aligned with the model's internal decision making) NL explanations and propose NILE, a method that jointly produces NLI labels and

faithful NL explanations.
Generating explanations in the context of news and fact-checking is a timely and novel topic in the NLP community [129, 161, 170]. In [129] the authors proposed a supervised BERT [104] based model for jointly predicting the veracity of a claim by extracting supporting explanations from fact-checked claims in the LIAR-PLUS [171] dataset. Kotonya and Toni [170] constructed a dataset for a similar task in the public health domain and provided baseline models for explainable fact verification using this dataset. Fan et al. [161] used explanations about a claim to assist fact-checkers and showed that explanations improved both the efficiency and the accuracy of the fact-checking process.

7.3 Methods

We explore two methods for producing natural language explanations: an extractive unsupervised method based on Biased TextRank, and an abstractive method based on GPT-2.

7.3.1 Extractive: Biased TextRank

Introduced by Kazemi et al. [162] and based on the TextRank algorithm [172], Biased TextRank is a targeted content extraction algorithm with a range of applications in keyword and sentence extraction. The TextRank algorithm ranks text segments for their importance by running a random walk algorithm on a graph built by including a node for each text segment (e.g., sentence), and drawing weighted edges by linking the text segment using a measure of similarity.
The Biased TextRank algorithm takes an extra "bias" input and ranks the input text segments considering both their own importance and their relevance to the bias term. The bias query is embedded into Biased TextRank using a similar idea introduced by Haveliwala [173] for topic-sensitive PageRank. The similarity between the text segments that form the graph and the "bias" is used to set the restart probabilities of the random walker in a run of PageRank over the text graph. That means the more similar each text segment is to the bias query, the more likely it is for that node to be visited in each restart and therefore, it has a better chance of ranking higher than the less similar nodes to the bias query. During our experiments, we use SBERT [105] contextual embeddings to transform text into sentence vectors and cosine similarity as similarity measure.

7.3.2 Abstractive: GPT-2 Based

We implement an abstractive explanation generation method based on GPT-2, a transformer-based language model introduced in Radford et al. [163] and trained on 8 million web pages containing 40 GBs of text.

Aside from success in language generation tasks [174, 175], the pretrained GPT-2 model enables us to generate abstractive explanations for a relatively small dataset through transfer learning.
In order to generate explanations that are closer in domain and style to the reference explanation, we conduct an initial fine-tuning step. While fine tuning, we provide the news article, the claim, and its corresponding explanation as an input to the model and explicitly mark the beginning and the end of each input argument with bespoke tokens. At test time, we provide the article and query inputs in similar format but leave the explanation field to be completed by the model. We use top-k sampling to generate explanations. We stop the generation after the model outputs the explicit end of the text token introduced in the fine-tuning process.
Overall, this fine-tuning strategy is able to generate explanations that follow a style similar to the reference explanation. However, we identify cases where the model generates gibberish and/or repetitive text, which are problems previously reported in the literature while using GPT-2 [176, 177]. To address these issues, we devise a strategy to remove unimportant sentences that could introduce noise to the generation process. We first use Biased TextRank to rank the importance of the article sentences towards the question/claim. Then, we repeatedly remove the least important sentence (up to 5 times) and input the modified text into the GPT-2 generator. This approach keeps the text generation time complexity in the same order of magnitude as before and reduces the generation noise rate to close to zero.

7.4 Evaluation

7.4.1 Experimental Setup

We use a medium (355M hyper parameters) GPT-2 model [163] as implemented in the Huggingface transformers [178] library. We use ROUGE [179], a common measure for language generation assessment as our main evaluation metric for the generated explanations and report the F score on three variations of ROUGE: ROUGE-1, ROUGE-2 and ROUGE-L.
We compare our methods against two baselines. The first is an explanation obtained by applying TextRank on the input text. The second, called "embedding similarity", ranks the input sentences by their embedding cosine similarity to the question and takes the top five sentences as an explanation.

7.4.2 Datasets

LIAR-PLUS. The LIAR-PLUS [171] dataset contains 10,146 train, 1,278 validation and 1,255 test data points collected from PolitiFact.com, a political fact-checking website in the U.S. A datapoint in this dataset contains a claim, its verdict, a news-length fact-check report justifying

Dataset	Total count	Av. Words	Av. Sent.
LIAR-PLUS	12,679	98.89	5.20
HNR	16,500	87.82	4.63

Table 7.2: Dataset statistics for explanations; total count, average words and sentences per explanation.

Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank	27.74	7.42	23.24
GPT-2 Based	24.01	5.78	21.15
Biased TextRank	**30.90**	**10.39**	**26.22**

Table 7.3: ROUGE-N scores of generated explanations on the LIAR-PLUS dataset.

the verdict and a short explanation called "Our ruling" that summarizes the fact-check report and the verdict on the claim. General statistics on this dataset are presented in Table 7.2.

Health News Reviews (HNR). We collect health news reviews along with ratings and explanations from healthnewsreview.org, a website dedicated to evaluating healthcare journalism in the US. [1] The news articles are rated with a 1 to 5 star scale and the explanations, which justify the news' rating, consist of short answers for 10 evaluative questions on the quality of information reported in the article. The questions cover informative aspects that should be included in the news such as intervention costs, treatment benefits, discussion of harms and benefits, clinical evidence, and availability of treatment among others. Answers to these questions are further evaluated as either satisfactory, non-satisfactory or non-applicable to the given news item. For our experiments, we select 1,650 reviews that include both the original article and the accompanying metadata as well as explanations. Explanations' statistics are presented in Table 7.2.
To further study explanations in this dataset, we randomly select 50 articles along with their corresponding questions and explanations. We then manually label sentences in the original article that are relevant to the quality aspect being measured.[2] During this process we only include explanations that are deemed as "satisfactory," which means that relevant information is included in the original article.

7.4.3 Producing Explanations

We use the Biased TextRank and the GPT-2 based models to automatically generate explanations for each dataset. With LIAR-PLUS, we seek to generate the explanation provided in the "Our

[1] We followed the restrictions in the site's *robots.txt* file.

[2] The annotation was conducted by two annotators, with a Pearson's correlation score of 0.62 and a Jaccard similarity of 0.75.

Model	Explanations			Relevant Sentences		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Embedding Similarity	18.32	2.96	15.25	22.02	8.79	20.21
GPT-2 Based	**20.02**	**4.32**	**17.67**	15.74	2.58	13.32
Biased TextRank	19.41	3.41	15.87	**23.54**	**10.15**	**21.88**

Table 7.4: ROUGE evaluation on the HNR dataset. Left columns under "Explanations" have the actual explanations as reference and the columns on the right provide results for comparison against question-relevant sentences.

Model	Acc.	F1 (+)	F1 (-)
GPT-2 Based	64.40%	49.04%	54.67%
Biased TextRank	**65.70%**	**56.69%**	**57.96%**

Table 7.5: Downstream evaluation results on the HNR dataset, averaged over 10 runs and 9 questions.

ruling" section. For HNR we aim to generate the explanation provided for the different evaluative questions described in section 7.4.2. We use the provided training, validation and test splits for the LIAR-PLUS dataset. For HNR, we use 20% of the data as the test set and we study the first nine questions for each article only and exclude question #10 as answering it requires information beyond the news article. We use explanations and question-related article sentences as our references in ROUGE evaluations over the HNR dataset, and the section labeled "Our ruling" as ground truth for LIAR-PLUS.

Extractive Explanations. To generate extractive explanations for the LIAR dataset, we apply Biased TextRank on the original article and its corresponding claim and pick the top 5 ranked sentences as the explanation (based on the average length of explanations in the dataset). To generate explanations on the HNR dataset, we apply Biased TextRank on each news article and question pair for 9 of the evaluative questions and select the top 5 ranked sentences as the extracted explanation (matching the dataset average explanation length).

Abstractive Explanations. We apply the GPT-2 based model to generate abstractive explanations for each dataset using the original article and the corresponding claim or question as an input. We apply this method directly on the LIAR-PLUS dataset. On the HNT dataset, since we have several questions, we train separate GPT-2 based models per question. In addition, each model is trained using the articles corresponding to questions labeled as "satisfactory" only as the "unsatisfactory" or "not applicable" questions do not contain information within the scope of the original article.

7.4.4 Downstream Evaluation

We also conduct a set of experiments to evaluate to what extent we can answer the evaluation questions in the HNR dataset with the generated explanations. For each question, we assign binary labels to the articles (1 for satisfactory answers, 0 for not satisfactory and NA answers) and train individual classifiers aiming to discriminate between these two labels. During these experiments each classifier is trained and evaluated ten times on the test set and the results are averaged over the ten runs.

7.5 Experimental Results

As results in Table 7.3 suggest, while our abstractive GPT-2 based model fails to surpass extractive baselines on the LIAR-PLUS dataset, Biased TextRank outperforms the unsupervised TextRank baseline. Biased TextRank's improvements over TextRank suggest that a claim-focused summary of the article is better at generating supporting explanations than a regular summary produced by TextRank. Note that the current state-of-the-art results for this dataset, presented in [129] achieve 35.70, 13.51 and 31.58 in ROUGE-1, 2 and L scores respectively. However, a direct comparison with their method would not be accurate as it is a method that is *supervised* (versus the unsupervised Biased TextRank) and *extractive* (versus the abstractive GPT-2 based model).
Table 7.4 presents results on automatic evaluation of generated explanations for the HNR dataset, showing that the GPT-2 based model outperforms Biased TextRank when evaluated against actual explanations and Biased TextRank beats GPT-2 against the extractive baseline. This indicates the GPT-2 based method is more effective in this dataset and performs comparably with Biased TextRank. Results for the downstream task using both methods are shown in Table 7.5. As observed, results are significantly different and demonstrate that Biased TextRank significantly outperforms (t-test $p = 0.05$) the GPT-2-based abstractive method, thus suggesting that Biased TextRank generates good quality explanations for the HNR dataset.

7.6 Discussion

Our evaluations indicate that Biased TextRank shows the most promise, while the GPT-2 based model mostly follows in performance. Keeping in mind that the GPT-2 based model is solving the harder problem of *generating* language, it is worth noting the little supervision it receives on both datasets, especially on the HNR dataset where the average size of the training data is 849.
In terms of resource efficiency and speed, Biased TextRank is faster and lighter than the GPT-2 based model. Excluding the time needed to fine-tune the GPT-2 model, it takes approximately 60

seconds on a GPU to generate a coherent abstractive explanation on average on the LIAR-PLUS dataset, while Biased TextRank extracts explanations in the order of milliseconds and can even do it without a GPU in a few seconds. We find Biased TextRank's efficiency as another advantage of the unsupervised algorithm over the GPT-2 based model.

7.7 Conclusion

In this chapter, we presented extractive and abstractive methods for generating supporting explanations for more convenient and transparent human consumption of news. We evaluated our methods on two domains and found promising results for producing explanations. In particular, Biased Text-Rank (an extractive method) outperformed the unsupervised baselines on the LIAR-PLUS dataset and performed reasonably close to the extractive ground-truth on the HNR dataset. Biased TextRank is also easy to adapt to work in multilingual and non-English settings where only the embedding vectors encoding the text spans need to be updated.
For future work, we believe generating abstractive explanations should be a priority, since intuitively an increase in the readability and coherence of the supporting explanations will result in improvements in the delivery and perception of news.
This chapter, alongside chapter 2, contribute to the third research question around building human-centered NLP for helping users safely navigate around misinformation. This chapter did so through contextualizing health news and fact-check reports, and chapter 2 used alongside the models and approach from chapter 4 have been helping real users in India, Brazil, and the Philipinnes to flag and learn more about potentially misleading information on end-to-end encrypted social media. The following chapters include our conclusionary remarks. In the next chapter we aim to answer a high level question, that if all human-centered language and AI technology to understand and quantify misinformation on the internet existed, how can we put those pieces together and leverage the insights to collectively fight back against misinformation.

CHAPTER 8

Conclusion

This dissertation has been a deep dive into the world of human-centered Natural Language Processing (NLP), highlighting its crucial role in helping people understand, identify, and navigate through misinformation. Our mission to enhance the quality of online information began by developing NLP methods to explore how misinformation operates across various platforms and user groups. Inspired by these insights, we developed multilingual models and datasets aimed at empowering fact-checkers and journalists globally to identify misinformation at scale. Alongside this, we created NLP tools to make online experiences safer for users, aiding them in steering clear of misinformation. This comprehensive research emphasizes the vital role of human-centered NLP in bolstering our collective ability to tackle the challenges posed by misinformation in the digital realm.
We now revisit the thesis' goals and discuss our findings and contributions to human-centered NLP for countering misinformation in details:

8.1 Revisiting the Goal of RQ1: How to use language technology to gain a human-centered understanding of misinformation?

To develop language technology aimed at countering and navigating through misinformation, it is essential to grasp its underlying mechanisms. In chapters 2 and 3, we delved into how NLP can be harnessed to gain a deeper understanding of the phenomenon of misinformation.
In chapter 2, we conducted the first study on misinformation within the realm of end-to-end encrypted social media, specifically focusing on WhatsApp during the 2019 Indian general elections. Our investigation revealed that when users have the option to participate in tiplines, a significant portion of viral content in public WhatsApp groups is reported to these tiplines before reaching widespread circulation. An in-depth analysis of the texts, images, and URLs submitted to our study's tipline uncovered that the most frequently shared content across these conversations is indeed misinformation. Our research shed light on the dynamics of misinformation within end-to-end encrypted social media and underscored the valuable role tiplines can play in

bringing transparency to these otherwise opaque platforms.
Chapter 3 demonstrated that individuals from various backgrounds exhibit distinct belief patterns regarding misinformation. Through a series of user-centered modeling experiments, we ascertained that information about a user's demographics, including factors such as age, gender, education, and race, can enhance models for understanding misinformation perception and elucidating the reasons behind individual beliefs. Moreover, our findings revealed that the greater the homogeneity within a group's perception of misinformation, the more the demographic characteristics of the group members can explain their individual views on misinformation. This suggested a link between susceptibility to misinformation and the demographic profile of the user. Ultimately, the insights from chapter 3 underscored the need to consider individual differences in the fight against misinformation, recognizing that people from diverse backgrounds may be susceptible to different forms of misinformation.

8.2 Revisiting the Goal of RQ2: How to utilize language technology to identify misinformation at scale while keeping humans in the loop?

In chapters 4, 5, and 6, we researched the utilization of NLP for identifying misinformation at scale, with a key requirement that human oversight remained pivotal. These interventions enhanced the expertise of fact-checkers and were rooted in a human-centered comprehension of misinformation.
In chapter 4, we introduced the concept of claim matching across five languages, spanning both high and low resource domains, with the aid of two novel datasets. Leveraging knowledge distillation, we trained a student embedding model to align with a teacher model (XLM-RoBERTa) by using parallel data between the teacher and student languages. Our findings revealed that our multilingual model consistently outperforms state-of-the-art multilingual embeddings across various settings, including retrieval and classification. The work in chapter 4 formed the foundation for improved search and claim matching performance. It empowered fact-checkers to operate more efficiently, enabling them to fact-check similar claims only once. Moreover, this model has been successfully deployed in large-scale scenarios, providing crucial fact-checking support during elections in Brazil, the Philippines, and France.
Additionally, in chapter 5, we investigated the application of multilingual embeddings in retrieving relevant fact-checks for tweets in English, Spanish, Portuguese, and Hindi, both in single-language and cross-language contexts. We observed that state-of-the-art multilingual embedding models often struggle when processing lengthy fact-check reports, falling short compared to BM25, except in the cross-language setting where embedding models prove their utility. We have also made our novel multilingual dataset available to support future research in this critical area,

facilitating open research and assisting fact-checkers in locating relevant fact-checks for their queries within fact-checked claim reports.
Manually identifying the most effective search terms for fact-checkers across diverse platforms when researching potentially misleading claims is a challenging task. In chapter 6, we addressed this issue by aiding fact-checkers in transforming their initial claim into an effective search query. We accomplished this by training an offline reinforcement learning agent using a Decision Transformer, to optimize edit actions that enhance a potentially misleading claim. Using the FEVER evidence retrieval dataset, we applied straightforward editing actions (e.g., word removal, synonym substitution) to iteratively improve the initial claim's effectiveness for desired performance metrics, such as precision or recall. Our research further investigated model behaviors and conducted ablation studies to uncover the hyperparameters contributing to enhanced performance. Our work provided a tangible demonstration of how fact-checkers can employ a human-interpretable query rewriter adaptable to diverse social media search platforms. This tool transformed claims under fact-checking into effective queries based on adjustable performance and platform criteria, thereby facilitating the scalability and efficiency of fact-checkers.
The models, systems, and datasets presented in these three chapters, which addressed the second research question (language technology for aiding humans in identifying misinformation at scale), substantially enhanced the efficiency and effectiveness of fact-checkers through claim matching, linking claims with fact-check reports, and optimizing claim queries.

8.3 Revisiting the Goal of RQ3: How can NLP help users safely navigate around misinformation in online environments?

Despite the existence of fact-checking pipelines overseen by humans, internet users may still encounter misleading information on the web. In chapter 7, we introduced novel tasks and NLP models aimed at generating explanations to assist users in navigating misinformation. This chapter explored both extractive (Biased TextRank) and abstractive (fine-tuned GPT-2) methods for extracting and generating explanations related to health news and fact-check reports.
Our findings revealed that, for the HNR dataset, which features abstractive explanations as ground truth, GPT-2 outperformed other methods, albeit marginally. However, in the LIAR-PLUS dataset, which predominantly contains extractive explanations from longer fact-check reports, Biased TextRank demonstrated significant superiority. Both methods exhibited improvements over existing baselines for explanation generation, offering the capability to contextualize internet information, thereby shielding users from deceptive content.
In chapter 2, we delved into the realm of misinformation on end-to-end encrypted social media, focusing on a case study of the 2019 Indian general elections. Our investigation highlights the

effectiveness of a crowdsourced opt-in tipline, operational during the election period, in assisting users in seeking additional information regarding potentially misleading claims in public WhatsApp groups. When combined with the work described in chapter 4, these tiplines empowered social media users to access information from fact-checkers, enabling the scaling of fact-checking efforts to effectively respond to the surge in tipline requests. Thus, as detailed in chapter 2, these tiplines are valuable for social media users, particularly those on end-to-end encrypted platforms like WhatsApp and Telegram, enabling them to navigate through misinformation safely.
Taken together, these findings addressed the third research question (utilizing NLP to facilitate safe navigation through misinformation) by equipping internet users with tools that enhance their online environments with factually accurate information.

8.4 The Way Forward

This thesis has shed light on research questions around human-centered NLP to counter misinformation through contributing NLP models, systems, algorithms, and datasets. We demonstrated that NLP can help us gain a deeper understanding of the phenomena of misinformation, build systems that help scale up human fact-checking efforts, and help users navigate safely around misinformation on the internet.
Fully automated misinformation detection is considered a challenging NLP and AI task, and so far it has not been operationalized in any major way as the state-of-the-art performs poorly on real world data. Every automatic misinformation detection pipeline must achieve at least the five following elements: (i) reliable and up-to-date data streams, (ii) parse text into check-worthy claims, (iii) identify relevant context, (iv) reason over contextual knowledge to assign a veracity label or score, (v) provide explanations or demonstrate why the decision is a sane one. State-of-the-art NLP and fake news detection systems currently do none of these tasks well enough to make a fully automatic pipeline practical. If the goal is to achieve full automation, these five subproblems are important questions for future research.
In contrast with automatic misinformation detection, leveraging the intelligence of crowds to conduct fact-checks is another scalable way of identifying misinformation as recent research [180] has demonstrated that a panel of sixteen or more crowd fact-checkers can outperform fact-checking performance equivalent of three expert journalists, provided the crowd is ideologically balanced. Crowd fact-checking comes with new NLP challenges such as finding the right jury to fact-check a post or aggregating opinions of crowd fact-checkers to explain the veracity of claims. Not without its shortcomings [181], X's community notes is one example of how such strategies can be carried out at scale to enable various individuals fact-check a large number of daily posts. Furthermore, network science can be leveraged to understand the dynamics of misinformation

spread, and identify coordinated disinformation campaigns in social networks.
Additionally, future research in areas of multi and cross linguality (misinformation is a global issue), multimodality (misinformation is not just text), and retrieval (fact-checkers benefit from search functionalities) within the domain of misinformation are necessary for improving human-centered NLP to counter misinformation more broadly and effectively.
With the increase in popularity of large language models in recent years, several risks and opportunities rise in regards to misinformation [182]. Large language models can help with many of the subproblems previously mentioned as important to countering misinformation such as claim detection, reasoning over contextual knowledge, and better retrieval support for fact-checkers. LLMs can however be used to generate fake news and deepfake videos for targeting internet users with disinformation, and are known to hallucinate or produce factually inaccurate and inconsistent content. Such issues are important areas for future research, as they currently pose hurdles to the applicability of AI in domains and tasks where factuality of LLMs is highlighted.
Finally, misinformation is a term commonly overused outside of its domain of application and while categorization does exist for different types of misleading information [183], our fundamental understanding of what actually constitutes misinformation is limited. Sometimes the presence of correct but manipulative information could have more malicious outcomes than false information, and so the need for understanding the fundamental mechanisms that drive human beliefs is necessary. In many cultures, there are generational folklore myths that have been taken as truths by a significant portion of the population, and while they are sometimes not based in reality, they are also considered truths by entire factions of societies. Other scholars have compared fake news with deception research, focusing more on an intent to deceive rather than investigating the accuracy and reliability of the information. It is still unclear why humans believe misinformation, even after they have been exposed to the correct version of stories [184]. Cross and trans disciplinary research is required for a deeper understanding of how individuals interact with and are affected by this phenomenon and quantifying misinformation's socioeconomic impact is an important area for future research.

BIBLIOGRAPHY

[1] Joseph B. Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel A. Centeno, Iain D. Couzin, Jonathan F. Donges, Mirta Galesic, Andrew S. Gersick, Jennifer Jacquet, Albert B. Kao, Rachel E. Moran, Pawel Romanczuk, Daniel I. Rubenstein, Kaia J. Tombak, Jay J. Van Bavel, and Elke U. Weber. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118(27):e2025764118, 2021. doi: 10.1073/pnas.2025764118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2025764118.

[2] Amnesty International. Myanmar: Facebook's systems promoted violence against rohingya; meta owes reparations. *Amnesty International*, 2022. URL https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/.

[3] Maria Mercedes Ferreira Caceres, Juan Pablo Sosa, Jannel A Lawrence, Cristina Sestacovschi, Atiyah Tidd-Johnson, Muhammad Haseeb UI Rasool, Vinay Kumar Gadamidi, Saleha Ozair, Krunal Pandav, Claudia Cuevas-Lou, et al. The impact of misinformation on the covid-19 pandemic. *AIMS public health*, 9(2):262, 2022.

[4] Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, 2022.

[5] Ullrich KH Ecker, Stephan Lewandowsky, Olivia Fenton, and Kelsey Martin. Do people keep believing because they want to? preexisting attitudes and the continued influence of misinformation. *Memory & cognition*, 42:292–304, 2014.

[6] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[7] Federal Communications Commission et al. Obscene, indecent and profane broadcasts. *Federal Communications Commission Consumer Guides*, 13, 2017.

[8] Sharon Kann and Angelo Carusone. In less than a month, elon musk has driven away half of twitter's top 100 advertisers. *Media Matters for America*, 2022. URL https://www.mediamatters.org/elon-musk/less-month-elon-musk-has-driven-away-half-twitters-top-100-advertisers.

[9] R Cavazos. The economic cost of bad actors on the internet: Fake news, 2019.

[10] Heidi Shierholz and Ben Zipperer. Here is what's at stake with the conflict of interest ('fiduciary') rule. 2017.

[11] Ron Carson. Retirement savers are losing $17 billion a year from fake news and conflicts of interest. *Forbes*, 2018. URL https://www.forbes.com/sites/rcarson/2018/10/14/retirement-savers-are-losing-17-billion-a-year-from-fake-news-bad-advice-conflicts-of-interest/?sh=4c6e6e6abec7.

[12] Dara Kerr. Tech companies' newest cause celebre? boycott breitbart. *CNET*, 2017. URL https://www.cnet.com/tech/tech-industry/boycott-breitbart-lyft-hewlett-packard-t-mobile-autodesk-uber-amazon/.

[13] Jack Queen. Alex jones ordered to pay $473 million in punitive damages in sandy hook defamation case. *Reuters*, 2022. URL https://www.reuters.com/legal/alex-jones-must-pay-473-million-punitive-damages-sandy-hook-defamation-case-2022-11-10/.

[14] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[15] Sun Kyong Lee, Juhyung Sun, Seulki Jang, and Shane Connelly. Misinformation of covid-19 vaccines and vaccine hesitancy. *Scientific Reports*, 12(1):13681, 2022.

[16] Polydor Ngoy Mutombo, Mosoka P Fallah, Davison Munodawafa, Ahmed Kabel, David Houeto, Tinashe Goronga, Oliver Mweemba, Gladys Balance, Hans Onya, Roger S Kamba, et al. Covid-19 vaccine hesitancy in africa: a call to action. *The Lancet Global Health*, 10(3):e320–e321, 2022.

[17] Renee Garett and Sean D Young. Online misinformation and vaccine hesitancy. *Translational behavioral medicine*, 11(12):2194–2199, 2021.

[18] Alex J Xu, Jacob Taylor, Tian Gao, Rada Mihalcea, Veronica Perez-Rosas, and Stacy Loeb. Tiktok and prostate cancer: misinformation and quality of information using validated questionnaires. 2021.

[19] Stacy Loeb, Jacob Taylor, James F Borin, Rada Mihalcea, Veronica Perez-Rosas, Nataliya Byrne, Austin L Chiang, and Aisha Langford. Fake news: spread of misinformation about urological conditions on social media. *European urology focus*, 6(3):437–439, 2020.

[20] Stacy Loeb, Hala T Borno, Scarlett Gomez, Joseph Ravenell, Akya Myrie, Tatiana Sanchez Nolasco, Nataliya Byrne, Renee Cole, Kristian Black, Sabrina Stair, et al. Representation in online prostate cancer content lacks racial and ethnic diversity: implications for black and latinx men. *The Journal of Urology*, 207(3):559–564, 2022.

[21] Carly M Goldstein, Eleanor J Murray, Jennifer Beard, Alexandra M Schnoes, and Monica L Wang. Science communication in the age of misinformation. *Annals of Behavioral Medicine*, 54(12):985–990, 2020.

[22] Jevin D. West and Carl T. Bergstrom. Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15):e1912444117, 2021. doi: 10.1073/pnas.1912444117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1912444117.

[23] Dietram A. Scheufele and Nicole M. Krause. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669, 2019. doi: 10.1073/pnas.1805871115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1805871115.

[24] Office of the Surgeon General et al. Confronting health misinformation: The us surgeon general's advisory on building a healthy information environment [internet]. 2021.

[25] Vivek H. Murthy. Confronting health worker burnout and well-being. *New England Journal of Medicine*, 387(7):577–579, 2022. doi: 10.1056/NEJMp2207252. URL https://doi.org/10.1056/NEJMp2207252.

[26] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1287.

[27] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

[28] Rui Hou, Veronica Perez-Rosas, Stacy Loeb, and Rada Mihalcea. Towards automatic detection of misinformation in online medical videos. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 235–243, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368605. doi: 10.1145/3340555.3353763. URL https://doi.org/10.1145/3340555.3353763.

[29] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[30] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

[31] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

[32] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/619. URL https://doi.org/10.24963/ijcai.2021/619. Survey Track.

[33] Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1073. URL https://aclanthology.org/P15-1073.

[34] Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1119. URL https://aclanthology.org/D17-1119.

[35] David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2134. URL https://aclanthology.org/P14-2134.

[36] Aparna Garimella, Carmen Banea, and Rada Mihalcea. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1242. URL https://aclanthology.org/D17-1242.

[37] Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.334. URL https://aclanthology.org/2020.emnlp-main.334.

[38] Jenny Fan and Amy X. Zhang. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

Systems, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376293. URL https://doi.org/10.1145/3313831.3376293.

[39] Peter Elkind, Jack Gillum, and Craig Silverman. How facebook undermines privacy protections for its 2 billion whatsapp users, 2021.

[40] Chinmayi Arun. On whatsapp, rumours, and lynchings. *Economic & Political Weekly*, 54(6): 30–35, 2019.

[41] Meedan. One year of running the whatsapp end-to-end fact-checking project, 2020. URL https://meedan.com/blog/one-of-year-of-running-the-end-end-to-fact-checking-project/.

[42] Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.347. URL https://aclanthology.org/2021.acl-long.347.

[43] Gowhar Farooq. Politics of fake news: how whatsapp became a potent propaganda tool in india. *Media Watch*, 9(1):106–117, 2017.

[44] CSDS Lokniti. How widespread is whatsapp's usage in india? Live Mint, 2018. URL https://livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html.

[45] Kiran Garimella and Dean Eckles. Images and misinformation in political groups: Evidence from whatsapp in india. *Harvard Kennedy School Misinformation Review*, 2020.

[46] Pushkal Agarwal, Kiran Garimella, Sagar Joglekar, Nishanth Sastry, and Gareth Tyson. Characterising user content on a multi-lingual social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 2–11, 2020.

[47] Gautam Kishore Shahi. Amused: An annotation framework of multi-modal social media data, 2020.

[48] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media*, page 100104, 2021.

[49] Julio Reis, Philipe Melo, Kiran Garimella, and Fabricio Benevenuto. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *The Harvard Kennedy School (HKS) Misinformation Review*, 2020. URL https://doi.org/10.37016/mr-2020-035.

[50] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[51] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[52] Matthijs Douze, Jeff Johnson, and Hervé Jegou. Faiss: A library for efficient similarity search, 2017.

[53] Jonathan Bright. The Social News Gap: How News Reading and News Sharing Diverge. *Journal of Communication*, 66(3):343–365, 06 2016. ISSN 0021-9916. doi: 10.1111/jcom.12232. URL https://doi.org/10.1111/jcom.12232.

[54] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[55] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2), apr 2021. ISSN 2692-1626. doi: 10.1145/3412869. URL https://doi.org/10.1145/3412869.

[56] Helen Margetts, Peter John, Scott Hale, and Taha Yasseri. *Political Turbulence: How Social Media Shape Collective Action*. Princeton University Press, 2015.

[57] Jonathan Bright. The social news gap: How news reading and news sharing diverge. *Journal of Communication*, 66(3):343–365, 06 2016. ISSN 0021-9916. doi: 10.1111/jcom.12232. URL https://doi.org/10.1111/jcom.12232.

[58] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*, WWW '19, page 818–828, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313688. URL https://doi.org/10.1145/3308558.3313688.

[59] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098131. URL https://doi.org/10.1145/3097983.3098131.

[60] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims,

and fake news. In *Proceedings of the 43rd European Conference on Information Retrieval*, ECIR '21, pages 639–649, Lucca, Italy, March 2021. URL https://link.springer.com/chapter/10.1007/978-3-030-72240-1_75.

[61] Kat Lo. Fact-checking and mental health. *Meedan*, 2020. URL https://meedan.com/post/fact-checking-and-mental-health.

[62] Meedan. Fact champ: New project to increase collaboration between fact-checkers, academics, and community leaders to counter misinformation online. *Meedan*, 2021. URL https://meedan.com/blog/fact-champ-launch/.

[63] Philipe Melo, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. Whatsapp monitor: A fact-checking system for whatsapp. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):676–677, Jul. 2019. URL https://ojs.aaai.org/index.php/ICWSM/article/view/3271.

[64] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.332. URL https://aclanthology.org/2020.acl-main.332.

[65] Maldita.es. Disinformation on whatsapp: Maldita.es' chatbot and the "frequently forwarded" attribute. *Maldita.es*, 2021. URL https://web.archive.org/web/20211129201556/https:/maldita.es/uploads/public/docs/disinformation_on_whatsapp_ff.pdf.

[66] Keith Payne. *The broken ladder: How inequality affects the way we think, live, and die*. Penguin, 2018.

[67] Eleanor D Brown, Mariam D Seyler, Andrea M Knorr, Mallory L Garnett, and Jean-Philippe Laurenceau. Daily poverty-related stress and coping: Associations with child learned helplessness. *Family Relations*, 65(4):591–602, 2016.

[68] Jerome Rabow, Sherry L Berkman, and Ronald Kessler. The culture of poverty and learned helplessness: A social psychological perspective. *Sociological Inquiry*, 53(4):419–434, 1983.

[69] Jennifer A. Whitson and Adam D. Galinsky. Lacking control increases illusory pattern perception. *Science*, 322(5898):115–117, 2008. doi: 10.1126/science.1159845. URL https://www.science.org/doi/abs/10.1126/science.1159845.

[70] Sander Van der Linden. The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and Individual Differences*, 87:171–173, 2015.

[71] Andrea Pereira, Elizabeth Harris, and Jay J. Van Bavel. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations*, 26(1):24–47, 2023. doi: 10.1177/13684302211030004. URL https://doi.org/10.1177/13684302211030004.

[72] Claire E Robertson, Clara Pretus, Steve Rathje, Elizabeth A Harris, and Jay J Van Bavel. How social identity shapes conspiratorial belief. *Current Opinion in Psychology*, 47:101423, 2022.

[73] Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.88. URL https://aclanthology.org/2023.acl-short.88.

[74] Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.nlperspectives-1.2.

[75] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.222. URL https://aclanthology.org/2022.acl-long.222.

[76] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[77] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199, 2020.

[78] Clara Pretus, Camila Servin-Barthet, Elizabeth Harris, William Brady, Oscar Vilarroya, and Jay Van Bavel. The role of political devotion in sharing partisan misinformation. 2022.

[79] Manjul Gupta, Denis Dennehy, Carlos M Parra, Matti Mäntymäki, and Yogesh K Dwivedi. Fake news believability: The effects of political beliefs and espoused cultural values. *Information & Management*, 60(2):103745, 2023.

[80] Laura D Scherer, Jon McPhetres, Gordon Pennycook, Allison Kempe, Larry A Allen, Christopher E Knoepke, Channing E Tate, and Daniel D Matlock. Who is susceptible to

online health misinformation? a test of four psychosocial hypotheses. *Health Psychology*, 40(4):274, 2021.

[81] Milton King and Paul Cook. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.299.

[82] Charles Welch, Verónica Pérez-Rosas, Jonathan K Kummerfeld, and Rada Mihalcea. Learning from personal longitudinal dialog data. *IEEE Intelligent systems*, 34(4):16–23, 2019.

[83] Charles Welch, Verónica Pérez-Rosas, Jonathan K Kummerfeld, and Rada Mihalcea. Look who's talking: Inferring speaker attributes from personal longitudinal dialog. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 476–490. Springer, 2019.

[84] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[85] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1542. URL https://aclanthology.org/P19-1542.

[86] Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*, 2023.

[87] William A Falcon. Pytorch lightning. *GitHub*, 3, 2019.

[88] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

[89] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, 2020.

[90] Whitney Phillips. The oxygen of amplification. *Data & Society*, 22:1–128, 2018.

[91] Claire Wardle. Lessons for reporting in an age of disinformation. *First Draft*, 28, 2018.

[92] C Wardle, A Pimenta, G Conter, and ND Pedro Burgos. Comprova: An evaluation of the impact of a collaborative journalism project on brazilian journalists and audiences. *First Draft*, 2019.

[93] Meedan. Press release: New whatsapp tip line launched to understand and respond to misinformation during elections in india, 2019. URL https://medium.com/@meedan/press-release-new-whatsapp-tip-line-launched-to-understand-and-respond-to-misinformation-during-f4fce616adf4.

[94] Raúl Magallón Rosa. Verificado méxico 2018: Desinformación y fact-checking en campaña electoral. *Revista de comunicación*, 18(1):234–258, 2019.

[95] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 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, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

[96] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1081. URL https://aclanthology.org/S16-1081.

[97] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2045. URL https://aclanthology.org/S15-2045.

[98] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2010. URL https://aclanthology.org/S14-2010.

[99] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June 2013.

Association for Computational Linguistics. URL https://aclanthology.org/S13-1004.

[100] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1051.

[101] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

[102] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[103] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[104] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[105] Nils Reimers and Iryna Gurevych. 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

[106] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029. URL https://aclanthology.org/D18-2029.

[107] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin

Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[108] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038.

[109] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[110] Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5370–5378. AAAI Press, 2019.

[111] Nguyen Vo and Kyumin Lee. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.621. URL https://aclanthology.org/2020.emnlp-main.621.

[112] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer, 2020.

[113] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. Claimskg: A knowledge graph of fact-checked claims. In *International Semantic Web Conference*, pages 309–324. Springer, 2019.

[114] J. J. Randolph. Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. 2005. URL https://eric.ed.gov/?id=ED490661.

[115] Matthijs J. Warrens. Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, (4), 2010. doi: 10.1007/s11634-010-0073-4. URL https://doi.org/10.1007/s11634-010-0073-4.

[116] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017.

[117] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. 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 4667–4676, 2019.

[118] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365.

[119] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[120] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[121] Norbert Fuhr. Some common mistakes in ir evaluation, and how they can be avoided. In *ACM SIGIR Forum*, volume 51, pages 32–41. ACM New York, NY, USA, 2018.

[122] Caio Sacramento de Britto Almeida and Débora Abdalla Santos. Text similarity using word embeddings to classify misinformation. In *Workshop on Digital Humanities and Natural Language Processing*, DHandNLP 2020, 2020. URL https://arxiv.org/abs/2003.06634.

[123] Julie Tibshirani. Text similarity search with vector fields, 2019. URL https://www.elastic.co/blog/text-similarity-search-with-vectors-in-elasticsearch.

[124] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[125] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

[126] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1283.

[127] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812, 2017.

[128] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. Extractive and abstractive explanations for fact-checking and evaluation of news. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–50, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4if-1.7. URL https://aclanthology.org/2021.nlp4if-1.7.

[129] Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.256. URL https://aclanthology.org/2020.emnlp-main.256.

[130] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.474. URL https://aclanthology.org/2020.coling-main.474.

[131] Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. 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 2026–2039, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.158. URL https://aclanthology.org/2021.acl-long.158.

[132] Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*, 2021.

[133] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.

[134] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[135] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[136] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf.

[137] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

[138] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A Hale, and Rada Mihalcea. Matching tweets with applicable fact-checks across languages. *arXiv preprint arXiv:2202.07094*, 2022.

[139] Victor Lavrenko and W Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[140] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250, 2008.

[141] Cheng Li, Yue Wang, Paul Resnick, and Qiaozhu Mei. Req-rec: High recall retrieval with query pooling and interactive classification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 163–172, 2014.

[142] Rodrigo Nogueira and Kyunghyun Cho. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1061. URL https://aclanthology.org/D17-1061.

[143] Karthik Narasimhan, Adam Yala, and Regina Barzilay. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2365, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1261. URL https://aclanthology.org/D16-1261.

[144] Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, and Gaurav Singh Tomar. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. *arXiv preprint arXiv:2112.08558*, 2021.

[145] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.44. URL https://aclanthology.org/2021.naacl-main.44.

[146] Eric Malmi, Yue Dong, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, and Aliaksei Severyn. Text generation with text-editing models. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 1–7, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-tutorials.1. URL https://aclanthology.org/2022.naacl-tutorials.1.

[147] Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Edit5: Semi-autoregressive text-editing with t5 warm-start. *arXiv preprint arXiv:2205.12209*, 2022.

[148] Machel Reid and Victor Zhong. LEWIS: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.344. URL https://aclanthology.org/2021.findings-acl.344.

[149] Felix Stahlberg and Shankar Kumar. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.418. URL https://aclanthology.org/2020.emnlp-main.418.

[150] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[151] Stefan Riezler and Yi Liu. Query rewriting using monolingual statistical machine translation. *Computational Linguistics*, 36(3):569–582, September 2010. doi: 10.1162/coli_a_00010. URL https://aclanthology.org/J10-3010.

[152] Aritra Mandal, Ishita K. Khan, and Prathyusha Senthil Kumar. Query rewriting using automatic synonym extraction for e-commerce search. In *eCOM@SIGIR*, 2019.

[153] Davood Rafiei and Haobin Li. Wild card queries for searching resources on the web. *arXiv preprint arXiv:0908.2588*, 2009.

[154] Adi Haviv, Jonathan Berant, and Amir Globerson. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.316. URL https://aclanthology.org/2021.eacl-main.316.

[155] R. Jones and Daniel C. Fain. Query word deletion prediction. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[156] George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://aclanthology.org/H94-1111.

[157] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1341. URL https://aclanthology.org/D19-1341.

[158] Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pages 41–75. Springer, 2018.

[159] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. URL https://doi.org/10.5281/zenodo.4461265.

[160] Yi-Ju Lu and Cheng-Te Li. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.48. URL https://aclanthology.org/2020.acl-main.48.

[161] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.580. URL https://aclanthology.org/2020.emnlp-main.580.

[162] Ashkan Kazemi, Verónica Pérez-Rosas, and Rada Mihalcea. Biased TextRank: Unsupervised graph-based content extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1642–1652, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.144. URL https://aclanthology.org/2020.coling-main.144.

[163] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[164] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

[165] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[166] Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building explainable artificial intelligence systems. In *AAAI*, pages 1766–1773, 2006.

[167] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8163-e-snli-natural-language-inference-with-natural-language-explanations.pdf.

[168] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL https://aclanthology.org/D17-1070.

[169] Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771. URL https://aclanthology.org/2020.acl-main.771.

[170] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623. URL https://aclanthology.org/2020.emnlp-main.623.

[171] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5513. URL https://aclanthology.org/W18-5513.

[172] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[173] Taher H Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

[174] Paweł Budzianowski and Ivan Vulić. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5602. URL https://aclanthology.org/D19-5602.

[175] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 583–592, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.54. URL https://aclanthology.org/2020.acl-main.54.

[176] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

[177] Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. Consistency of a recurrent language model with respect to incomplete decoding. *arXiv preprint arXiv:2002.02492*, 2020.

[178] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[179] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[180] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. Searching for or reviewing evidence improves crowdworkers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, 2(2):26339137231173407, 2023. doi: 10.1177/26339137231173407. URL https://doi.org/10.1177/26339137231173407.

[181] Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[182] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models, 2023.

[183] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 27, 2017.

[184] Jay J Van Bavel and Dominic J Packer. *The power of us: Harnessing our shared identities to improve performance, increase cooperation, and promote social harmony*. Little, Brown Spark, 2021.