

Università degli Studi di Napoli Federico II
Ph.D. Program in
Information Technology and Electrical Engineering
XXXVI Cycle

Thesis for the Degree of Doctor of Philosophy

Knowledge-informed Disinformation Mining: From Fact-Checking to Content Moderation

by
Valerio La Gatta

Advisor: Prof. Vincenzo Moscato

Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

Live as if you were to die tomorrow;
learn as if you were to live forever

Gandhi

Knowledge-informed Disinformation Mining: From Fact-Checking to Content Moderation

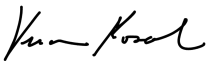
Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information Technology and Electrical Engineering
by

Valerio La Gatta

October 2023

Approved as to style and content by

Prof. Vincenzo Moscato, Advisor



Università degli Studi di Napoli Federico II
Ph.D. Program in Information Technology and Electrical Engineering
XXXVI cycle - Chairman: Prof. Stefano Russo

http://itee.dieti.unina.it

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information Technology and Electrical Engineering is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 11, 2023

Valerio La Gatta

Abstract

In the digital era, the widespread propagation of disinformation presents significant threats to societal, economic, and political stability, a concern underscored by recent global events such as the COVID-19 pandemic. This thesis adopts an integrative approach, combining computer science, network science, artificial intelligence, and knowledge-informed methodologies to tackle online disinformation. Recognizing disinformation as a complex phenomenon, intertwined with human cognition, social dynamics, and emotional responses, the research focuses on leveraging diverse forms of contextual knowledge to combat this challenge.

Focusing on the enduring importance of manual fact-checking processes, we introduce an AI-driven system to expedite fact-checking by utilizing *knowledge from previously fact-checked information*. Additionally, the thesis also presents KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), an innovative methodology that combines internal meme content with *background and cultural knowledge* for harmful meme detection. Furthermore, we explore how simultaneously addressing various disinformation-related tasks, such as fake news detection and sentiment analysis, can bolster overall detection performance and provide a deeper understanding of disinformation content. Lastly, the thesis investigates how *the knowledge of content moderation on a source platform* can inform the moderation strategies of the other social media platforms, enhancing the integrity of the overall digital information ecosystem.

All in all, this thesis advances the understanding of online disinformation and underscores the need for holistic, knowledge-driven approaches to address this pervasive issue.

Keywords: disinformation mining, content moderation, fact-checking, knowledge-informed methods

Sintesi in lingua italiana

Nell'attuale epoca digitale, la diffusione della disinformazione rappresenta una minaccia per la stabilità sociale, economica e politica, una preoccupazione evidenziata da eventi globali recenti come la pandemia di COVID-19. Questa tesi adotta un approccio multidisciplinare, unendo metodologie informatiche, della scienza delle reti, intelligenza artificiale e approcci basasti sulla conoscenza per affrontare la problematica della disinformazione online. Riconoscendo la disinformazione come un fenomeno complesso, intrecciato con la cognizione umana, le dinamiche sociali e le risposte emotive, la ricerca si concentra sullo sfruttamento di diverse forme di conoscenza contestuale per contrastare questa sfida.

Focalizzandosi sull'importanza dei processi di fact-checking, introduciamo un sistema basato su intelligenza artificiale per accelerare il fact-checking utilizzando *la conoscenza derivata da informazioni precedentemente verificate*. Inoltre, la tesi presenta KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), una metodologia innovativa che combina il contenuto interno dei meme con *conoscenze di background e culturale* per il riconoscimento di meme offensivi. Inoltre, esploriamo come l'indirizzo simultaneo di vari task correlati alla disinformazione possa migliorare le prestazioni complessive di rilevamento e fornire una maggiore comprensione del contenuto della disinformazione. Infine, la tesi indaga su come *la conoscenza della moderazione dei contenuti su una piattaforma sorgente* possa informare le strategie di moderazione di altri social media, migliorando l'integrità dell'ecosistema informativo digitale nel suo complesso. In sintesi, questa tesi contribuisce ad avanzare la comprensione della disinformazione online e sottolinea la necessità di approcci olistici per affrontare questo problema.

Parole chiave: disinformation mining, content moderation, fact-checking, knowledge-informed methods

Acknowledgements

Mi trovo alla conclusione del mio percorso di dottorato ma anche di un capitolo importante della mia vita, dove ho imparato tanto ma anche sudato e vacillato come non mai. Prima di proseguire verso nuove avventure, desidero esprimere la mia più profonda gratitudine a coloro che mi hanno aiutato, supportato e sopportato durante questo percorso.

Prima fra tutti, la mia famiglia. Ai miei genitori, Rosy ed Enzo, per non avermi mai fatto mancare nulla ed avermi dato la possibilità di fare tutto ciò che mi appassionasse. A mia sorella Noemi, per avermi incoraggiato giorno per giorno e sostenuto nei momenti di maggiore difficoltá. Un pensiero speciale va anche ai miei nonni. A Nonna Mery, Nonno Peppe e Nonna Annamaria, che mi hanno sempre circondato con il loro affetto e le loro storie. A Nonno Gigi, che non è più con noi, ma la cui memoria continua a guidarmi nei momenti bui. E un affettuoso pensiero va anche agli zii e cugini, per il loro affetto incondizionato.

Ringrazio inoltre il prof. Vincenzo Moscato, la cui guida durante questi tre anni è stata fondamentale per crescere professionalmente e come persona. Un grazie anche al prof. Giancarlo Sperlì per i suoi consigli e insegnamenti.

Come non ringraziare Marco, compagno di innumerevoli sfide, sia durante il percorso di laurea che nel corso di questo dottorato. La tua amicizia e il tuo impegno sono stati per me fonte di ispirazione e di grande aiuto. Insieme abbiamo condiviso stress e successi, spero che il futuro ti riservi tante altre novitá.

Estendo i miei ringraziamenti ad Antonio, Michela, Antonino e a tutti i colleghi del PICUSLab. La solidarietà e l'entusiasmo che ho trovato tra voi sono stati pilastri della mia esperienza di dottorato. Un tributo speciale va al Prof. Antonio Picariello, che, pur non essendo più tra noi, ha lasciato un'impronta indelebile introducendomi nel mondo della ricerca con la mia

tesi di laurea magistrale.

Ringrazio poi tutti i colleghi e amici che ho avuto il piacere di incontrare durante il mio periodo a Los Angeles. Un ringraziamento speciale va al prof. Emilio Ferrara, per aver creduto in me, offrendomi un'opportunità unica che ha arricchito il mio percorso di ricerca e di vita. Un grazie a Luca, Francesco e Goran, per aver condiviso con me la loro preziosa esperienza e conoscenza. Non posso dimenticare di ringraziare Margherita, per la sua compagnia e per i momenti di condivisione che hanno reso questo periodo statunitense stimolante e piacevole. Un pensiero va anche a Phillip, il mio coinquilino, per aver reso la mia prima esperienza di vita lontano da casa indimenticabile. Infine, ma non meno importanti, ringrazio Julie, Alex, Bisjean, Maja e tutti gli altri colleghi e amici per aver contribuito a rendere il mio soggiorno a Los Angeles un'esperienza unica.

Ringrazio poi i miei amici di sempre, Gianluigi, Luca, Valeria, Geremia, Cicca, Nicola, Nunzia, Nunù, Antonio, Lucia, Enrica e Marco, grazie per essere stati la mia rete di sicurezza, per le risate e i consigli, e per tutti i momenti che abbiamo condiviso.

Un ringraziamento particolare va ai miei compagni apneisti (e non solo) della piscina San Mauro. A Enrico, un maestro non solo nell'arte dell'apnea, ma anche nella disciplina e nella pazienza che essa richiede. Ti conosco da non so più quanto tempo e le tue lezioni sono state fondamentali non solo sott'acqua, ma anche nella vita di tutti i giorni. A Giulia ed Antonio, che con il loro entusiasmo contagioso per il mare hanno trasformato la mia passione in un amore profondo per l'immensità del blu. E non posso dimenticare Filippo, Anna, Eva, Giangi, Rino e Giggino con cui ho condiviso tantissime immersioni e momenti preziosi che rimarranno sempre impressi nella mia memoria.

Infine, un grazie speciale a Maria, che ha condiviso con me gli ultimi due anni, supportandomi e sopportandomi nonostante tutto. La sua presenza è stata fondamentale in questo viaggio.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the digital age, the rapid dissemination of information through online platforms has transformed the way we consume and share knowledge. Yet, this revolution in information exchange is a double-edged sword, culminating in a surge of disinformation – a phenomenon where false and misleading narratives are deliberately crafted and disseminated with malevolent intent. The ramifications of disinformation permeate societal, economic, and political realms, casting shadows over electoral processes [16], increasing polarization [290], and engendering public health crises [41]. Notably, the misinformation maelstrom during the COVID-19 outbreak, which questioned mask efficacy, exacerbated transmission risks [156], while spurious vaccine narratives fueled hesitancy [149]. In the geopolitical arena, the Ukraine-Russia standoff highlighted the potential of online disinformation campaigns to reshape narratives, either by dubbing the conflict a special operation against alleged Nazis or attributing it to NATO expansion dynamics [92, 197].

Deciphering the drivers of this disinformation epidemic is paramount. Contemporary digital infrastructures, shaped by an intricate interplay of technological, societal, and cognitive determinants, serve as conducive environments for the propagation of disinformation [57]. The ubiquity of platforms such as social media and instant news portals has rendered global communication seamless [122]. Yet, this democratization harbors pitfalls as it allows bypassing traditional journalistic standards with no third-party verification or editorial oversight for online content [11].

Furthermore, the inherent human cognitive biases such as confirmation bias [174] and the illusory truth effect [64] result in individuals gravitating towards, and blindly accepting, information that aligns with their existing beliefs. Additionally, in our post-truth era, where emotive narratives often eclipse factual rigor [163], discerning truth becomes even more challenging. This complexity, coupled with inadequate media literacy [160], renders people more vulnerable to unquestioningly embrace disinformation. Financial incentives further muddy the waters, with click-bait content and state actors propaganda adding layers of complexity [222, 169, 263, 187].

Last but not least, tackling online disinformation also begets ethical quandaries associated with content moderation. Striking a balance between filtering out falsehoods, while preserving free speech and preventing inadvertent suppression of legitimate discourse, demands scrupulous judgment [121]. These efforts, whether powered by algorithms or human intervention, are not immune to unintended biases, thereby eroding trust in online platforms [89].

This disinformation conundrum explains the latest explosion of interest around issue of disinformation, misinformation and other forms of harmful content spreading in digital environment. The research community needs to confront with multifaceted challenges: not only the sheer volume of misleading content and the nuanced strategies employed by adversarial agents to replicate authentic sources, but also the limitations instituted by primary information platforms. These platforms, in their operations, manifest a notable opacity in content moderation strategies and have recently imposed stringent data accessibility restrictions, hampering academic research.

Traditionally, the research community has approached the detection and mitigation of disinformation predominantly through algorithmic methodologies. These methods, often rooted in traditional machine learning or more contemporary deep neural networks, predominantly utilize content-based features, such as textual or visual content, user profiles attributes and source credibility information. Yet, a significant limitation of these efforts has been their reliance on small, ad-hoc datasets, leading to models that may not generalize effectively in real-world, dynamic environments.

Conversely, large-scale analyses have shed light on the disinformation

dynamics, probing into the distinctive roles of entities such as social bots, trolls, and public figures. These observational studies further underscore the amplifying effects of echo chambers in perpetuating both truth and untruth and the paradoxical existence of specific digital spaces where harmful content is intentionally not moderated.

Collectively, this highlights a tangible divergence between developing state-of-the-art detection methodologies – for identifying either deceptive news or malicious actors – and grasping the nuanced mechanisms of disinformation dissemination in the digital sphere.

1.1 Contributions of the Thesis

This thesis aims to contribute a holistic, scientifically rigorous approach to combat the pervasive challenge of disinformation by adopting a multifaceted approach, blending computer science and artificial intelligence methodologies. Within this landscape, the term "disinformation" encompasses the wide spectrum of false, misleading, and potentially hateful content proliferating across the digital information ecosystem. In particular, we integrate several thematic threads, each dedicated to addressing pivotal issues related to the detection and mitigation of disinformation. These threads utilize a blend of AI techniques and knowledge-informed strategies to achieve two complementary objectives: while the development of state-of-the-art predictive models remains a cornerstone, the research equally focuses on a holistic comprehension of the disinformation phenomenon. Our research encompasses the analyses and integration of several forms of *contextual knowledge*. This includes (i) databases of fact-checked information for accurate information verification, (ii) cultural and common-sense knowledge to understand complex information piece, (iii) computational patterns from various disinformation-related tasks to improve disinformation detection and (iv) platform-specific interventions to craft cross-platform moderation strategies.

Our contributions can be categorized and summarised as follows:

Fact-Checking Recognizing the enduring role of manual fact-checking, our investigation delves into the utility of identifying previously fact-checked information before embarking on a more in-depth verification pro-

cess. Specifically, we introduce a novel AI-driven information retrieval mechanism capable of managing this task across multimodal data, integrating both textual and visual components. Our results underscore the capability of the system to expedite the fact-checking process, especially for pivotal and ongoing geopolitical events, as illustrated with a case study on the Ukraine-Russia conflict.

Harmful Content Detection Modern disinformation detection grapples with the challenge of complex multimodal information objects, such as internet memes and short videos. These objects ingeniously blend visual and textual elements, harnessing cultural references and social context to propagate deceptive narratives or offensive messages. In response, we propose KERMIT – Knowledge-EmpoweRed Model In harmful meme deTection – a pioneering approach that synergizes common-sense knowledge with memory-augmented neural networks for comprehensive meme understanding. Central to KERMIT is the construction of a knowledge-enriched information network tailored for each meme, integrating its internal entities with relevant background knowledge. Our results demonstrate KERMIT's proficiency in detecting harmful memes as well as the assimilated knowledge further facilitates an assessment of the meme's congruence with established facts and logical consistency.

Knowledge Transfer across Disinformation Tasks Disinformation campaigns deftly harness emotional triggers and prey on cognitive biases. To combat such sophisticated manipulations, it is imperative to leverage the inter-relatedness of diverse disinformation-related tasks, such as sentiment analysis, stance identification and topic detection. In light of this, we present an interpretable multi-task learning framework to explore the dynamics of inter-task knowledge transfer in disinformation detection. Our results not only reveal multi-task performance improvements but also underscore that this enhancement is not merely attributed to computational factors such as data augmentation or data similarity. Instead, our analysis indicates that the enhancement arises from the absorption of supplementary patterns across tasks, highlighting that deceptive narratives intricately meld sentiment cues, stance markers, and topical elements.

Cross-Platform Moderation Disinformation is seldom an isolated phenomenon; rather, it flourishes within online communities, often exacerbated by echo chambers spanning multiple social networks. To break this cycle, we extend our strategies to encompass cross-platform moderation interventions. By examining Twitter discussions surrounding harmful and moderated YouTube content, we unveil the interconnected nature of online harm and moderation. Notably, we discern that Twitter users disseminating harmful YouTube content frequently advocate extreme and conspiratorial ideologies, culminating in subsequent suspensions from Twitter. This underscores the potential benefits of sharing moderation interventions across different social media platforms within the information ecosystem.

Impact and Broader Perspective Together, these contributions offer a broader perspective on the intricacies of disinformation in the digital age. By weaving computational techniques with nuanced understandings of human cognition and behavior, our research endeavors to pioneer holistic strategies that are not only effective in detecting and mitigating disinformation but also insightful in understanding its multifarious roots. The implications of these advances reach beyond academia, informing technology developers, policy-makers, and digital platform designers on best practices to safeguard information ecosystems.

1.2 Structure of the Thesis

In Chapter 2, we provide the reader with an in-depth analysis of the disinformation research landscape. We meticulously explore the principal drivers catalyzing the proliferation of disinformation and elucidate the inherent challenges of addressing disinformation detection without a comprehensive perspective. Subsequently, we present the *knowledge-informed disinformation mining* paradigm, advocating for a holistic methodology in the battle against disinformation. Concluding the chapter, we highlight the pivotal role of AI in this field, detailing its application in disinformation identification and mitigation. Additionally, we discuss the potential risks associated with the deployment of modern AI tools, emphasizing their ambivalent nature in the context of disinformation.

In Chapter 3, we investigate the optimization of the fact-checking pro-

cess through the identification of previously fact-checked information. We commence by formally defining the challenge as an information retrieval task and subsequently survey pertinent literature in the domain. Then, state-of-the-art retrieval models are benchmarked, with a focused exploration into the synergistic potential of merging traditional information retrieval techniques with advanced deep learning frameworks. Stemming from this analysis, we introduce a novel multimodal system which achieves state-of-the-art performance for this task. The chapter concludes by assessing the strategic advantage of recognizing fact-checked data during the preliminary phases of geopolitical incidents, with the Ukraine-Russia conflict serving as a contextual case study. The material presented in the chapter is mostly based on the following articles:

- Chakraborty Tanmoy, **La Gatta Valerio**, Moscato Vincenzo, Sperli Giancarlo; Information retrieval algorithms and neural ranking models to detect previously fact-checked information; Neurocomputing 2023; DOI: 10.1016/j.neucom.2023.126680

- Formisano Raffaele, **La Gatta Valerio**, Moscato Vincenzo, Sperli Giancarlo; Fact-checked Information Retrieval using Multimodal Machine Learning; under review

- **La Gatta, Valerio** and Wei, Chiyu and Luceri, Luca and Pierri, Francesco and Ferrara, Emilio; Retrieving false claims on Twitter during the Russia-Ukraine conflict; Companion Proceedings of the ACM Web Conference 2023; DOI: 10.1145/3543873.3587571

In Chapter 4, we pivot towards the domain of multimodal disinformation detection, with a specific focus on identifying harmful memes. Following a meticulous review of existing literature in this field, we present KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), a pioneering methodology designed to integrate the intrinsic meme entities with external, common-sense knowledge. We conclude the chapter by presenting quantitative results that substantiates KERMIT's proficiency in retrieving contextual knowledge and employing it effectively for classification, thereby amplifying predictive performance. The material presented in the chapter is mostly based on the following article:

- Grasso Biago, **La Gatta Valerio**, Vincenzo Moscato, Giancarlo Sperlì; KERMIT: Knowledge-EmpoweRed Model In harmful meme deTection; under review

In Chapter 5, we embark on a comprehensive examination of knowledge transfer across an array of disinformation-related tasks. Building upon a detailed review of recent literature, we introduce a multi-task learning paradigm, designed to elucidate the intricacies of knowledge interchange among different tasks, including sentiment analysis, fake news detection, stance identification, and topic detection. The chapter culminates with quantitative results, shedding light on the conditions and underlying mechanisms fostering positive knowledge transfer. The material presented in the chapter is mostly based on the following articles:

- **La Gatta Valerio**, De Cegli Luigi, Vincenzo Moscato, Giancarlo Sperlì; From Single-Task to Multi-Task: Unveiling the Dynamics of Knowledge Transfers in Disinformation Detection; under review

- **La Gatta, Valerio** and Moscato, Vincenzo and Postiglione, Marco and Sperli, Giancarlo; COVID-19 Sentiment Analysis Based on Tweets; IEEE Intelligent Systems 2023; DOI: 10.1109/MIS.2023.3239180

In Chapter 6, we undertake a rigorous exploration of the cross-platform propagation of harmful YouTube content. Emphasizing the pressing need for collaborative moderation interventions across platforms, we subsequently elucidate our empirical analyses of Twitter discourses linked to moderated YouTube videos, specifically contextualized within the 2020 US election period. The material presented in the chapter is mostly based on the following article, which was nominated for the ACM Hypertext Ted Nelson Award 2023:

- **La Gatta, Valerio** and Luceri, Luca and Fabbri, Francesco and Ferrara, Emilio; The Interconnected Nature of Online Harm and Moderation: Investigating the Cross-Platform Spread of Harmful Content between YouTube and Twitter; ACM Hypertext 2023; DOI: 10.1145/3603163.3609058

Finally, in Chapter 7, we provide a comprehensive synthesis of the research findings and contributions delineated throughout this thesis. Additionally, we elucidate the overarching implications of our work and delineate prospective avenues for further exploration in the realm of disinformation studies.

Chapter 2

Towards Knowledge-Informed Disinformation Mining

2.1 Terminology

In the past, terms like disinformation, misinformation, fake news, rumor, and propaganda have often been entangled, leading to confusion and misinterpretation. Different forms of incorrect information were frequently lumped together, blurring the lines between various types and hindering effective management strategies.

Recognizing the need for clarity, there have been several concerted efforts across academia, industry, and government bodies to standardize the definitions of key terms related to disinformation. These efforts aim to distinguish subtly different types of information that are often erroneously grouped together. The objective is to create a common language that can be universally understood and applied, facilitating more effective strategies for identification, analysis, and mitigation of these information-related threats.

Stakeholders in the field have now largely converged on definitions that hinge on the purpose and context of the information in question. Notably, the Cybersecurity and Infrastructure Security Agency (CISA) offers precise and distinct definitions[1], foundational to the discourse in knowledge-informed disinformation mining:

[1] https://www.cisa.gov/sites/default/files/publications/mdm.pdf

- Misinformation: Defined as *false information not intended to cause harm*. Its origins often lie in miscommunication, erroneous interpretation, or dissemination errors. This type of information proliferates in the digital age, as the rapid circulation of information online frequently bypasses rigorous verification.

- Disinformation: Identified as *intentionally fabricated to deceive, harm, or manipulate*. The crafting of disinformation often employs elaborate strategies to appear credible, posing challenges in detection and response. Such campaigns, often utilized in cyber warfare and political propaganda, can have profound socio-political effects.

- Malinformation: This involves *fact-based information used out of context to mislead, harm, or manipulate*.

Understanding these definitions is vital in the field of disinformation mining. Each type presents unique challenges in detection and mitigation, necessitating tailored approaches. For instance, while misinformation might be countered through fact-checking and public awareness campaigns, disinformation requires a more complex strategy involving the identification of malicious intent, often hidden behind sophisticated masking techniques. Similarly, combating malinformation requires a deep understanding of the context and the ability to discern the subtleties in the presentation of factual information.

2.2 Disinformation Drivers

News Consumption and Social Networks The dissemination of false information, although rooted in the early days of the free press, has evolved considerably in the digital era. Marquis de Condorcet's "Outlines of an Historical View of the Progress of the Human Mind" [44] references a critique by U.S. President Adam Smith – "There has been more new error propagated by the press in the last ten years than in an hundred years before 1798". Yet, the complexities of false narratives have extended beyond inaccurate reporting, encompassing financial narratives to coordinated influence campaigns [129].

Modern social media platforms, including Facebook, Twitter, YouTube, Instagram, and TikTok, have fundamentally altered news dissemination and consumption. The accessibility and allure of these platforms, combined with declining barriers to content creation, have democratized information dissemination. However, this expansion also ushers in concerns about the quality and accuracy of information [11]. Furthermore, as trust in traditional media is rapidly declining, there is an increasing spread of fake news on these online platforms [129].

The complexity is further heightened by digital agents, including social bots, cyborgs, and trolls. Their coordinated efforts magnify the reach and penetration of misleading narratives [237]. Additionally, while mainstream platforms have diversified their multimedia offerings, they exhibit varied moderation and data access policies [78]. In parallel, the emergence of fringe low-moderated platforms, such as Gab, Bitchute and 4chan, serves as echo chambers for extremist ideologies [267, 314], adding another layer to the intricate tapestry of digital disinformation.

Human Factors The influence of culture and demographics on an individual's perception and interaction with information cannot be understated. Distinct cultural backgrounds and demographic attributes contribute to unique cognitive frameworks, which, in turn, affect how both true, false, and potentially harmful content is received and interpreted [100, 29].

Central to this discourse is the concept of *naive realism*, the belief that individuals perceive the world objectively [296]. Those subscribing to this notion tend to believe that people with opposing views are either uninformed, irrational, or biased. This self-assured perspective often means that conflicting information is casually dismissed, irrespective of its veracity. Complementing *naive realism* is the widely acknowledged *confirmation bias*, where individuals have a tendency to seek out, interpret, and remember information that aligns seamlessly with their pre-existing beliefs [188]. Such tendencies can hinder a balanced evaluation of new or contrasting information.

Beyond individual biases, social dynamics also play a crucial role in information processing. The *normative social influence* describes the human tendency to conform to group expectations in order to be accepted [51].

Even if individuals internally disagree, the external pressure to align with group norms might lead them to endorse or disseminate certain viewpoints. This is further reinforced by the *social identity theory*, which posits that individuals categorize themselves into specific social groups [13]. Invariably, beliefs and attitudes of the in-group are viewed as superior compared to those of the out-groups [269]. This demarcation can further skew the consumption and distribution of information.

Another challenge in the modern digital landscape is *information overload*. Individuals are constantly bombarded with an overwhelming volume of information, making comprehensive processing challenging [59]. Often, this can result in reliance on shortcuts in decision-making or a preference for misleading or sensationalized content.

These individual and societal tendencies pave the way for the emergence of echo chambers, environments where homogeneity of opinion reigns supreme [50, 277]. Within these chambers, similar beliefs and views are perpetuated, with little to no exposure to alternative perspectives, further entrenching biases.

Furthermore, the digital tools and platforms of our era introduce additional complexities in the form of the *algorithmic bias*. As social media platforms aim to enhance user engagement through content personalization, they may inadvertently amplify the echo chamber effect [42], further distorting the information landscape. Pertinently, recent investigations, particularly focused on YouTube, have delved deeply into the influence of recommendation algorithms in directing users towards more extreme or biased content [218, 130, 61], further complicating the digital information ecosystem.

2.3 Beyond the Deficit Model

Historically, the challenge of disinformation was interpreted through the lens of the deficit model in science communication [253]. This model was grounded in the belief that rectifying misconceptions could be achieved merely by supplying accurate information [39]. However, this model's shortcomings became evident, as it did not sufficiently address the complex psychological, emotional, and social factors that contribute to the allure of disinformation [57].

In response to the multifaceted challenges posed by the proliferation of disinformation and its societal ramifications, stakeholders across the digital information ecosystem – including academics, policymakers, technology platforms, civil society groups, and the media – have collectively gravitated towards a more holistic paradigm, which we term *knowledge-informed disinformation mining*. Rather than viewing disinformation as mere isolated instances of false or harmful content, this approach aims to understand the complex drivers underpinning an individual's or collective's susceptibility to misleading narratives. It recognizes that the factors influencing the assimilation and spread of such narratives often transcend a mere deficit of accurate knowledge, encapsulating deeper cognitive and sociocultural intricacies.

Furthermore, this broader conception of disinformation mining acknowledges the challenge as extending beyond merely identifying harmful content, malicious actors or susceptible users. It perceives the issue as embedded within a larger ecosystem where information, motives, platforms, and stakeholders intermingle. To navigate this challenging terrain requires a comprehensive strategy. Effectively navigating this complex scenario necessitates a multifaceted strategy. While the foundation remains solid with traditional fact-checking, there is an imperative to integrate wider governance tools, such as the EU Code of Practice[2] and the US Disinformation Governance Board[3]. Moreover, collaborations with platform-specific initiatives, like the Twitter Moderation Research Consortium[4] and Meta's Integrity and Transparency Reports[5], are crucial. This collective endeavor aims not only to mitigate the reach and impact of disinformation but also to cultivate a more informed and critical digital populace.

The Role of Knowledge In the paradigm of *knowledge-informed disinformation mining*, knowledge emerges as a paramount tool in countering deceptive content. This framework integrates resources such as fact-checking databases, which act as vaults of corroborated information, bolstering the identification and refutation of erroneous claims [88]. It also

[2]https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
[3]https://www.dhs.gov/publication/disinformation-governance-board
[4]https://transparency.twitter.com/en/reports/moderation-research.html
[5]https://transparency.fb.com

employs common-sense reasoning, influenced by cross-cultural nuances, user-specific background knowledge, and general societal understanding, to evaluate the logical consistency and veracity of information. This nuanced approach acknowledges the varying interpretations of content based on cultural contexts and individual experiences. By considering these factors, it crafts a more robust shield against disinformation that deftly manipulates both textual and graphical elements, potentially mitigating the risk of harmful or biased content that might otherwise be overlooked in a singular cultural or knowledge framework.

Moreover, this methodology promotes cross-platform synergies. Insights derived from moderation interventions on a particular platform can be extrapolated to influence strategies on another, enhancing the efficacy of interventions [127]. It is imperative to understand that information does not exist in isolation but it is nested within a dense network of context, relationships, and user activities. Consequently, knowledge takes on a dynamic dimension, materializing as contextual information intricately intertwined with variables such as temporal factors and user behaviors [329]. Temporal aspects involve meticulous tracking of when and how information emerges, evolves, and exerts its influence over time [286, 258]. User behaviors entail the astute consideration of distinct user archetypes, including bots or trolls [162], whose coordinated efforts can fuel disinformation campaigns [191].

Understanding the Emotional Landscape The emotional dimensions of content significantly influence the dynamics of disinformation dissemination. Expressions of anger can fuel divisive discourse [184] and intensify social polarization [72], while content designed to incite excitement can go viral, rapidly spreading across networks [215]. Knowledge-informed methodologies delve into these emotional catalysts, furnishing both analysts and users with the tools to discern and counterbalance the emotive lures driving the propagation of harmful content [76, 75]. Moreover, in the context of the post-truth era, the disentanglement of emotional manipulation from factual information assumes paramount importance as it is instrumental in enhancing media literacy, particularly within the realm of social media platforms [133].

Social Dynamics and Resilience Building The prevalence of disinformation is often accentuated within the confines of social networks and resonant ideological spaces. A key facet of knowledge-informed disinformation mining is the rigorous examination of societal interconnections, the dynamics of peer-driven influence, and collective behavioral patterns. Delving into the mechanisms of peer influence is paramount, elucidating the modalities by which individuals' perceptions and beliefs are shaped by their immediate social connections [56, 232]. Informed by such insights, interventions can be designed to harness constructive peer influences, thereby enabling individuals to withstand coercive nudges towards accepting spurious information [328]. In tandem, the knowledge-informed approach probes the intricacies of collective behavior, discerning how attributes like group cohesion, shared identity, and mutual values can potentiate the reception and propagation of misleading narratives [167, 19]. This understanding guides the creation of strategies aimed at promoting open discourse, fostering critical thinking, and disrupting echo chambers [231, 330, 275].

2.4 The Role of Artificial Intelligence

Over the past decade, Artificial Intelligence (AI) and Machine Learning (ML) have become pivotal in countering disinformation. These technologies not only enhance automatic detection mechanisms, pinpoint malicious actors and early radicalization indicators, but they also play a crucial role in disinformation mitigation strategies. Moreover, they furnish researchers and regulators with robust tools to quantitatively analyze the mechanisms behind online disinformation, from the role of social bots and echo chambers to the demographics of users who are more vulnerable to deceptive content.

2.4.1 Automatic Detection of Online Disinformation

Historically, automated disinformation detection has been conceptualized as a supervised classification challenge. Depending on the nature of the disinformation, a piece of news can be categorized as true or false [329], offensive or benign [69], and propagandistic or neutral [181]. The classifica-

tion scheme could unfold as multi-class to encapsulate a finer granularity of disinformation (e.g. partial false news) [179, 291], or multi-label to capture overlapping disinformation signals (e.g., multiple persuasion techniques) [53, 68]. Annotation processes predominantly harness evaluations from fact-checking news outlets [256, 247], whereas the identification of offensive or discriminatory content often necessitates direct manual annotation by researchers [221, 14].

The data spectrum exploited for crafting classification algorithms and models includes information pertinent to the scrutinized news article and its accompanying social media shares. This data landscape prompted the adoption of two principal detection methodologies [248]: content-based and context-based techniques.

Content-based methodologies delve into the explicit content of news articles, encapsulating elements such as text, imagery, and videos [8], alongside the linguistic style employed (e.g., emotive, persuasive, reportorial) and the credibility of the disseminating source or promoting group [248]. Conversely, context-based methodologies are anchored in comprehending the environment and conditions enveloping content dissemination rather than the content per se [329]. These methodologies accentuate user interactions like likes, comments, and re-shares [268], and leverage diverse data structures like news propagation cascades, temporal networks, and self-defined graphs depicting news propagation in conjunction with user-related information [329].

Within this realm, the research community has primarily ventured into representation learning techniques, aspiring to furnish vector representations, or embeddings, of semantically complex content [248, 168]. Initially, content-based methodologies predominantly fixated on textual content, utilizing statistical methodologies for embedding, such as n-grams and TF-IDF [325, 250]. In contrast, context-based techinques drew inspiration from observational studies [286, 26] delineating the discrepancies in false and true news propagation, centering on the extraction of hand-crafted features [328] from these networks (e.g., user count in a propagation cascade, maximal depth or breadth of the propagation cascade, average time span between consecutive shares) or crafting graph kernels to ascertain graph similarity [301].

Upon obtaining the news representation, conventional machine learn-

ing classifiers like logistic regression, decision trees, and support vector machines (SVMs) were the preferred options for classification tasks [329].

Nonetheless, the emergence of pre-trained deep neural networks ushered in efficacious and automatic modalities to represent complex information [23, 319]. Specifically, these pre-trained networks were engineered to concurrently refine the (pre-trained) representation and tackle the classification task in an end-to-end fashion [327, 111, 293]. Content-based methodologies began to extend their focus beyond text to more complex entities like images, via convolutional neural networks (CNNs), and videos, through recurrent neural networks (RNNs) [8]. Similarly, context-based approaches aimed directly at the news propagation graph, via graph neural networks (GNNs).

Overall, the experimental evidence suggests that while content-based methodologies attain superior detection performance, their scope remains constrained. On the flip side, context-based approaches exhibit broader generalization capabilities [249, 186]. Consequently, contemporary research is steering towards hybrid methodologies that amalgamate the merits of both solutions. This fusion is realized through the creation of increasingly sophisticated data structures (e.g., hypergraphs, spatio-temporal graphs) encompassing the news content, the user network sharing the news, and the news propagation cascade [258, 266, 102].

Finally, the recent advancements in self-supervised models, pre-trained on extensive datasets, have redirected research emphasis towards more nuanced challenges beyond the disinformation classification task. For content-based techniques, these challenges encompass endeavors such as detecting inconsistencies between textual and visual elements [242, 327] or unveiling manipulated media [238, 101]. For context-based techniques, challenges span areas like estimating unobserved propagation networks [104] and differentiating between intentional and unintentional disinformation spreaders [326].

Notably, there is a sustained initiative to discern methods to augment model generalization across diverse domains (e.g., politics, healthcare, entertainment) [251, 183, 54] and, on a broader scale, tackle disinformation detection challenges under few-shot settings [320, 132]. The overarching narrative also showcases a burgeoning interest in the architectural innovation of interpretable detection systems [152, 246, 15, 1]. These systems

aim not only to execute the classification task adeptly but also to elucidate the rationale underpinning the classification decisions.

2.4.2 Mitigation of Online Disinformation

In addition to introducing new policies and regulations, efficiently blocking and mitigating the spread of fake news also demands technical innovations brought by AI-empowered tools.

Historically, the sphere of online disinformation mitigation pivoted around the tactic of integrating true news within a social network. In this context, traditional information diffusion models [310], such as the Independent Cascade and Linear Threshold models [109], have been employed to optimize the spread of factual information and contain the proliferation of false narratives. The subsequent evolution of this field introduced sophisticated models entrenched in multivariate point process theories and reinforcement learning. These techiniques were tailored to discern the differences in the propagation of diverse news [63] as well as to capture the unique consumption patterns of individual users [6], particularly accounting for their varying exposure levels to disinformation [7]. Notably, an inherent challenge with these methodologies is the prerequisite identification and continual monitoring of fake news trajectories across networks, a task that remains non-trivial. Consequently, the prevalent mitigation method has thus capitalized on crowd-sourced mechanisms, facilitating users to actively report false news or potential policy breaches within digital platforms [115, 279].

More recently, academic pursuits are converging towards the development of custom recommender systems that account for user susceptibility to disinformation. Proactive measures can be taken by personalizing content suggestions in accordance with a user's radicalization level [61, 99]. Conversely, reactive approaches steer users towards high-quality content subsequent to their engagement with harmful materials [147].

Finally, a burgeoning corpus of research emphasizes the importance of disinformation correction, where entities (regular users or organizations) counteract misleading content by underscoring its inaccuracies and citing fact-validated articles. In this vein, [284] proposed a URL recommendation algorithm to incentive proactive Twitter users, named "guardians", to disseminate fact-checked data to combat online misinformation. Despite the

intuitive benefits, empirical studies present mixed results on correction efficacy. Some research [233, 287, 288] suggests that corrective interventions can significantly reduce belief in misinformation, while others [176, 105] indicate potential adverse effects, such as the backfire effect, where corrections inadvertently reinforce misinformation. In particular, the effectiveness of disinformation correction may depend on factors like the source's credibility and the relationship between the disinformation spreader and the corrector.

2.4.3 The Double-Edged Sword: AI as a Vector for Disinformation

In the quest to leverage AI for disinformation detection and countermeasures, recent technological strides in Generative AI have simultaneously cast light on the potential threats brought by these advancements [66]. This dichotomy is not novel but rather represents a modern manifestation of age-old issues of deception and manipulation. Indeed, the historical trajectory of AI has always been shadowed by the potential for its dual use, where technological advancements intended for progress and enlightenment also open doors to sophisticated forms of deceit. The current focal point of this duality lies in the capabilities of deep fakes and Large Language Models (LLMs) such as GPT [211], LLaMa [278] and DALL-E [214].

These foundation models, for instance, can be exploited to generate scam emails [199] or entire (fabricated) news articles[6] that fuel disinformation campaigns. These campaigns employ a blend of factual and fallacious elements to craft messages that convincingly malign public figures, potentially undermining public trust in digital communications [18]. The persuasive power of LLMs lies in their ability to produce content that closely mimics human writing, allowing for the propagation of personalized and misleading narratives. Deep fakes further exacerbate this issue by creating hyper-realistic videos, making individuals appear to say or do things they have not, thus compromising the integrity of digital content [185].

In addition, the sophistication of LLMs in imitating writing styles and

[6] https://www.newsguardtech.com/misinformation-monitor/august-2023

evading detection creates significant hurdles for the verification of information. They enable the mass production of varied, misleading content, rendering traditional fact-checking processes insufficient. Moreover, the ease of automating fake profiles with these AI technologies raises concerns about their potential to distort online conversations and societal norms [308].

On the flip side, even LLMs conceived for constructive purposes may inadvertently distribute false information [201] due to the limitations of their training datasets, model biases, or the creation of baseless content – commonly known as "hallucinations" [103]. This can lead AI-driven interfaces to produce flawed advice or information, posing risks in domains where accuracy is paramount, such as finance or healthcare. The real danger lies in the potential for users to trust, share, and act upon such AI-generated misinformation, thus amplifying its impact [316].

2.5 Discussion

In this chapter, we have established a foundational understanding of the disinformation landscape, setting the stage for the deeper analysis that follows in this thesis.

Our exploration began with a clear definition of key terms in this field, specifically misinformation, disinformation, and malinformation. We then provided a historical perspective, tracing how disinformation has evolved from the early days of free press, as critiqued by figures like U.S. President Adam Smith, to the complex and pervasive entity it is in the digital age. The proliferation of global digital platforms has broadened the reach and sophistication of disinformation campaigns, impacting society at multiple levels. The rise of social media has granted unprecedented access to information, yet this has come with the heightened challenge of discerning truth from falsehood. This dilemma is intensified by the proliferation of bots and unregulated platforms, which often serve as conduits for extremist views and misinformation. Additionally, ingrained human biases and cultural factors contribute to the perpetuation of disinformation, leading to echo chambers that resist outside scrutiny.

At its core, the chapter highlighted the limitations of the information deficit model, which assumed that misinformation can be corrected

by merely providing accurate information. In its place, the holistic approach of knowledge-informed disinformation mining has been presented, which emphasizes understanding the psychological, emotional, and societal drivers behind susceptibility to disinformation.

AI and ML have surfaced as double-edged swords in this narrative. They hold the potential to significantly advance the detection and mitigation of disinformation, yet they also pose new risks. The threat of deep fakes and the misuse of LLMs illustrate the darker possibilities of these technologies, highlighting the need for an ethical framework for AI development and usage.

As the thesis moves forward, it delves into defining knowledge-informed methodologies that address knowledge-informed disinformation mining across various dimensions. It positions AI as a force for good, sharpening our fact-checking and content analysis capabilities. By avoiding AI's malicious uses, the thesis advocates for its informed and ethical application, essential for constructing proactive defenses against the ongoing surge of digital misinformation.

Chapter 3

Detecting previously fact-checked information

3.1 Research Context and Problem Definition

The advent of social media has brought about a transformative shift in human communication, providing individuals with expedited channels to express their viewpoints and emotions. This shift is especially pronounced in the context of the digital information ecosystem, where sharing occurs swiftly and widely, often facilitated by multimodal content like memes and short animated frames, which have demonstrated heightened appeal and perceived credibility [90]. However, the misuse of this freedom of expression, often referred to as disinformation, has supported fake news dissemination to mislead people decisions and has encouraged hostility behaviors in the form of hate speech and cyber-bulling [9].

Whilst representatives of online platforms, leading social networks and advertising industry, also in accordance to new governmental policies[1], are increasingly adapting to mitigate this problems, fact-checking still represents the leading strategy to debunk false information through domain experts' analyses and semi-automatic systems assessing news truthfulness [88]. Indeed, the Duke Reportes' Lab[2] counts more than 400 fact-checking

[1] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

[2] https://reporterslab.org/fact-checking/

world-wide organizations which focus on specific domains (e.g. *politifact.com*) or social networks (e.g. *snopes.com*).

In particular, the fact-checking process comprises a four-stages pipeline [180]: (i) claim detection, i.e. selecting and prioritizing content according to its importance and relevance; (ii) verified claim retrieval, i.e. detecting previously fact-checked information; (iii) evidence retrieval, i.e. finding the evidences which support or refute a claim; (iv) claim verification, i.e. assessing claim's veracity (even partially) based on the retrieved evidences.

Considering the proliferation of fact-checked news and the frequent reposting of viral claims, the verified claim retrieval task offers a promising avenue for enhancing the fact-checking process. For example, during the third presidential debate preceding the 2016 US Presidential Election, Donald Trump stated that "$6 billion went missing at State Department under Clinton." Politifact's assessment[3] not only verified the claim but also explicitly mentioned that it had been previously debunked on the platform. Thus, identifying previously fact-checked information can streamline the manual fact-checkers' workflow by filtering out content that has already been verified. This not only enhances efficiency but also ensures the delivery of pertinent and trustworthy information, ultimately boosting their productivity and effectiveness.

Formally, the problem of detecting previously fact-checked information can be phrased as an information retrieval task, i.e., retrieving and re-ranking a list of verified documents according to their relevance with an input claim. Previous work has predominantly tackled this task by focusing on textual content and re-ranking techniques [235, 244, 285], often employing intricate models to re-assess a limited set of verified documents. These efforts typically commence with standard information retrieval algorithms like BM25 [220] for the initial document retrieval step. Nevertheless, recent advancements, particularly those harnessing pre-trained language models, are introducing promising alternatives to traditional sparse representation-based retrieval and re-ranking approaches.

Furthermore, the majority of prior research pertaining to the verified claim retrieval task has concentrated on its isolated evaluation using benchmark datasets. However, these assessments frequently lack an examination

[3] https://www.politifact.com/factchecks/2016/oct/20/donald-trump/trump-wrongly-says-6-billion-went-missing-state-de/

of the practical viability of the task in real-world contexts. In such scenarios, not only does the volume of verified information undergo dynamic changes over time, but the task also interfaces with multiple other phases within the fact-checking pipeline. This necessitates a more comprehensive evaluation that considers the task's real-world applicability and its interplay with the evolving information landscape.

Given these limitations, our contributions can be summarized as follows:

- Considering the information retrieval nature of the verified claim retrieval task, we conduct an extensive benchmark of the most recent methods and models from information retrieval and question answering literature to unveil their practical relevance for the task under analysis [32].

- We curate a dataset comprising 83 false claims that proliferated on Twitter during the initial weeks of the Ukraine-Russia conflict. Subsequently, we manually annotate 5,872 original tweets to ascertain whether they discuss any of these false claims, providing invaluable data for task evaluation. We then engineer an automated pipeline, primarily based on the verified claim retrieval task, to identify and retrieve tweets that engage with any of the 83 false claims within our dataset [127].

- We propose a simple, yet effective, multimodal retrieval system based on modern visual-language models to detect previously fact-checked information. In particular, our design can be deployed seamlessly under retrieval and re-ranking settings.

- Our empirical findings underscore the practical applicability of traditional and neural methodologies drawn from pertinent literature when it comes to detecting previously fact-checked information. Furthermore, our proposed multimodal systems establish themselves as state-of-the-art performers in both retrieval and re-ranking tasks. Ultimately, we demonstrate that the verified claim task effectively identifies already debunked false claims that disseminated on Twitter during the Ukraine-Russia conflict.

3.2 Related works

3.2.1 Fact-checking Panorama

The fact-checking problem, which involves assessing the truthfulness of a claim, has been a subject of extensive research across various contexts. Recently, there has been a growing focus on evidence-aware fact-checking, a methodology that involves determining the veracity of a given claim by considering supporting or refuting evidence.

In this domain, the FEVER dataset introduced by [273] stands out, as it aims to evaluate fact-checking performance on altered claims derived from Wikipedia articles. Other notable approaches, such as those presented by [20] and [178], employ web search engines to identify real-time potential evidence and assess their alignment with the input claim. Additionally, [207] utilizes LSTM models and attention mechanisms for document retrieval and identifying the most relevant sentences within those documents. Building upon this foundation, [190] introduces neural semantic matching networks to tackle both document retrieval and evidence selection tasks. Inspired by the remarkable performance of transformer architectures in various natural language processing (NLP) tasks, [257] adopts the BERT model to determine the relevance of evidence and evaluate the veracity of the input claim. Moreover, [34] and [33] employ reasoning techniques over an entity-graph and a hierarchical hypergraph, respectively, to conduct fine-grained verification using evidential information.

Another avenue of research in fact-checking involves leveraging knowledge bases. [271] constructs a knowledge graph containing fact-checked information that can be queried to assess the veracity of a given claim. Meanwhile, [245] encodes background knowledge in the form of Horn rules and generates rule-based explanations that support veracity predictions for claims. In a novel approach, [245] determines claim truthfulness by treating the knowledge graph as a flow network. Lastly, [202] proposes the use of language models as a knowledge base, capitalizing on their factual knowledge acquired during pretraining for improved fact-checking performance.

3.2.2 Verified Claim Retrieval

While identifying check-worthy claims, a.k.a. claim detection, finding the evidences supporting or refuting a claim, a.k.a. evidence retrieval, and detecting claim truthfulness, a.k.a. claim verification, have been extensively studied so far [180, 273], there has been a notable gap in exploring the phenomenon of recurring claims, particularly those that have potentially been previously verified and spread across different contexts or time periods.

Only recently, the claim retrieval task has been proposed to detect previously fact-checked information [235]. In essence, this task can be framed as an information retrieval (IR) problem where a collection of verified documents must be ranked in relevance to a given input claim. However, unlike traditional ad-hoc retrieval scenarios, the corpus of documents, representing verified information, is dynamic, and ideally, it should be updated for each new assessment of truthfulness.

Building upon this conceptualization, participants in the CheckThat!2021 competition [182] demonstrated that fine-tuning state-of-the-art transformer models for re-ranking purposes yields significant performance enhancements compared to applying standard IR algorithms (e.g., BM25 [220]) in isolation. Additionally, MTM [244] improved transformer-based re-ranking performance by identifying crucial sentences and common pattern templates within the document corpus. Similarly, [159] achieved comparable results in the context of multilingual Covid-19 claims, covering both English and Arabic. Conversely, focusing on a debate scenario, [234] evaluated the impact of modeling the claim's global and local contexts on (re-)ranking performance. Finally, MAN [285] introduced a custom neural network-based architecture for claim retrieval using multimodal data, incorporating both textual and image-based information from the claims and verified documents.

Despite the promising performance exhibited by the aforementioned methodologies, they primarily address the re-ranking stage, assuming the presence of a dependable retrieval algorithm, typically BM25, for the initial document selection process. In contrast, our focus extends to the comprehensive design of systems capable of excelling in both retrieval and re-ranking contexts. Furthermore, our evaluation approach delves into the real-world applicability of the task and its intricate interactions within the

dynamic information landscape, surpassing the limitations of benchmark datasets. Moreover, we explore the potential enhancements introduced by incorporating multimodal elements into the task, but, unlike [285], our model harnesses state-of-the-art visual-language models, consistently outperforming baseline methods, even in the re-ranking scenario.

3.3 Background

The task of ranking a list of documents in response to specific queries is a commonplace challenge encountered in information retrieval endeavors. This becomes particularly evident when dealing with extensive document corpora, where the prevailing approach involves the adoption of multi-stage pipelines, as observed in [30]. In this context, the initial stage, referred to as the *retriever*, is responsible for the preliminary selection of the top-k documents deemed potentially relevant to the query. Subsequently, the second stage (and potentially subsequent stages) known as the *reranker* comes into play, with the objective of reordering this set of candidate documents using more sophisticated and computationally intensive models.

3.3.1 Retriever

The first-stage retrieval task has long been dominated by the classical term-based probabilistic models (e.g. BM25 [220]) due to their efficiency and effectiveness even with million-scale corpus of documents. Nevertheless, they still suffer from the vocabulary mismatch problem ([71]) and do not model the document semantics which is essential when considering text's meaning. While in the past decades term dependency and topic models ([166, 25, 131]) have addressed the former problem, the unprecedented performance improvements that transformer architectures and representation learning strategies are achieving in NLP, have determined an explosive growth of works proposing their neural network-based semantic first-stage retriever. These neural retrievers can be categorized into two main types [30]: *sparse retrieval methods* and *dense retrieval methods*. The former employ efficient sparse representations for queries and documents, enhancing the weighting scheme of traditional term-based methods. Examples of these strategies include DeepCT [47] and docT5query [194].

In contrast, *dense retrieval methods* usually employ a dual-encoder architecture where queries and documents are independently embedded. The final relevance score is calculated using a similarity function denoted as f. These methods can be further categorized into *term-level representation learning* and *document-level representation learning* [30]. In *term-level representation learning* approaches, queries and documents are represented as sequences of term embeddings, and the similarity function f operates at the term level, aggregating the results to compute the final score. Examples of such methods include DC-BERT [189] and ColBERT [110]. Conversely, *document-level representation learning* approaches aim to find a single global representation for each query and document. Examples of this approach include Sentence-BERT [217] and DPR [108].

It is worth to note that even if the above-mentioned methods are categorized as first-stage retriever for their efficiency, they can still be used for end-to-end retrieval, performing jointly the retrieval and reranking tasks.

3.3.2 Reranker

While certain retriever models have demonstrated discrete ranking performance [220, 110], there is an ongoing effort to develop specialized learning-to-rank systems. Over the past decade, there has been a significant surge in the utilization of deep neural networks for constructing ranking models, commonly referred to as neural ranking models (NRMs). These NRMs can be broadly categorized into two classes: *representation-based* and *interaction-based* approaches [87].

The former methods employ a similar bi-encoder plus matching layer architecture as adopted by *dense retrieval methods*. Notable examples include DSMN [98] and ESIM [36], which utilize fully-connected networks and chained LSTMs, respectively, for tasks such as Natural Language Inference. In the realm of fact-checking, NSMN [190] combines bidirectional LSTMs and a pooling strategy to jointly perform evidence retrieval and fact verification.

Conversely, *interaction-based* NRMs aim to capture relevant matching signals between a query and a document based on word interactions. Early works like MatchPyramid [195] and KNRM [305] applied deep neural networks to represent word interaction matrices. More recently, pre-trained transformers [333, 52] have achieved state-of-the-art performance

in ranking-related tasks. In particular, [91] demonstrates the effectiveness of ensembling different BERT models and combining point-wise, pair-wise, and list-wise loss functions. Similarly, [193] proposes a two-stages re-ranking pipeline with point-wise (monoBERT) and pair-wise (duoBERT) classification models, respectively.

Additionally, hybrid architectures like DUET [171] have been proposed, combining outputs from models of different categories to produce relevance scores.

While *interaction-based* approaches often outperform representation-based ones in terms of ranking performance, their application for end-to-end retrieval remains constrained due to their reduced efficiency in online ranking scenarios [87].

3.4 Benchmarking IR and Q&A Models for Verified Claim Retrieval

3.4.1 Datasets & Metrics

We utilize a dataset provided by [235], comprising 1000 tweets sourced from Snopes[4] fact-checking articles, along with 10396 verified claims from the ClaimsKG dataset [271]. This dataset covers diverse domains, including politics and gossip. Notably, tweets and their corresponding verified documents sometimes exhibit similar phrasing, facilitating approximate matching, while in other cases, differing terms necessitate more advanced semantic matching.

Our data split follows the standard 60%20%20% division provided by the authors for training, validation, and testing sets. Similar to many information retrieval tasks, numerous verified claims do not have related original tweets.

For evaluation, we employ Mean Reciprocal Rank (MRR), Mean Average Precision truncated at k (MAP@k), and the hit ratio [94] truncated at k (HasPositives@k). While the first two metrics consider ranking order, the latter assesses the system's ability to retrieve correct matches. HasPositives@k is essentially Recall@k since most tweets have only one relevant document. We conduct a statistical t-test between top-ranked

[4] https://www.snopes.com/

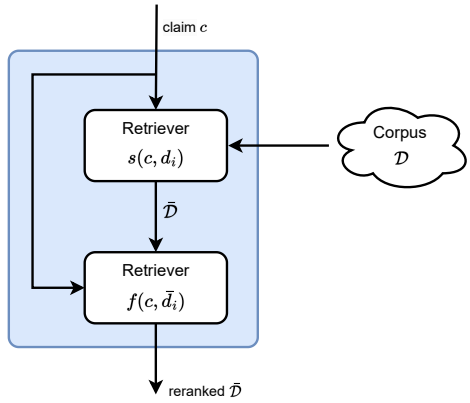


Figure 3.1. Pipeline of our benchmarking architecture.

models to validate our results. From the application perspective, metrics at lower k values (e.g., $k \in 1, 3, 5$) can gauge the system's utility in assisting manual fact-checkers by quickly identifying relevant results. Conversely, metrics at higher k values (e.g., $k \in 10, 20$) are pertinent in offline settings or automated fact-checking pipelines, where results serve as evidence for veracity predictions.

3.4.2 Benchmarking architecture

As highlighted in the preceding section, ranking challenges are pervasive in information retrieval, driving extensive exploration of machine learning methods to devise efficient solutions. To facilitate the integration and comparative assessment of contemporary neural ranking and retrieval models alongside conventional information retrieval techniques, we employ a two-stage learning-to-rank framework, illustrated in Figure 3.1.

The first-stage retriever aims at selecting the subset $\bar{\mathscr{D}}$ of the documents corpus. Specifically, it operates under the assumption that the input claim and the most pertinent documents exhibit shared fundamental characteristics, such as mentioning the same entities, possessing similar statistical representations (e.g., TF-IDF features), or addressing corresponding concepts and topics. In essence, the retriever's role is to filter out entirely unrelated verified information contained within $\mathscr{D}$. While we do not an-

ticipate exceptionally high ranking performance from the retriever, we do require it to achieve commendable recall scores, ensuring that it does not adversely impact the reranking process executed in the subsequent stage. In simpler terms, a pre-selection algorithm that excludes a substantial number of relevant documents would act as a bottleneck, hindering the overall system's performance. Furthermore, given that the retriever operates on extensive document collections, it must exhibit efficiency and scalability relative to the corpus size. The impact of the algorithm choice will be assessed in the experimentation section.

The second-stage reranker represents an advanced NRM that delves into the inherent semantics of both the claim and the (selected subset of) documents. Its primary objective is to achieve high-performance reranking. In essence, after the retriever has filtered out documents that exhibit a significant correlation with the input claim at a broad level, the reranker engages in semantic matching. Its aim is to determine whether the input query and document, represented as the $(c, \bar{d}_i)$ pair, convey, even to a partial extent, the same meaning or concepts.

It is important to underscore that our chosen multi-stage pipeline provides a means to systematically benchmark both interaction-based and representation-based rerankers, while imposing minimal computational overhead. Although the NRM incurs computational demands, its primary task is to predict the relevance between the input claim and a substantially reduced subset of verified documents pre-selected by the retriever algorithm. The ramifications of this reranking procedure on both the effectiveness and computational efficiency of the overall framework needs to be meticulously assessed through empirical experiments.

Despite recent attempts to construct end-to-end neural retrieval systems [110, 281], we posit that our multi-stage pipeline, apart from enhancing computational efficiency, may yield performance benefits for the reranker. This hypothesis stems from the relatively simplified problem the reranker encounters when operating in conjunction with the retriever. To elaborate, when the ranking model operates in isolation, it must effectively discriminate between the semantic content of the input claim and the extensive knowledge encompassed within the document corpus. In contrast, our experimental setup furnishes a more controlled environment, where the training process can presuppose a certain degree of semantic relevance

between the claim and the documents being (re)raked. This controlled environment may facilitate more efficient learning and potentially lead to performance improvements.

3.4.3 Experimental Protocol

In our endeavor to synthesize the extensive literature encompassing retrieval and ranking models for the detection of previously fact-checked documents, we formulate our research objectives. These objectives aim to elucidate the most suitable methods for our two-stage pipeline while simultaneously scrutinizing both the effectiveness and efficiency of the framework. Specifically, we seek to address the following research questions:

- Which are the best retrievers? Can modern neural semantic techniques replace the standard term-based approaches?
- Which are the best neural (re-)ranking models?
- What is the benefit of combining retrievers and rerankers with respect to the overall performance?

Experimental Setup

In the following subsection we detail which are the retrievers/rerankers considered in the benchmark, explaining how they have been trained and configured in order to promote reproducibility.

We select a wide range of retrievers dividing them in four groups.

First, we consider classical probabilistic approaches including BM25 [220], TF-IDF [300] and Language Model with Dirichlet smoothing [317]. These algorithms assign a score to each tweet-claim pair based on exact matching between the words in the tweet and the words in a target verified claim. They have been long studied and applied in various information retrieval tasks, thus representing the baseline for the other retrievers. We adopted the Elasticsearch[5] (version 7.10.1) implementation for BM25 and LM Dirichlet, with default parameters, and used Haystack library[6] for TF-IDF.

[5] https://www.elastic.co/

[6] https://github.com/deepset-ai/haystack

Second, we select docT5query [37] as neural sparse retrieval models because expanding the documents with auto-generated queries seems profitable in this context because the query, i.e., the (false) information, is often repeated with a few differences over times. Specifically, we adapted the official code[7] and use the provided *T5-base* model to generate three queries for each document. We then used BM25 to reindex the expanded documents.

Third, we choose ColBERT [110] as neural sparse retrieval models. It is worth noting that ColBERT can be used for reranking as well, due to the interaction mechanism it performs between query and document terms. In particular, we used the official repository[8] and retrained the *bert-based-uncased* model using the default hyper-parameters.

Fourth, we picked SentenceBERT [217] and DPR [108] as neural document-based dense retrieval techniques. The former is the first attempt to leverage transformer-based models to perform text similarity and thus represents our "neural" baseline. Specifically, we used the sentence-transformer library[9], fine tuning (for 4 epochs and a batch size of 16) the *stsb-distilbert-base* model using cosine similarity loss.

On the other hand, the latter adopts the in-batch negative strategy to reuse negative examples already in the training batch rather than creating new ones. In particular, we used the Haystack library[8], fine tuning (for 10 epochs and a batch size of 16) the *bert-base-uncased* model.

With the exception of DPR which customizes the batch generation strategy, the training dataset has always been built considering the positive query-document pair and 10 random negative ones.

Considering the second stage of the pipeline, we considered 9 rerankers, divided in the categories mentioned in Section 3.3.2. Specifically, we choose MatchPyramid [195], KNMR [305], ConvKNMR [48] and BERT models [333], as interaction-based algorithms; ESIM [36] and HAR [331], as representation-based algorithms; DUET [171] as hybrid model.

For HAR we used the official implementation[10], and for all others methods we adopted the Pytorch implementation of the Matchzoo framework

[7] https://github.com/castorini/docTTTTTquery
[8] https://github.com/stanford-futuredata/ColBERT
[9] https://github.com/UKPLab/sentence-transformers
[10] https://github.com/mingzhu0527/HAR

[86]. All hyper-parameters have been set to default with the exception of the number of kernels in KNRM and ConvKNRM which was set to 11. All models have been trained until convergence on the validation set. Finally, for BERT, we adopted the *stsb-distilroberta-base* cross-encoder provided by the sentence-transformer library[11], fine tuning (for 4 epochs and a batch size of 16) using the cross-entropy loss.

When training rerankers, we need to select k negative samples for each tweet-claim pair. The choice of k might be decisive for the performance of the model: low values might determine poor performance because the model would see few pairs representing non-matching knowledge. On the other hand, since there is just one verified claim matching most of the tweets in our dataset, increasing k too much might lead to imbalanced training set, making the learning task more difficult. We select 50 random negative documents from the top-100 ones retrieved in the first stage. However, in the experiments we also evaluate the effect of a completely random choice.

3.4.4 Results

Which are the best retrievers?

Table 3.1 presents the results of the retriever models. It is noteworthy that we have not included latency performance metrics, as the document corpus is relatively concise, making it challenging to observe substantial differences among the chosen models in this regard.

Although they do not attain the level of performance achieved by BM25, the progress of neural retrievers is unmistakable as they surpass the TF-IDF baseline and perform on par with the LM Dirichlet model.

The sparse model, docT5Query [37], emerges as the first runner up among the retriever models, exhibiting significant enhancements in comparison to the BM25 baseline it relies upon. Our conjecture is that this notable improvement is attributed to the augmentation of fact-checked documents through artificially generated queries, subsequently indexed through conventional techniques (in our case, BM25). This approach is evidently effective as the generation process adeptly extracts subjects, topics, and events, consequently increasing the likelihood of detecting matching queries referencing these concepts. Regrettably, it is imperative to acknowledge

Table 3.1. Performance of retrievers (bold indicates the best results, underline the first runner up)

Category	Model	MRR	HasPositives@k						
		all	$k=1$	$k=3$	$k=5$	$k=10$	$k=20$	$k=50$	$k=100$
Classical	TF-IDF	0.681	0.593	0.739	0.789	0.829	0.869	0.914	0.924
	LM Dirichlet [318]	<u>0.799</u>	<u>0.770</u>	0.825	<u>0.860</u>	0.890	0.915	<u>0.95</u>	0.960
	BM25 [220]	**0.817**	**0.785**	**0.865**	**0.880**	**0.895**	**0.915**	**0.950**	**0.960**
Neural sparse	docT5query [194]	0.786	0.754	<u>0.834</u>	0.844	<u>0.894</u>	<u>0.919</u>	0.945	<u>0.960</u>
Term-based	ColBERT [110]	0.765	0.708	0.793	0.819	0.874	0.904	0.944	0.949
Document-level	SentenceBERT [217]	0.669	0.592	0.713	0.763	0.804	0.834	0.884	0.924
	DPR [108]	0.624	0.547	0.673	0.718	0.753	0.788	0.859	0.909

that the query generation process, reliant on the T5 transformer [212], is computationally intensive and may not be practical for online scenarios that necessitate frequent updates to the document corpus.

Conversely, ColBERT [110] attains compelling performance without necessitating any preprocessing steps. Furthermore, its late interaction mechanism between query and document words appears to be sufficiently efficient and scalable even with document corpora of substantial magnitude.

Lastly, the document-level neural retrievers, namely SentenceBERT [217] and DPR [108], are one step behind the other approaches. This discrepancy can likely be attributed to the fact that representing the entire document/query with a single embedding yields a coarse representation, insufficient to capture the intricate details necessary for inferring the relationship between the claim and its corresponding verified document. Specifically, fact-checked documents often comprise extensive texts citing numerous concepts and entities to assess the veracity of the claim. In such a context, it becomes challenging to provide an insightful representation by considering the document as a whole, rather than delving into more granular information, such as individual terms and sentences within the text.

To sum up, recent progress in neural information retrieval methods appears to be narrowing the performance gap with classical retrieval approaches. However, our findings demonstrate that even the most contemporary retrievers cannot entirely supplant the efficacy of traditional

methods in practical applications. Furthermore, our research underscores that the amalgamation of these two retrieval paradigms, along with the development of more efficient interaction functions, represent the most auspicious avenues for future investigation in this domain.

Which are the best neural re-ranking models?

Table 3.2 presents the performance of rerankers, specifically focusing on queries that have at least one relevant article within the top 50 documents retrieved by BM25 in the initial stage. As expected, interaction-based approaches generally outperform representation-based ones, given their explicit search for relevant matching signals within query-document pairs.

Across reranker categories, it is evident that transformer-based models, namely BERT [333] and colBERT [110], significantly outperform other algorithms. Notably, they yield impressive results even when considering truncated rankings at the top positions, indicating their ability to effectively capture the relationship between fact-checked documents and input claims. However, it is worth noting that the execution time of these models has a substantial impact on the number of documents they can practically rerank, an aspect we will delve into in the subsequent section.

When observing the huge performance difference between transformer-based systems and other NRMs, we conjecture that it depends on the transformers' pre-training procedure, which allows these models to acquire not only language syntax and semantics but also factual and relational knowledge [202]. By contrast, other NRMs (e.g. MatchPyramid [195], KNRM [305]) are trained from scratch, thus requiring more training (labelled) data and time to achieve good reranking performance.

Lastly, BERT [333] outperforms colBERT [110] due to its more intricate interaction mechanism, which captures matching signals between the input claim and verified document more comprehensively. Indeed, although colBERT's late interaction mechanism prioritizes (computational) efficiency, it cannot compete with the full self-attention mechanism BERT relies on.

To sum up, fine-tuning pre-trained language models appears to be the most effective and straightforward approach for obtaining high-quality rankings. A more equitable comparison with other neural ranking models

Table 3.2. Performance of Neural Ranking Models (NRMs) (bold indicates the best results, underline the first runner up, * statistical significance at $p = 0.001$ w.r.t. the second best)

Category	Model	MRR	MAP@k				
		all	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
Interaction -based	BERT [333]	**0.968***	**0.942***	**0.968***	**0.968***	**0.968***	**0.968***
	ColBERT [110]	<u>0.903</u>	<u>0.847</u>	<u>0.893</u>	<u>0.901</u>	<u>0.902</u>	<u>0.903</u>
	MAN [285]	0.509	0.386	0.470	0.484	0.501	0.509
	MatchPyramid [195]	0.495	0.413	0.444	0.462	0.479	0.489
	KNRM [305]	0.319	0.212	0.272	0.287	0.298	0.307
	ConvKNRM [49]	0.744	0.677	0.721	0.729	0.738	0.742
Representation -based	ESIM [36]	0.507	0.370	0.451	0.482	0.498	0.504
	HAR [331]	0.602	0.331	0.508	0.557	0.557	0.560
Hybrid-based	DUET [171]	0.392	0.233	0.302	0.313	0.323	0.330

may become feasible when a million-scale dataset of fact-checked information is made available.

What is the benefit of combining retrievers and rerankers?

As highlighted in the previous section, the two steps within the retriever-ranker framework capture distinct types of information. Therefore, it is valuable to explore the combined performance of these steps. Tables 3.3 and 3.4 present an overview of the overall system's performance, measured in terms of HasPositive@k and MAP@k, respectively. This evaluation considers the integration of two transformer rerankers, namely BERT and ColBERT, with the top-performing retriever algorithm, BM25. The system's configuration involves the selection of the top 100 verified claims by the latter model from the document corpus.

The results underscore the effectiveness of combining a robust retriever with a more powerful reranker (BERT). In fact, the overall combination surpasses the performance of both individual components considered in isolation. Conversely, when weaker rerankers (ColBERT) or less proficient retrievers are employed, the system's performance may be compromised, potentially acting as a bottleneck for the entire framework. To be completely fair, it is worth noting that while the results of the two

Table 3.3. Performance of the overall pipeline (bold indicates the best results, underline the first runner up)

Model	HasPositive@k				
	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
BM25 [220]	0.785	0.865	0.880	0.895	0.915
BERT [333]	**0.865**	**0.935**	**0.960**	**0.970**	**0.985**
ColBERT [110]	0.793	0.819	0.874	0.904	0.944
BM25 (100) + BERT	<u>0.862</u>	<u>0.925</u>	<u>0.935</u>	<u>0.945</u>	<u>0.955</u>
BM25 (100) + ColBERT	0.779	0.794	0.804	0.804	0.804

Table 3.4. Performance of the overall pipeline (bold indicates the best results, underline the first runner up)

Model	MRR	MAP@k					
	all	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$	all
BM25 [220]	0.817	0.816	0.819	0.821	0.822	0.817	0.785
BERT [333]	<u>0.901</u>	<u>0.865</u>	<u>0.895</u>	<u>0.901</u>	<u>0.902</u>	<u>0.903</u>	<u>0.903</u>
ColBERT [110]	0.709	0.749	0.754	0.762	0.765	0.765	0.708
BM25 (100) + BERT	**0.906**	**0.873**	**0.905**	**0.908**	**0.908**	**0.908**	**0.909**
BM25 (100) + ColBERT	0.739	0.756	0.760	0.761	0.761	0.762	0.738

top-performing models are statistically indistinguishable, they are still significant. Even though these two solutions exhibit comparable performance, the combination of BM25 and BERT proves to be considerably more efficient than BERT alone, as we will elucidate in the efficiency experiment.

Additionally, Table 3.5 provides compelling evidence that retraining the reranker on the claims retrieved during the first stage has a positive impact on ranking performance. This confirms our initial hypothesis that employing the prefiltering information retrieval algorithm simplifies the learning task. This simplification occurs because the neural ranking model no longer needs to align the semantic content of the input tweet with the entirety of the knowledge encoded within the verified claims.

Lastly, we conduct an evaluation to scrutinize the influence of the number of documents selected by the first-stage retriever on the efficiency and effectiveness of the overall system. Table 3.6 presents the runtimes, accompanied by their corresponding 95% confidence intervals, for the system resulting from the combination of the BM25 retriever with each transformer model. We also consider the scenario where no reranker is applied. Notably, representation-based models like colBERT exhibit favorable scalability characteristics, surpassing even the performance of the BM25 baseline in terms of runtime efficiency. Conversely, the BERT model experiences a substantial increase in runtimes, sometimes reaching up to 30 seconds per query as the number of retrieved documents expands. To contextualize these runtime considerations, we emphasize that for a system aimed at facilitating manual fact-checkers' efforts, maintaining a response time of approximately five seconds is deemed advantageous. Consequently, the application of the computationally intensive BERT algorithm becomes infeasible when dealing with more than 1000 documents that require reranking. Consequently, the practical use of complex transformers is closely tied to the adoption of a high-recall retriever, which effectively filters out a significant portion of the documents in the corpus.

Furthermore, Figure 3.2 presents the MAP metrics while varying the top-k fact-checked documents retrieved by the BM25 baseline. As expected, the performance generally increases (or decreases) with the integration of the BERT (or colBERT) reranker. This behavior arises from the fact that, as the number of retrieved documents increases, the system gradually converges toward the performance of the second-stage reranker

Table 3.5. Effect of negative pairs' selection during reranker training

Model	MAP@k				
	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
BERT (random negatives)	0.365	0.525	0.556	0.573	0.575
BERT (top-k negatives)	0.865	0.895	0.901	0.902	0.903

Table 3.6. Runtimes (in seconds) varying the number of claims to rerank

	Without rerank	**colBERT**	**BERT**
BM25 (10)	0.0170 ± 0.0017	0.0634 ± 0.0014	0.0500 ± 0.0100
BM25 (100)	0.0233 ± 0.0010	0.0703 ± 0.0014	0.3483 ± 0.0153
BM25 (1000)	0.1156 ± 0.0054	0.1688 ± 0.0053	3.3851 ± 0.1709
BM25 (10000)	0.6122 ± 0.0900	0.7225 ± 0.0908	30.8846 ± 1.5110

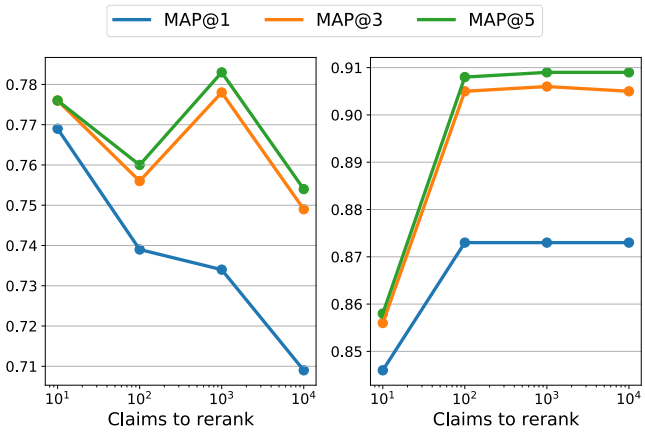


Figure 3.2. Performance varying the number of claims retrieved by the first stage and reranked by ColBERT (left) and BERT (right).

when applied in isolation. Consequently, the retriever's role becomes less prominent. Nevertheless, it is noteworthy that for the BERT model, performance no longer exhibits improvement beyond the retrieval of more than 100 documents.

In conclusion, the realm of information retrieval literature offers a diverse array of methods and models that hold promise for effectively addressing the challenge of identifying previously fact-checked information. More precisely, the adoption of a multi-stage ranking pipeline demonstrates the capability to attain satisfactory quality performance by seamlessly integrating efficient retrievers with more intricate rerankers. This integration effectively manages the delicate balance between ranking performance and computational runtimes.

In the specific context of our benchmark study, we arrive at the determination that the optimal system configuration comprises the BM25 model responsible for retrieving up to 100 fact-checked documents, followed by the BERT model, which excels in high-performance reranking tasks.

3.4.5 Discussion

Fact-checking continues to serve as the primary line of defense against countering online disinformation. Nevertheless, despite the considerable intensification of their endeavors, fact-checking organizations confront a formidable challenge in keeping stride with the vast deluge of false information disseminated across social media platforms. In this prevailing milieu, the identification of previously fact-checked information emerges as a strategic imperative. Such an endeavor holds the potential to enhance the verification process, thereby augmenting the productivity of fact-checkers. Furthermore, it offers a reservoir of more dependable evidence upon which assessments can be founded.

Conceptually, our study serves as a bridge between recent ad-hoc initiatives targeting the verified claim retrieval task and the extensive array of techniques and models introduced in the domains of information retrieval and Q&A literature. To facilitate this alignment, we have devised a retriever-reranking framework geared towards evaluating the efficiency and effectiveness of the scrutinized methodologies. Diverging from existing studies, our investigation seeks to uncover the optimal equilibrium between ranking performance and computational execution times.

Our findings underscore the promise of amalgamating conventional and neural methodologies as the most auspicious avenue for enhancing retrieval performance. Fine-tuned transformer architectures emerge as stalwarts, capable of delivering high-caliber reranking performance. Furthermore, the computational efficiency of these sophisticated reranking models remains feasible in practice, as the necessity to process an extensive volume of documents for ranking improvement is obviated.

3.5 Our Multimodal Proposal

3.5.1 Problem Formulation

Let's assume $\mathscr{C} = \{c_1, c_2, \ldots, c_N\}$ represents the collection of claims to undergo fact-checking, where each claim c_i comprises its textual content c_{text} along with associated images $\{c_{im_1}, c_{im_2}, \ldots, c_{im_k}\}$. Similarly, let $\mathscr{D} = \{d_1, d_2, \ldots, d_N\}$ denote a substantial repository of validated documents, with each document d_i consisting of textual content $d$text and accompanying images $\{d_{im_1}, d_{im_2}, \ldots, d_{im_k}\}$.

Building upon the promising outcomes elucidated from the benchmark study (Section 3.4), we formalise the task of multimodal verified claim retrieval under the same retriever-reranker paradigm. In this setting, given an input claim c, the initial-stage retriever endeavors to learn a function $s : (c, d_i); |; d_i \in \mathscr{D} \rightarrow \mathbb{R}$. This function is devised to impart elevated scores to (c, d) pairs that exhibit relevance and conversely, to assign lower scores to those that manifest irrelevance. Essentially, the retriever's primary objective is to distill a subset $\bar{\mathscr{D}} = \{\bar{d_1}, \bar{d_2}, \ldots, \bar{d_M}; |; M \ll N\}$ consisting of documents potentially pertinent to c. As a consequence, this operation results in $\bar{\mathscr{D}}$ being a proper subset of $\mathscr{D}$.

Subsequently, the secondary-stage reranker undertakes the task of learning a function $f : (c, \bar{d_i}); |; \bar{d_i} \in \bar{\mathscr{D}}); \rightarrow \mathbb{R}$. This function is responsible for the rearrangement of elements within $\bar{\mathscr{D}}$ based on their respective degrees of similarity to the input claim c.

3.5.2 Our framework

Motivated by the remarkable efficiency demonstrated by two-tower ranking architectures [30], we have devised a straightforward yet highly

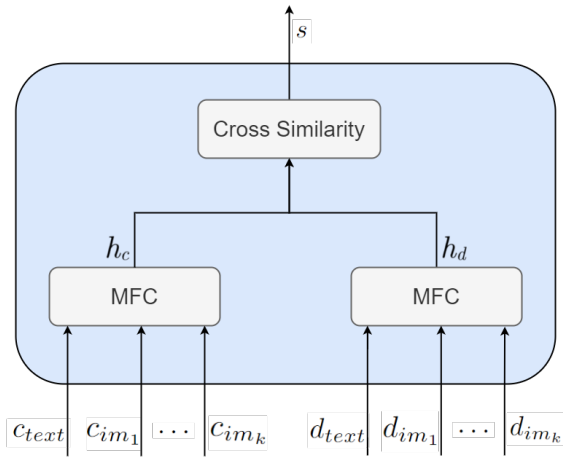


Figure 3.3. Our framework

effective model with versatility as both a retriever and a reranker. The schematic representation of our framework is presented in Figure 3.3.

Following the initial stage where the *Multimodal Fusion Component* (MFC) is responsible for extracting meaningful representations from both multimodal claims and verified documents, the subsequent phase is managed by the *Cross Similarity* (CS) module. This module takes charge of implementing an efficient similarity function that calculates the ultimate relevance score.

Multimodal Fusion Component (MFC) The integration of various data modalities represents an open challenge within the realm of machine learning research, inciting exploration into new avenues of multimodal tasks such as visual question answering (VQA) [260] and topic learning [139]. Moreover, the specific nature of our task demands consideration, as both input claims and documents may contain multiple images.

In response to these demands, we have elected to employ multimodal transformers as the core of our MFC module. This choice is grounded in their inherent flexibility to accept inputs encompassing images, achieved by projecting the image embeddings into the text token space [112]. Furthermore, these multimodal transformers are endowed with common-sense and

factual knowledge acquired during the pre-training process [202], rendering them adept at aligning query content with the corresponding document. This capability is particularly valuable when specific contextual information cannot be readily gleaned from the inputs themselves.

Formally, the MFC module implements the function

$$h : \{x_{text}, [x_{im_1}, \ldots, x_{im_k}]\} \rightarrow h_x \in \mathbb{R}^n,$$

n being the latent space's size and x being either a claim $c \in \mathscr{C}$ or a document $d \in \mathscr{D}$. It is important to emphasize that the output vector is not a mere amalgamation of independently computed text and image embeddings. Instead, it emerges from a sophisticated and robust computation involving self-attention and encoder-decoder attention mechanisms, which capture significant information from individual modalities and their interactions, respectively.

Cross Similarity (CS) The CS module serves the purpose of gauging the degree of similarity between an input claim c and a document d, grounding its analysis in their respective latent representations. Formally, it embodies the function $f : (h_c \in \mathbb{R}^n, h_d \in \mathbb{R}^n) \rightarrow s \in \mathbb{R}$, where s stands as the ultimate relevance score that quantifies the relationship between the two inputs.

Assessing the relevance between an input claim and a verified document extends beyond a mere scrutiny of overlapping words or the computation of image similarity. In the realm of fact-checking, instances arise where the claim and its corresponding verified document employ divergent lexical choices to convey identical concepts. Moreover, intricate claims may amalgamate various verified claims, rendering partial matches relevant for the task at hand [235]. Furthermore, when considering multimodal content, the image contained within the input claim typically mirrors that in its verified document. However, the manner in which the text references the image can vary contingent on how the image was originally utilized, including scenarios involving decontextualized images.

The inherent intricacies of these interaction mechanisms have led to the development of potent yet computationally intensive CS functions, which capture interaction signals across varying granularities (e.g., word, sentence) or rely on specially extracted keywords from the verified documents

[285, 244]. However, the computational demands of such intricate similarity functions preclude their adoption in retrieval scenarios, as they are not scalable to corpora comprising millions of documents.

Conversely, our objective is to devise a system possessing both retrieval and reranking capabilities. Recent strides in neural semantic retrievers have demonstrated that even simpler CS functions (e.g., cosine similarity, dot product) can yield favorable outcomes when complemented by an effective learning regimen, particularly when the input encoder – represented by the MFC module – generates highly representative features [30]. In light of these insights, we leverage these findings to execute our investigative task, subsequently scrutinizing the impact of various efficient similarity metrics (i.e., cosine similarity, dot product, negative Euclidean distance) within our experiments.

3.6 Experimental evaluation

3.6.1 Dataset & Metrics

For all experiments, we leveraged the datasets provided by [285]. The dataset from Politifact pertains to the political domain and encompasses a corpus of 2026 queries and 467 documents. Conversely, the Snopes dataset is oriented towards the realm of social networks, specifically Twitter, encompassing a larger volume with 11167 queries and 1703 documents. For an in-depth view of these datasets, Table 3.7 presents a comprehensive summary, denoting $E[|\cdot|]$ and $vocab|\cdot|$ as the average length and the number of unique words, respectively. Additionally, each document, on average, contains two images, while each query is consistently associated with a single image. At a fundamental level, these (verified) documents constitute rather formal textual compositions wherein experts have meticulously assessed the veracity of claims, providing a rationale for their determinations. As a consequence, most claims are substantiated by a single document. For illustrative purposes, Figure 3.5 provides exemplars from each dataset.

In our ranking formulation, we employ the Normalized Discounted Cumulative Gain truncated at k (NDCG@k) and the hit ratio [94] truncated at k (HIT@k) as our primary evaluation metrics. Specifically, when evalu-

Table 3.7. Datasets statistics

Statistic	Politifact	Snopes
$\mid \mathscr{C} \mid$	2 026	11 167
$\mid \mathscr{D} \mid$	467	1 703
$E[\mid c \mid]$	58	70
$E[\mid d \mid]$	1 822	7 246
$\mid vocab\{\mathscr{C}\} \mid$	1 043	4 446
$\mid vocab\{\mathscr{D}\} \mid$	38 899	116 254

ating retriever performance, we resort to HIT@k with $k \in 50, 100, 200$ for the Politifact dataset and $k \in 50, 200, 500$ for the Snopes dataset. Conversely, for the analysis of reranker performance, we consider HIT@k and NDCG@k with $k \in 1, 3, 5$ for both datasets. This dichotomy arises from the distinct roles of retrievers, which yield a subset of potentially relevant documents, necessitating elevated recall scores for larger k values, and rerankers, tasked with delivering the final ranked output, thus requiring the maximization of the relevance ranking for claim-document pairs.

Finally, to estimate efficiency, we collected runtimes for 50 random claims and performed the statistical t-test to assess the reliability of the results under both retrieval and reranking settings.

3.6.2 Experimental protocol

We conduct an extensive set of experiments to comprehensively address a set of pivotal research inquiries, namely:

- *Modality Fusion Strategies*: We seek to ascertain whether there exists an optimal approach for amalgamating image and text modalities tailored to our specific use case.

- *Reranking Evaluation*: Our focus extends to the performance evaluation of our system in a reranking context. This evaluation encompasses a comparative analysis encompassing both the efficiency

and effectiveness of our proposed system concerning contemporary reranking systems. These systems are either explicitly devised for the verified claim retrieval task or originate from the ad-hoc information retrieval domain.

- *Retrieval Evaluation*: Our experimental purview ventures into a retrieval scenario, an unexplored territory within the domain of detecting previously fact-checked information. We endeavor to evaluate the performance of our system in this novel context.

- *Ablation Study*: We strive to dissect the contribution of each modality to the overall system performance and, concurrently, appraise how their collaborative integration enhances system efficacy

- *Error Analysis*: This qualitative assessment serves to identify areas where our model encounters challenges in correctly identifying claim-document matches within authentic scenarios.

Experimental Setup

All experiments have been conducted on a machine equipped with CPU Intel i9-9900K, RAM 32 GB and one NVIDIA GeForce RTX 2070.

We conduct a comprehensive evaluation of our system's performance in both reranking and retrieval tasks. In both scenarios, the training process necessitates the selection of k negative samples for each gold claim-document pair. The choice of the parameter k carries significant weight as it can substantially impact model performance. Low values of k may result in diminished performance, as the model encounters too few pairs representing non-matching knowledge. Conversely, excessively increasing k can lead to an imbalanced training set, rendering the learning task more challenging. Our empirical observations indicate that utilizing $k = 50$ negative documents yields favorable results.

Specifically, when undertaking the reranking task, we employ the BM25 algorithm as the first-stage retriever, ensuring comparability with competing methods [285, 244]. In contrast, under retrieval settings, we randomly select 50 negative documents during our system's training process.

The dataset division comprises training (60%), validation (20%), and testing sets (20%) to embark on the training procedure, hyperparameter

Table 3.8. Effect of similarity functions: NDCG@3 score on the validation set under reranking settings.

Similarity function	Politifact	Snopes
Cosine	**.712**	**.860**
Neg. Euclidean Distance	.695	.846
Dot product	.692	.857

tuning, and performance evaluation, respectively. In our hyperparameter tuning, we explore different values for hyperparameters, including the learning rate ($\alpha \in [10^{-6}, 10^{-5}]$), embedding sizes ($n \in 100, 200, 300$), and the loss function (cross-entropy, hinge, and approxNDCG [87]). We opt for uniform sampling to select hyperparameter values and conduct a total of 20 experimental trials. Furthermore, in the construction of our framework's CS module, we experiment with four similarity functions: cosine similarity, negative Euclidean distance, and dot product. Notably, Table 3.8 illustrates that the cosine similarity function outperforms the others.

Lastly, in the retrieval task, we employ the Faiss library [107] to index the embeddings of (verified) documents, thereby enhancing the speed of similarity searches. The ramifications of this indexing technique will also be subjected to evaluation.

3.6.3 Results

Modality Fusion Strategy

In our initial experiment, we assess the suitability of various (multimodal) architectures for integrating text and image modalities. We compare straightforward multimodal fusion methods, such as ARCNN [213] (employing late fusion of independent text and image classifiers) and ConcatBERT [254] (applying early fusion of BERT and ResNet embeddings), with more recent approaches like SAFE [327], VisualBERT [137], and MMBT [112]. Our evaluation focuses on the multimodal fake news detection task, using the FakeNewsNet dataset [247], which includes news articles from political (Politifact, 624 instances) and gossip (GossipCop,

Table 3.9. Classification results for the preliminary study. (bold indicates the best result, underline the first runner up)

		Politifact		GossipCop	
Strategy	Method	Accuracy	F1	Accuracy	F1
simple	ARCNN	.791	.766	.804	.713
	ConcatBERT	.794	.824	.793	.805
advanced	SAFE	.874	**.896**	.838	.895
	VisualBERT	.790	.835	.847	<u>.897</u>
	MMBT	.744	<u>.836</u>	.863	**.901**

10,259 instances) domains. We assess the models based on their accuracy and F1-score performance metrics. The results, as presented in Table 3.9, consistently demonstrate the superior performance of advanced fusion strategies, particularly pronounced on the larger GossipCop dataset, whose size is much bigger and thus better suited for a data-driven training process.

Overall, these findings indicate that advanced fusion strategies for combining image and text modalities and that leveraging the knowledge embedded in pre-trained visual-language models could significantly enhance the capabilities of contemporary fake news detection systems. Consequently, we chose to implement the Multimodal Fusion Component (MFC) using the MMBT model [112]. Notably, we employed two distinct MMBT models, each with untied weights, to represent claims (c) and documents (d) due to their distinct lengths [321].

Overall, these results underscore the potential benefits of employing advanced fusion techniques for integrating image and text modalities and capitalizing on the knowledge ingrained in pre-trained visual-language models to enhance the capabilities of contemporary fake news detection systems. From our perspective, we have opted to implement the MFC component using the MMBT model [112]. Notably, we have employed two separate MMBT models, each with independent weights, to represent claims (c) and documents (d) due to their varying lengths [321].

Reranking Evaluation

We select 7 state-of-the-art competitors representing all rerankers categories mentioned in Section 3.3.2: (1) BM25 algorithm [220] is the standard baseline; (2) MatchPyramid [195], KNRM [305], BERT [52] and MAN [285] are interaction-based approaches; (3) NSMN [190] and sentenceBERT [217] are representation-based approaches.

Table 3.10 provides a comprehensive overview of the reranking performance for both datasets. Notably, the BM25 baseline demonstrates suboptimal results, which is unsurprising given that the fact-checking domain demands more nuanced semantic matching strategies, rather than relying solely on word-level matching techniques. This distinction is pivotal because input claims, particularly those sourced from social media, often undergo paraphrasing using different terminology and are discussed in a more formalized manner within corresponding verified documents.

Additionally, a compelling competition emerges between BERT and MAN for the second-best position in the rankings, despite BERT's lack of explicit utilization of multimodal inputs. To elaborate, on the Politifact dataset, BERT outperforms MAN, largely attributable to its ability to leverage unsupervised pre-trained knowledge. The relatively modest dataset size poses challenges for MAN, which must train its network from scratch, ultimately hindering its capacity to extract robust representations, even with multimodal data. In contrast, the scenario undergoes a reversal when evaluating the Snopes dataset: its larger size empowers MAN to effectively discern latent and intricate relationships between textual and visual elements, resulting in superior performance relative to BERT.

Nevertheless, our model outperforms all competitors because it effectively combines the great advantages of the two front-runner methods, i.e. MAN and BERT. As a matter of fact, it employs both text and images modalities and leverages the wide knowledge of transformer-based architectures.

Shifting the focus to efficiency analysis, Table 3.11 shows the average runtime values, along with their corresponding 95% confidence intervals, required for reranking the top-k potentially relevant documents within the Snopes dataset. Specifically, we compare our system with the top-two reranking performers, namely BERT and MAN. Our method consistently excels, primarily due to the remarkable efficiency exhibited by the selected

Table 3.10. Re-ranking performance: BM25 represents our baseline, the second and third groups refer to *interaction-based* and *representation-based* methods, respectively. The second column highlights multimodal approaches. (bold indicates the best result, underline the first runner up)

		Politifact					Snopes				
Method	MM	HIT@3	HIT@5	NDCG@1	NDCG@3	NDCG@5	HIT@3	HIT@5	NDCG@1	NDCG@3	NDCG@5
BM25		.379	.433	.182	.292	.313	.329	.395	.206	.276	.304
MatchPyramid		.455	.503	.294	.389	.408	.660	.733	.481	.585	.616
KNRM		.636	.722	.422	.549	.585	.697	.761	.489	.610	.639
BERT		<u>.786</u>	<u>.856</u>	<u>.505</u>	<u>.675</u>	<u>.704</u>	.778	.828	.588	.699	.720
MAN	✓	.732	.786	.551	.654	.676	<u>.877</u>	**.906**	<u>.743</u>	<u>.822</u>	<u>.834</u>
NSMN		.551	.679	.379	.477	.531	.703	.778	.458	.601	.632
sentence-BERT		.139	.176	.059	.098	.113	.123	.197	.034	.084	.114
Ours	✓	**.918**	**.922**	**.701**	**.712**	**.721**	**.882**	<u>.902</u>	**.851**	**.860**	**.862**

Table 3.11. Re-ranking runtimes per claim. * indicates statistical significance, at $p = .05$, between the best and the second best methods.

Method	Politifact	Snopes
BERT	$.9431 \pm .0232$	$.9833 \pm .0176$
MAN	<u>$.0934 \pm .0005$</u>	<u>$.0938 \pm .0012$</u>
Ours	$\mathbf{.0606 \pm .0002}^*$	$\mathbf{.0603 \pm .0001}^*$

CS module. However, in complete fairness, we highlight that from the perspective of practical applicability within the domain, all reported runtimes remain well within the bounds of acceptability for the reranking task. This underscores that MAN and BERT also represent efficient alternatives, even when tasked with reranking a more modest subset of 50 documents.

Retrieval Evaluation

When studying competitor retrievers, our choice has been driven by the efficiency requirement of such systems. We selected 4 state-of-the-art retrievers: BM25 [220], docT5query [194], colBERT [110] and concatBERT [254].

Figure 3.4 provides a visual representation of the hit ratios observed in both datasets while varying the number of retrieved documents. Notably, the performance metrics at $k \in 50, 100, 200$ are of particular significance

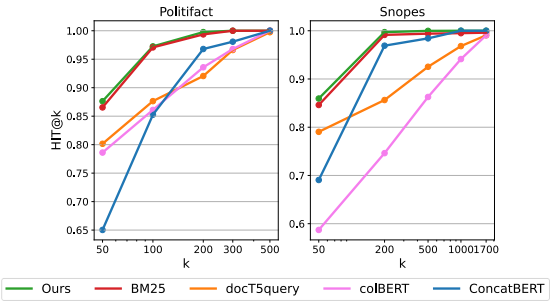


Figure 3.4. Retrieval hit ratios varying the number of retrieved documents from 50 to the full document corpus.

as they correspond to operational settings where the retriever operates in tandem with a robust reranker.

Interestingly, the BM25 baseline outperforms the majority of its competitors, which aligns with observations from other information retrieval tasks, where classical probabilistic approaches continue to deliver commendable retrieval performance [30]. However, it is noteworthy that our model distinguishes itself by being the sole approach that consistently matches or slightly surpasses the baseline's performance on both datasets.

Conversely, the neural network-based competitors exhibit a comparatively less competitive stance, consistently trailing our model by at least 5 HIT-points. Specifically, docT5query performs slightly better on the Snopes dataset because its claims are less formal, i.e. they refer to Twitter posts, thus making generating artificial query simpler. Finally, considering multimodal models, the lackluster performance of ConcatBERT underscores the critical significance of the modality fusion strategy within this domain.

Shifting the focus to efficiency analysis, Table 3.12 presents the average runtimes, along with their respective 95% confidence intervals, for the Snopes dataset, considering the retrieval of $k \in 50, 200$ documents per claim. ConcatBERT's performance is omitted from the report as the inference process is identical to that of our model.

As anticipated, it is evident that BM25 and docT5query exhibit statistically equivalent efficiency, with the latter method essentially employing the former algorithm with modified documents. In contrast, our model

Table 3.12. Retrieval runtimes per claim. * indicates statistical significance, at $p = .05$, between the best and the second best methods.

Method	$k = 50$	$k = 200$
BM25	$\underline{.0154 \pm .0002}$	$.0369 \pm .0003$
docT5query	$.0155 \pm .0016$	$\underline{.0296 \pm .0002}$
colBERT	$.1548 \pm .0003$	$.1556 \pm .0003$
Ours	$.0613 \pm .0001$	$.1316 \pm .0001$
Ours (index)	$\mathbf{.0001 \pm .0000}^*$	$\mathbf{.0001 \pm .0000}^*$

underscores the substantial potential for efficiency gains through indexing techniques, as evidenced by the two orders of magnitude difference in runtime between the inference step with and without indexing. However, it is worth noting that colBERT's efficiency remains the lowest due to the higher computational cost associated with its similarity function.

Finally, we assess the memory overhead, noting that the MMBT model size is 646MB, comprising 6.3 million parameters, and can comfortably accommodate server-side GPUs. The storage overhead depends primarily on the indexing technique and amounts to 1.2GB for our Faiss index. Nevertheless, we acknowledge that this analysis's applicability may be limited by the dataset's document corpus size.

Ablation Study

In the realm of multimodal systems, it is paramount to discern the individual contributions of each modality to the ultimate performance. To address this, we conduct an ablation study focusing on re-ranking settings, wherein we consider exclusively textual inputs, solely image inputs, and their amalgamation. It is worth emphasizing that the input modality exclusively affects our framework's MFC component, necessitating appropriate masking of its textual or visual layers.

Table 3.13 shows the results for the Snopes dataset. We omit the Politifact dataset since its analysis provides similar insights. Firstly, it is evident that the image modality offers significantly greater informativeness

Table 3.13. Ablation study: performance of re-ranking when using only text, only images and their combination. (bold indicates the best result, underline the first runner up)

Strategy	HIT@3	HIT@5	NDCG@3	NDCG@5
only text	.242	.251	.223	.233
only images	.635	.639	.618	.627
text + images	**.882**	**.902**	**.860**	**.862**

than the textual modality. This observation likely arises from the fact that fact-checking articles, while rephrasing and scrutinizing the text of input claims, frequently incorporate very similar images pertaining to the claims under examination. Secondly, the combination of both modalities substantially enhances the model's effectiveness compared to each modality in isolation. This outcome is contingent on the premise that multimodal content becomes consequential primarily when text is coupled with specific images. In these scenarios, the verified document evaluates the relationship between modalities, rather than assessing the individual modalities in isolation.

Error Analysis

To gain a qualitative understanding of our model's behavior, we conduct an error analysis. This analysis aims to elucidate the model's strengths and limitations, particularly in retrieval scenarios. We compared our results with those of BM25, seeking instances where our system outperforms BM25 and identifying the types of input claims that might still yield incorrect results.

Figure 3.5 presents two illustrative examples of such scenarios. In the left example, sourced from the Snopes dataset, the input claim explicitly references the attached image, providing no contextual information in its textual modality. Consequently, the BM25 baseline, along with most textual approaches, fails to retrieve the correct document within the top-50 positions. In contrast, our model accurately ranks it as the top result without requiring any re-ranking algorithm. In cases like this, the inclusion of

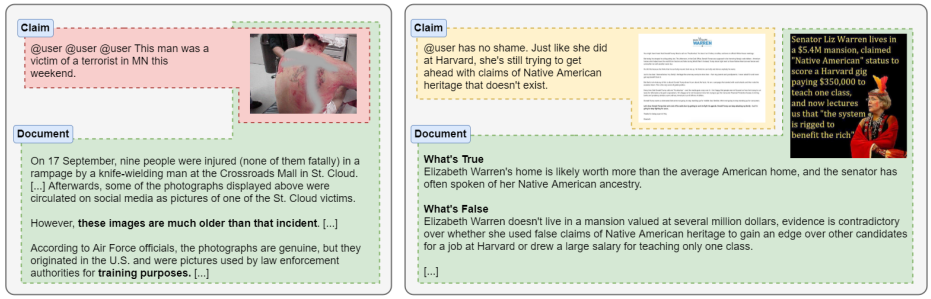


Figure 3.5. Two examples from Snopes (left) and Politifact (right) datasets. Each gray box includes the input claim with an extract of its verified document. Red (yellow) shade refers to a false (mixture) claim, while green one highlights the documents' content. The left image has been edited for its violent content.

image modalities proves invaluable in facilitating correct matches. Such scenarios are commonplace on social networks, where users frequently employ images to evoke emotional responses.

On the right-hand side, an example from the Politifact dataset illustrates a situation in which our model, although still retrieving the correct document within the top 50 positions, performs less effectively than the BM25 baseline. We attribute this outcome to two factors. First, the image modality is somewhat ambiguous, as its content primarily consists of textual elements, which can pose challenges for the MFC module. Extracting text from images could potentially enhance the system's performance, as suggested by prior research [285]. Second, the verified document is notably intricate, as it deliberates on the veracity of concepts presented in the input claim. Intuitively, capturing a semantic representation in such cases demands a more intricate and reasoning-based approach. Conversely, a simpler probabilistic approach, like BM25, can effectively match the most significant and frequently occurring words (e.g., "Harvard," "heritage") to yield optimal results.

3.6.4 Discussion & Limitations

The conducted experiments demonstrate the effectiveness and efficiency of our proposed system in both reranking and retrieval scenarios.

Specifically, in the reranking case, our system consistently outperforms all baseline methods. In the retrieval scenario, it stands out as the only model that achieves a slight improvement over the simplest BM25 algorithm [220]. Moreover, our ablation study underscores the substantial advantage of leveraging multimodal data to tackle this task. Overall, we posit that enhancing performance in the examined task hinges not only on the inclusion of images in the learning process but also on the design of an effective strategy to combine different modalities. Exploiting the extensive (pre-trained) knowledge of multimodal transformers proves significantly more effective than training a neural network from scratch, as seen in previous work [285].

While our proposed multimodal system has demonstrated strong performance in the context of detecting previously fact-checked information, we acknowledge two important limitations: (i) despite being real-world datasets, their relatively small size raises concerns about the seamless generalization of our results to operational settings; (ii) our system necessitates an initial (small) set of labeled data to facilitate the training process. Meeting this requirement may pose challenges in scenarios involving low-resource languages or evolving topics where verified information is scarce. However, it is crucial to emphasize that the task under examination constitutes a stage within a broader (multimodal) fact-checking pipeline. Consequently, when an input claim cannot be matched with any document, it can be flagged for further investigation by human experts.

3.7 Case study: Ukraine-Russia conflict

On February 24, 2022, Russia initiated its ongoing invasion of Ukraine, a geopolitical event that reverberated globally[11]. Subsequently, concerns arose regarding the presence of Russian disinformation campaigns on online social media platforms. These campaigns aimed to reframe the invasion as a "special operation" against alleged Nazis in Ukraine or to lay blame on NATO's expansion as the cause for the invasion [198, 205, 93, 204]. It is important to note that Russian interference in the democratic processes of other countries is not a recent phenomenon, as extensively documented in the context of the 2016 U.S. Presidential election [17, 154].

[11] https://en.wikipedia.org/wiki/2022_Russian_invasion_of_Ukraine

This case study focuses on utilizing the verified claim retrieval task to automatically identify false and unsubstantiated claims – verified by news agencies and fact-checking organizations (e.g., Politifact, Snopes) – that were shared on Twitter during the initial stages of the Ukraine-Russia conflict [127]. To this end, we employ an automated pipeline comprising two models: (i) the *claim detection* model determines whether an input tweet contains a false claim; (ii) assuming the input tweet reports a claim, the *verified claim retrieval* model ranks these claims based on their relevance to the input tweet. To evaluate our system, we collect a dataset consisting of 83 false claims that were verified in the initial weeks of the invasion and annotate 5,872 original tweets based on the claim(s) they discuss.

Our results demonstrate the effectiveness of our approach in retrieving claims referenced in the input tweets. Additionally, we assess how well our models generalize to new claims that were not part of the training data, providing insights into their adaptability and robustness. Overall, our results underscore the practical relevance of the claim retrieval task. It facilitates the retrieval of previously verified claims, obviating the need for redundant verification efforts, and offers insights into the dissemination patterns of these claims across social media platforms.

3.7.1 Case Study Design

In this section, we describe the data collected for the analysis and the methodology employed to annotate tweets and their corresponding claims. Then, we present our methodological framework by formally defining the *claim detection* and the *verified claim retrieval* tasks and corresponding models.

Data Collection To identify false and unsubstantiated claims disseminated online at the beginning of the invasion, we have harnessed the resources provided by the Russia-Ukraine ConflictMisinfo Dashboard[12]. This repository compiles a comprehensive inventory of verified claims and rumors, both true and false, corroborated by fact-checking authorities such as *USA Today* and *Snopes*. Specifically, we collect 83 English false claims that underwent verification within the period spanning from February 22nd

[12] https://conflictmisinfo.org

to March 1st.

Furthermore, we have tapped into an existing dataset [205] meticulously curated to encompass tweets related to the conflict. This dataset was constructed through a snowball sampling technique, targeting keywords in English, Russian, and Ukrainian languages, with data being collected via Twitter's Filter v1.1 Streaming API[13]. Our specific interest lies in English-language tweets disseminated during the initial weeks of the invasion, specifically, from February 22nd, 2022, to March 8th, 2022. It is important to note that we have extended our data collection window to encompass one week beyond the date when the aforementioned false claims were officially verified, which was March 1st. This extension allows us to capture the trajectory of these claims as they proliferate on the Twitter platform. In totality, during this observation period, our dataset encompasses over 2 million English tweets containing original content. To ensure the exclusion of duplicate content, we have deliberately omitted retweets, as their textual content is entirely identical.

Data Annotation In light of the amassed fact-checked information, our objective is to identify tweets that report these verified claims. Given the impracticality of manually annotating the entire corpus of tweets within our dataset, we employ a machine learning-driven annotation strategy tailored to maximize the likelihood of discovering tweets relevant to the claims under scrutiny.

To execute this strategy, we initially leverage the RoBERTa transformer [333] to extract vector embeddings for both claims and tweets. Subsequently, we calculate the cosine similarity between each claim and tweet in our dataset, retaining the top-100 most similar tweets for each claim. Consequently, this process yields 8,300 distinct tweet-claim pairs that underwent a meticulous manual labeling procedure. It is worth highlighting that our selection of the RoBERTa transformer is underpinned by its ability to provide higher similarity scores compared to alternative transformers, such as *ms-marco-MiniLM-L-4-v2* and *quora-roberta-base*. This choice allows us to maximize the likelihood of uncovering matching pairs.

Subsequently, we annotate these tweet-claim pairs, adhering to a stringent definition of relevance. In this context, a tweet is deemed relevant if it

[13] https://developer.twitter.com/en/docs/twitter-api/v1

Table 3.14. Examples of some tweet-claim pairs annotated in the dataset

No.	Claim	Tweet
1	Russian President Vladimir Putin threatened India against getting involved in the Ukraine crisis.	Putin has warned India that don't try to interfere in their matter, otherwise be ready to face the consequences
2	The President Of Ukraine, Volodymyr Zelenskyy, Is On The Ground With His Fellow Troops	Volodymyr Zelenskyy the president of Ukraine has decided to stay behind and fight among his people against the Russian army send to kyiv [...]
3	The Russian armed forces are not striking at the cities of Ukraine; they are not threatening the civilian population.	It is clear that the Russian army does not want to harm civilians, its strikes were directed only at military targets, [...] life seems almost normal in Kiev.
4	The Russian armed forces are not striking at the cities of Ukraine; they are not threatening the civilian population.	Russian forces continue strikes in multiple cities [...]. This is premeditated mass murder and must be responded to as such.

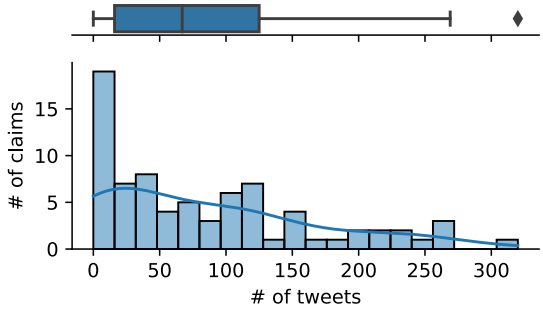


Figure 3.6. Distribution of the number of tweets with respect to the number of claims

explicitly references the same entities and events as those delineated in the fact-checked information. Some illustrative examples of false claims and corresponding matched tweets are presented in Table 3.14. These examples underscore how a matched tweet can discuss a claim without expressing a stance (examples 1 and 2) or may align with or contradict the claim (examples 3 and 4). Furthermore, it is noteworthy that the number of unique tweets differs from the number of pairs, as a single tweet can be associated with multiple claims. Specifically, we identify 5,872 unique tweets that are interconnected with the 83 claims.

Our manual annotation process yields 2,359 tweets (out of 5,872 – 40.2%) that are linked to at least one claim. Figure 3.6 elucidates the

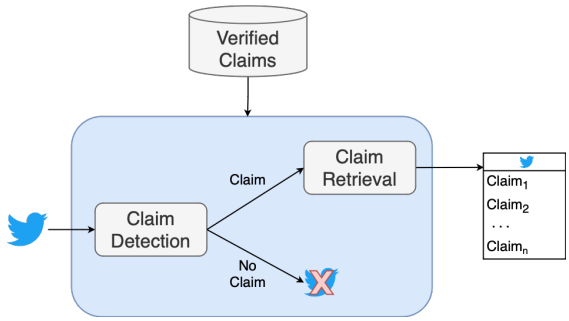


Figure 3.7. Our proposed methodological framework: the *claim detection* model detects whether the input tweet reports a fact-checked (false) claim. If a claim is detected, the *claim retrieval* model retrieves the most relevant claims (within the corpus of "Verified Claims") related to the tweet.

distribution of tweets concerning the number of claims: the majority of tweets pertain to fewer than five claims, with only 13 tweets engaging with more than 10 claims. Overall, each claim has at least one corresponding matching tweet, with the most frequently matched claim featuring 100 associated tweets. Conversely, we ascertain that 3,513 tweets out of 5,872 (59.8%) do not reference any claims.

Our Framework Figure 3.7 depicts the methodological framework underpinning our two-step pipeline. The initial step revolves around a *claim detection* model, tasked with determining whether an input tweet engages with a (false) claim from a predetermined catalog of already verified claims. Subsequently, if the input tweet is indeed associated with a false claim within the *Verified Claims* repository, the *verified claim retrieval* model takes on the role of ranking the claims in this repository in accordance with their relevance to the input tweet.

More specifically, we formulate the claim detection problem as a binary classification task. Given an input tweet t, the primary aim is to ascertain whether t encompasses a false claim from the compiled corpus of *Verified Claims*. To achieve this, we harness BERT-based language models [52]. These models have demonstrated their efficacy in numerous text classification tasks [142] by virtue of their comprehensive knowledge garnered during the pre-training phase.

In addition, assuming a realistic scenario where only a minority of tweets report specific factual events related to the war [12] (i.e., 2.3k out of 5.9k tweets identified by our annotation procedure), we evaluate our model by performing random oversampling of the minority class (i.e., tweets with a specific claim). In other words, we randomly replicate tweets containing a claim so as to have the same number of tweets belonging to the positive and negative classes.

Conversely, the *verified claim retrieval* model functions within the same parameters delineated for the detection of previously fact-checked information (see Section 3.5.1). In essence, when presented with an input tweet t containing a (false) claim, it orchestrates the ranking of a cluster of (already-)verified claims $\{c_1, c_2, \cdots, c_n\}$ contingent on their relevance to t. To perform the (supervised) training process, we find that random negative sampling yields good results in our use scenario, i.e., for each tweet we randomly select 10 unrelated claims. The final model is BERT-based cross-encoder model, which demonstrated superior ranking performance in our benchmark study (see Section 3.4.4).

3.7.2 Experimental Protocol

For each task, we perform a 5-fold cross-validation to evaluate the performance of our models. In particular, we consider two evaluation settings:

- Leave Tweet Out (LTO) Assessment: In both the claim detection and retrieval tasks, we ensure that tweets in the training, validation, and testing sets are mutually exclusive.
- Leave Claim Out (LCO) Assessment: In the claim detection task, we guarantee that tweets in the training, validation, and testing sets pertain to distinct claims. For the claim retrieval task, we ensure that claims in the tweet-claim pairs across the training, validation, and testing sets do not overlap.

The LTO assessment aligns with the conventional evaluation method applied in supervised machine learning contexts and remains consistent with established experiments in prior research [235, 285, 182]. Conversely, the LCO assessment emphasizes the model's ability to generalize to unseen claims, mirroring real-world conditions where claim detection and retrieval

should function effectively on tweets and claims that the models have not encountered during training.

In the claim detection task evaluation, we compare our BERT-based model with a TF-IDF baseline [118]. Furthermore, we assess the efficacy of the oversampling strategy used to balance the annotated dataset. For the claim retrieval task, we gauge the ranking performance of our BERT-based cross-encoder against that of Sentence-BERT [217].

Evaluation Data and Metrics For the claim detection task, we leverage the annotated tweets described earlier. Our dataset encompasses 2,359 tweets (out of 5,872, constituting 40.2%) that reference at least one (false) claim, while 3,513 tweets (out of 5,872, accounting for 59.8%) do not contain any claims. Given the classification nature of this task, we gauge our models' performance through binary classification metrics, encompassing precision, recall, F1-score, and accuracy.

Conversely, in the claim retrieval task, we utilize the subset of 2,359 tweets linked to the 83 claims obtained from the Russia-Ukraine Conflict-Misinfo Dashboard. As previously detailed, we construct a dataset comprising 40,007 tweet-claim pairs. This dataset encompasses 3,637 "positive" tweet-claim pairs (indicating tweets that match claims) and 36,370 "negative" tweet-claim pairs (indicating tweets that do not match any claims). Consistent with previous studies [285, 235], we evaluate the claim retrieval model's performance using the HitRatio@K metric, which assesses whether the correct claim (i.e., the claim corresponding to the tweet) is among the top-k results of the ranking.

Experimental Setup For the claim detection model, fine-tuning has been performed using the Huggingface library, specifically on the *BERT-base-cased* model. The fine-tuning process span five epochs, utilizing the Adam optimizer, categorical cross-entropy as the loss function, and employing a batch size of 20 tweets.

Regarding the verified claim retrieval model, fine-tuning is executed on the *stsb-roberta-base* cross-encoder available within the SBERT library. In particular, we train the model for three epochs using categorical cross-entropy loss and 16 tweet-claim pairs as batch size.

All experiments are conducted on a machine with hardware specifica-

Table 3.15. Claim detection: performance with and without random oversampling

Metric	No Oversampling	With Oversampling
Precision	79.90%	**81.39%**
Recall	80.19%	**80.24%**
F1-score	79.35%	**80.57%**
Accuracy	80.11%	**81.59%**

tions including an Intel Xeon-4116 CPU, 32 GB of RAM, and a NVIDIA A100 GPU.

3.7.3 Case Study Results

Claim Detection Table 3.15 presents a comparative analysis of classification performance, both with and without the random oversampling strategy. Evidently, the application of oversampling to the minority class consistently enhances the predictive capabilities of the claim detection model, resulting in improved performance across all classification metrics.

For a more granular perspective, Table 3.16 provides a breakdown of performance metrics for each class within the claim detection task. Notably, our BERT-based model consistently outperforms the TF-IDF baseline for both classes across all metrics. Interestingly, the model demonstrates superior performance on the negative class, indicating higher precision and recall in detecting tweets that do not contain any claims. We omit the accuracy metric as it is equivalent to the recall computed for the single class.

Lastly, Table 3.17 offers an aggregated performance assessment under both LTO and LCO evaluation settings. Remarkably, the LCO evaluation poses a more challenging scenario, as the model must contend with claims it has never encountered during its training phase.

Claim Retrieval Table 3.18 provides a comprehensive overview of the ranking performance of our claim retrieval model, in comparison with sentence-BERT [217]. In both evaluation settings, our model surpasses the baseline. This performance differential becomes even more pronounced

Table 3.16. Claim detection: performance comparison per class, and their 95% confidence interval, between TF-IDF baseline and our approach (bold indicates best on average, * indicates statistical significance ($p < 0.01$)

Metric	Claim (Positive class)		No Claim (Negative class)	
	TF-IDF	Ours	TF-IDF	Ours
Precision	$67.88\% \pm 1.57\%$	$\mathbf{79.75\% \pm 3.05\%}^*$	$80.97\% \pm 2.92\%$	$\mathbf{83.02\% \pm 1.87\%}$
Recall	$73.01\% \pm 4.99\%$	$\mathbf{73.31\% \pm 2.09\%}$	$76.84\% \pm 0.81\%$	$\mathbf{87.17\% \pm 3.85\%}^*$
F1-score	$70.33\% \pm 3.05\%$	$\mathbf{76.17\% \pm 1.71\%}^*$	$78.84\% \pm 1.40\%$	$\mathbf{84.97\% \pm 1.39\%}^*$

Table 3.17. Claim detection: LTO and LCO assessment

	Settings	
Metric	**LTO**	**LCO**
Precision	$81.39\% \pm 1.14\%$	$78.07\% \pm 1.48\%$
Recall	$80.24\% \pm 1.36\%$	$77.58\% \pm 1.67\%$
F1-score	$80.57\% \pm 1.38\%$	$77.63\% \pm 1.60\%$
Accuracy	$81.59\% \pm 1.42\%$	$78.03\% \pm 1.56\%$

when scrutinizing the top positions in the ranking ($k \in 1, 3$). Such results underscore the potential of our system in simplifying the tasks of fact-checkers, as it adeptly identifies tweets discussing previously verified false claims. Furthermore, it is noteworthy that the performance marginally favors the LTO setting over the LCO setting, albeit the difference is less pronounced compared to our observations in the claim detection task.

Table 3.18. Claim retrieval: performance comparison, and their 95% confidence interval, between the sentence-BERT baseline and our approach (bold indicates best on average, * indicates statistical significance ($p < 0.01$)

Setting	Model	HitRatio@k				
		$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
LTO	Sentence-BERT	$85.25\% \pm 2.07\%$	$94.87\% \pm 1.69\%$	$97.24\% \pm 0.96\%$	$98.77\% \pm 0.43\%1$	$99.27\% \pm 0.39\%$
	Ours	$\mathbf{86.05\% \pm 0.95\%}$	$\mathbf{96.35\% \pm 0.71\%}^*$	$\mathbf{98.04\% \pm 0.57\%}^*$	$\mathbf{99.27\% \pm 0.36\%}$	$\mathbf{99.78\% \pm 0.11\%}^*$
LCO	Sentence-BERT	$77.60\% \pm 0.1196$	$95.68\% \pm 6.74\%$	$98.01\% \pm 3.66\%$	$99.63\% \pm 0.66\%$	$99.78\% \pm 0.00\%$
	Ours	$\mathbf{82.25\% \pm 10.81\%}^*$	$\mathbf{96.42\% \pm 2.59\%}$	$\mathbf{98.26\% \pm 1.45\%}$	$\mathbf{99.88\% \pm 0.02\%}$	$\mathbf{99.96\% \pm 0.00\%}$

Evaluation in the wild We put our claim detection and retrieval models to the test on the 2M Twitter dataset to gauge their real-world applicability. In this instance, we lack the LTO and LCO annotations, precluding formal quantitative analysis. Nevertheless, we opt to randomly sample 100 tweets flagged by the claim detection model as reporting a claim, alongside another 100 designated as not discussing any claims. We then conduct manual verifications to ascertain the models' accuracy, determining whether a tweet indeed engages with a false claim and, if so, which specific false claim it pertains to.

Out of the 100 tweets identified by the claim detection model as unrelated to any false claims, we uncover 12 misclassifications, signifying instances where the tweets are, in fact, connected to a false claim. This finding reaffirms the claim detection model's high precision (exceeding 80%) for the negative class, as expounded in Table 3.16. Conversely, among the 100 tweets predicted by the claim detection model as discussing a false claim, we encounter 64 false positives. In other words, these tweets do not delve into any false claims but instead generally disseminate information concerning the Ukraine-Russia conflict. This outcome validates the observed performance divergence between the positive and negative classes of the claim detection model, potentially attributable to the imbalanced dataset.

Furthermore, we apply the claim retrieval model to the 36 tweets genuinely tied to false claims. Our analysis reveals that, among these, the correct corresponding false claim is retrieved 34 times out of 36. These results confirm the promising performance of the *verified claim retrieval* model but highlight the limitations of the *claim detection* model, which overestimates the number of tweets discussing false claims.

3.8 Conclusions and Future Works

In this chapter, we embarked on a comprehensive exploration of the multifaceted landscape surrounding the detection of previously fact-checked information in the vast expanse of social media. Our efforts encompassed a rigorous benchmark study, the development of a novel multimodal model, and a real-world case study applied to the context of the Ukraine-Russia conflict. These endeavors collectively unravel the complex-

ities of the domain, spotlighting its challenges and showcasing promising directions for future research.

Our initial contribution materialized through the benchmark study, an exhaustive examination of how existing methodologies from information retrieval literature could support the task of detecting previously fact-checked information. By meticulously evaluating various models under both re-ranking and retrieval scenarios, we gained vital insights into the performance landscape showing that neural methodologies as the most promising auspicious avenues for enhancing performance. This benchmark not only laid the foundation for our subsequent work but also provides an invaluable reference point for future research in the field.

In tandem with the benchmark, we introduced a novel framework to address the multimodal version of the task. The proposed system capitalizes on the synergy between textual and image-based information, offering a potent mechanism for enhancing the detection of previously fact-checked claims. Through a meticulous examination, we demonstrated that our model outperforms existing state-of-the-art techniques, underscoring the importance of multimodal fusion strategies in this domain.

Furthermore, we applied our methodological framework to a real-world scenario investigating false claims circulating on Twitter during the onset of the Ukraine-Russia conflict. This case study illuminated our models' practical utility in identifying and retrieving false claims within a dynamic and evolving information landscape.

Our research opens doors to a multitude of promising avenues. First, expanding the scope of our multimodal model to cater to a wider range of languages and social media platforms can significantly enhance its applicability. Second, exploring more advanced pre-training techniques and transfer learning strategies could further boost the model's performance. Finally, a crucial future endeavor should center on the acquisition of a more extensive dataset. This dataset should encompass a diverse array of claims and their corresponding verification documents, sourced from various origins and encompassing different evolving topics. This broader foundation will be instrumental in refining and enhancing the model's performance, especially in scenarios where knowledge is subject to rapid change.

Chapter 4

KERMIT: Knowledge-EmpoweRed Model In harmful meme deTection

4.1 Research context and Problem Definition

The utilization of internet memes, initially celebrated for their comedic qualities, has undergone a transformative evolution beyond their original intent, manifesting as a conduit for the dissemination of hate speech, noxious content, and disinformation within the realm of social media platforms. A salient exemplar of this phenomenon is the "Pepe The Frog"[1] meme, an entity that has undergone a symbolic metamorphosis intricately interwoven with far-right and white supremacist ideologies [80]. Similarly, the sharing of a meme by former US President Donald Trump, featuring two QAnon slogans, is often regarded as one of his most explicit acknowledgments of this conspiracy theory[2].

As a consequence, the precise and efficient identification of detrimental memes has emerged as an imperative endeavor in upholding online safety

[1] https://wikipedia/Pepe_The_Frog.jpg
[2] https://cnn.com/qanon-fans-donald-trump.html

and cultivating responsible online behavior. In recent years, this exigency has attracted attention from scholars and practitioners in diverse research domains, encompassing natural language processing [239, 53], computer vision [114, 68], and social media analysis [315].

Specifically, the detection of harmful memes presents unique challenges distinct from other forms of hate speech detection, as memes frequently leverage a fusion of visual and textual elements to convey their messages. Consequently, prior endeavors have predominantly concentrated on a meticulous examination of the intrinsic constituents of memes by employing multimodal representation learning methodologies aimed at scrutinizing intra- and inter-modality signals between image and text components [332, 242]. In particular, cutting-edge architectures harness pre-trained visual-language models (such as VisualBERT [137], CLIP [210], MMBT [112]) to leverage the amalgamated information stemming from both textual and visual elements, thereby enabling the capture of the nuanced meanings embedded within a meme.

However, harmful memes rely on cultural references, shared knowledge, and social context to convey their intended meaning [119, 141]. In essence, understanding the message embedded within a meme necessitates a background of tacit knowledge, which remains implicit within the meme itself, contingent upon the viewer's familiarity with specific contextualized facets of the world. To illustrate this concept, we can examine the meme depicted in Figure 4.2 as a case study. On initial inspection, the image of an attractive woman alongside accompanying text may not inherently appear offensive. However, upon deeper analysis and contextualization within the realm of real-world knowledge, it becomes evident that the meme conveys a contentious and provocative statement regarding beauty and race. It insinuates that being of white ethnicity intrinsically conveys superiority or desirability in terms of physical attractiveness, attributing the woman's beauty to her racial background.

Prior research has largely overlooked the challenge of incorporating this contextual knowledge into multimodal models for meme analysis. This oversight assumes that the pre-training process, which encompasses contrastive learning [313], masked-language modeling [270], and image-text matching [255], has already captured the requisite background knowledge. However, this approach presents limitations, as pre-training data

may not consistently encompass pertinent contextual information, particularly when addressing the subtle and nuanced aspects of meme content. Additionally, exclusive reliance on pre-trained knowledge can result in biases and inaccurate predictions, including those associated with religion, gender, and race [252]. Consequently, the integration of external knowledge into the classification process emerges as a pivotal necessity for the effective detection of harmful memes.

Based on these premises, we introduce a novel approach for the explicit integration of background knowledge into the process of identifying harmful memes. Our framework, named KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), comprises two pivotal steps. Firstly, KERMIT constructs a *knowledge-enriched information network* for the meme by amalgamating internal meme entities with pertinent external knowledge sourced from ConceptNet [259]. Subsequently, KERMIT employs a dynamic learning mechanism, harnessing memory-augmented neural networks and attention mechanisms, to discern the most informative segment of the *knowledge-enriched information network* for precise classification of harmful memes.

To the best of our knowledge, our study represents the first comprehensive endeavor to model meme-related knowledge, encompassing both the meme's internal entities and their interrelationships, in conjunction with external knowledge obtained from ConceptNet [259]. Significantly, while prior research has explored the inclusion of unstructured knowledge into the decision-making process, such as web entities detected within image modality [332] or semantic entities retrieved from ConceptNet [236], our proposed framework stands as the pioneering end-to-end solution dynamically learning the most pertinent knowledge for the task of hateful meme classification.

Our assessment is grounded in two benchmark datasets utilized in the Facebook Hateful Memes Challenge [114] and the Multimedia Automatic Misogyny Identification (MAMI) Challenge [68]. Notably, our findings substantiate that KERMIT proficiently retrieves pertinent contextual knowledge from ConceptNet and effectively employs it in the classification process, resulting in performance enhancements. Moreover, when compared with several other multimodal baselines, KERMIT consistently achieves superior classification performance.

To sum up, our contributions can be summarised as follows:

- We propose a novel approach, namely KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), which explicitly incorporates a comprehensive range of knowledge sources, including both internal meme entities and their relationships, as well as external knowledge from ConceptNet. This holistic knowledge-enriched information network sets the stage for a more nuanced understanding of memes and their contextual relevance.

- We design a dynamic learning mechanism powered by memory-augmented neural networks and attention mechanisms. This innovative approach enables KERMIT to autonomously identify and prioritize the most pertinent segments of the knowledge-enriched information network, enhancing the accuracy of harmful meme classification.

- Our experiments, conducted on benchmark datasets from the Facebook Hateful Memes Challenge and the Multimedia Automatic Misogyny Identification (MAMI) Challenge, underscore the effectiveness of KERMIT. It consistently outperforms several multimodal baselines, demonstrating its prowess in leveraging external knowledge for improved classification accuracy.

Overall, this work demonstrates the effectiveness of incorporating external knowledge into the classification process and sets a path for future research in the harmful meme detection, highlighting the significant role that artificial intelligence and knowledge discovery can play in improving content moderation.

4.2 Related Works

4.2.1 Harmful Meme Detection

The use of memes as a medium for disseminating harmful content has escalated in significance, prompting increasing concern within academic and practitioner circles in recent years [242]. The identification of detrimental memes has consequently garnered heightened attention, given the

distinct challenges posed by the multimodal nature of memes. These challenges often entail a fusion of textual and visual elements to convey messages, rendering conventional unimodal approaches that exclusively consider text or images inadequate for accurate detection of hateful content in memes [114]. Conversely, pre-trained visual-language models such as VisualBERT [137], CLIP [210], and MMBT [112] have emerged as promising solutions for detecting harmful memes. These models harness the amalgamated information from both textual and visual components to capture the nuanced meanings and contextual cues embedded within memes. Notably, participants in the *Facebook Hateful Meme Detection* challenge [113] have highlighted the potential performance enhancements achievable through ensemble approaches employing these multimodal models [332]. Furthermore, research has underscored the profound impact of the source domain of pre-training datasets on model capabilities [283].

Recent work by Dimitrov et al. [53] extends beyond conventional harmful meme detection by scrutinizing distinct persuasion techniques employed within memes, including the use of emotive language and derogatory terms. Similarly, Pramanick et al. [209] broadens the scope of hateful meme detection by introducing a novel classification framework that identifies the target of the meme, such as a public figure or a political party, through a blend of local and global contextual information gleaned from the meme's backdrop. Complementary investigations have demonstrated the efficacy of named entity recognition and object detection within these settings [241, 243]. Conversely, further studies [272, 95] address the interpretability limitations inherent in modern multimodal models and underscore the presence of biases against specific ethnic and gender groups, such as Muslims and women.

Notably, prior research [242, 113] has predominantly focused on the analysis of visual and textual aspects of memes, largely neglecting the substantial role that external knowledge plays in comprehending memes, especially harmful ones. Memes frequently rely on cultural references and social context to convey their intended meaning [119, 141]. Many studies presume that vision-language models acquire the requisite background knowledge during the pre-training phase. However, this approach presents two significant shortcomings. Firstly, it inherits any biases, including those related to religion, gender, and race, present in the pre-training dataset.

Secondly, there is no assurance that the pre-training dataset comprehensively encompasses relevant contextual information.

In light of these limitations, our approach diverges from the prevailing paradigm by advocating for the direct integration of human commonsense knowledge into the harmful meme classification process. To our knowledge, KnowMeme [236] currently represents the most pertinent research endeavor in pursuit of knowledge-informed hateful meme classification. This system constructs a graph that encapsulates meme content and associated knowledge, drawn from ConceptNet, subsequently performing graph classification to discern whether the meme qualifies as harmful. Nonetheless, our framework, KERMIT, distinguishes itself from KnowMeme in several fundamental aspects. Firstly, KnowMeme constructs the meme graph without considering relationships, while KERMIT incorporates a modeling approach that systematically incorporates relational information within the meme graph. Secondly, KERMIT introduces a dynamic learning mechanism to identify the most informative segments of the graph for meme classification. Finally, in contrast to KnowMeme's sole reliance on the meme graph for hateful classification, KERMIT embraces a hybrid approach that amalgamates explicit knowledge from the meme graph with a pre-trained vision-language model.

4.2.2 Memory-augmented Neural Networks

In the realm of informed decision-making, memory-augmented neural networks (MANNs) [298] represent a potent fusion of neural networks' proficiency in discerning intricate data patterns with the capacity to store and retrieve information from an external memory source. These models have showcased their efficacy across a diverse spectrum of tasks, spanning question answering [157], image captioning [65], and video analysis [196]. Concretely, MANNs extend the conventional neural network framework by incorporating an external *memory block* designed to house task-specific knowledge, encompassing contextual data, factual information, or domain expertise. Formally, this *memory block* comprises multiple *memory slots*, each serving as a fundamental repository of knowledge. The representation within each *memory bucket* is contingent upon the specific task and constitutes a deliberate design choice. Options for this representation range from unstructured text [225] to tables [84] and even graphs [173].

In the context of harmful meme detection, the incorporation of external knowledge assumes pivotal significance, facilitating a deeper comprehension of memes' intent, tonality, reliability, cultural references, and social context [236]. Consequently, our framework, KERMIT, capitalizes on a MANN framework to house the *knowledge-enriched information network*, representing the entities within a meme alongside their associated commonsense knowledge, drawn from ConceptNet. Specifically, KERMIT's *memory block* encompasses multiple *buckets*, each dedicated to storing a segment of the *knowledge-enriched information network*. Additionally, KERMIT leverages an attention mechanism to dynamically identify the most informative *bucket(s)* for the precise classification of harmful memes. This approach empowers our model to seamlessly integrate external knowledge into the decision-making process, augmenting its capacity to detect pernicious content in memes by factoring in contextual information and pertinent insights derived from external sources.

4.3 Problem Formulation

In the context of our study, we define an internet meme, denoted as $\mathscr{M}$, as a composite information entity comprising both an image, $\mathscr{I}_M$, and an embedded text, $\mathscr{T}_M$. Specifically, we characterize the set of entities within the meme as $\mathscr{S}_M = \{s_1, s_2, \cdots, s_h\}$, where each s_i represents a meaningful concept present either within the image $\mathscr{I}_M$ or the embedded text $\mathscr{T}_M$. These concepts may manifest as facial features within the image or named entities within the text, among other possibilities. Furthermore, we consider the relationships between these entities, forming the set $\mathscr{P}_M = \{p_{ij}\} = \{(s_i, s_j) | s_i, s_j \in \mathscr{S}_M\}$. For example, in the meme depicted in Figure 4.2, potential relationships may encompass the woman in the image smiling at the camera and the meme attributing her beauty to her white parents. Consequently, we define the *meme graph* as $\mathscr{G}_M = \{\mathscr{S}_M, \mathscr{P}_M\}$, which encapsulates the internal knowledge explicitly conveyed within the meme itself.

Given the premise that comprehending a meme necessitates background knowledge, we introduce an external knowledge base $\mathscr{K}$ and define the meme's background knowledge, denoted as $\mathscr{K}_M$, as the collection of knowledge facts within $\mathscr{K}$ that pertain to any entity within $\mathscr{M}$. Subse-

quently, we construct $\widetilde{\mathscr{G}}_M$ as the *knowledge-enriched information network* of the meme. In essence, $\widetilde{\mathscr{G}}_M$ augments the meme graph $\mathscr{G}_M$ with relevant external knowledge drawn from $\mathscr{K}_M$.

This formulation leads us to conceptualize the task of knowledge-informed harmful meme detection as a multimodal classification problem. Specifically, our objective is to learn a binary decision function denoted as $f : \mathscr{M} = \{\mathscr{I}_M, \mathscr{T}_M, \widetilde{\mathscr{G}}_M\} \rightarrow y \in \{0, 1\}$. Importantly, this formulation remains agnostic to the specific strategy employed for constructing $\widetilde{\mathscr{G}}_M$ and can be readily extended to accommodate multi-class scenarios.

4.4 Our Proposal

The architecture of the proposed framework is illustrated in Figure 4.1. KERMIT comprises three principal modules: the Knowledge Modeling and Organization Module (KMOM), the Knowledge Embedding Representation Module (KERM), and the Knowledge-Augmented Classification Module (KACM).

The KMOM module assumes responsibility for constructing the *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$ of the meme. This network amalgamates the meme's entities with pertinent background knowledge obtained from the external source $\mathscr{K}$. The objective here is to capture salient contextual information from external reservoirs, thereby enhancing the semantic understanding of the meme and ultimately improving the performance of hateful meme detection.

Subsequently, the KERM module is tasked with embedding $\widetilde{\mathscr{G}}_M$ into a lower-dimensional, continuous latent space. This embedding process retains the topological and semantic attributes of the network's nodes and edges while reducing the dimensionality.

Finally, the KACM module integrates information derived from the input meme $\mathscr{M}$ and its *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$. This integration enables the module to execute the hateful meme classification process, thereby determining whether the meme falls into the category of harmful content or not.

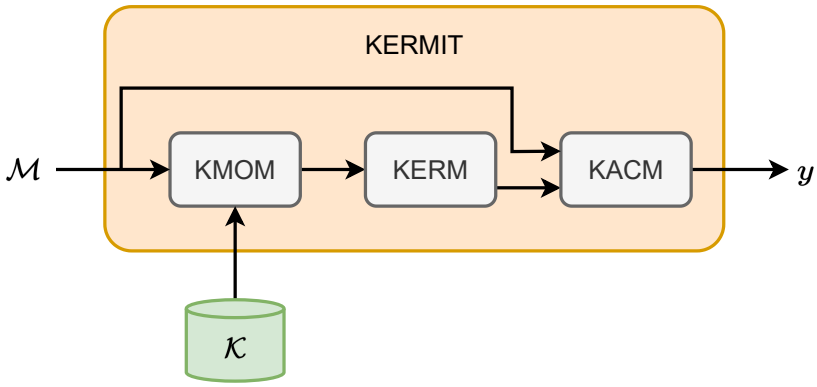


Figure 4.1. The high-level architecture of KERMIT: The KMOM module leverages the input meme $\mathscr{M}$ and an external knowledge base $\mathscr{K}$ and generates the meme's *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$. Subsequently, the KERM module embeds $\widetilde{\mathscr{G}}_M$ into a lower-dimensional space. Finally, the KACM module integrates the knowledge stored in $\widetilde{\mathscr{G}}_M$ together with the input meme for hateful classification.

4.4.1 Knowledge Modeling and Organization (KMOM)

This module encompasses a two-stage process aimed at creating the meme's *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$. In the initial stage, we establish the *meme graph* $\mathscr{G}_M$ based on the meme's inherent content. Subsequently, we augment $\mathscr{G}_M$ to produce $\widetilde{\mathscr{G}}_M$ by incorporating supplementary knowledge sourced from ConceptNet [259]. This amalgamation enriches the meme's original graph with external contextual information, fostering a deeper understanding of its semantics and augmenting the effectiveness of hateful meme detection.

Meme Graph The internal concepts encapsulated within a meme emerge from the entities and relationships embedded in its multimodal content. Figure 4.2 depicts the procedure for extracting these concepts and constructing $\mathscr{G}_M$.

Initially, we employ OCR techniques to retrieve the meme's embedded text, along with modern vision-language models [264] to extract the caption. Subsequently, drawing inspiration from prior work that concentrates on constructing knowledge graphs from unstructured text [46, 62, 223], the

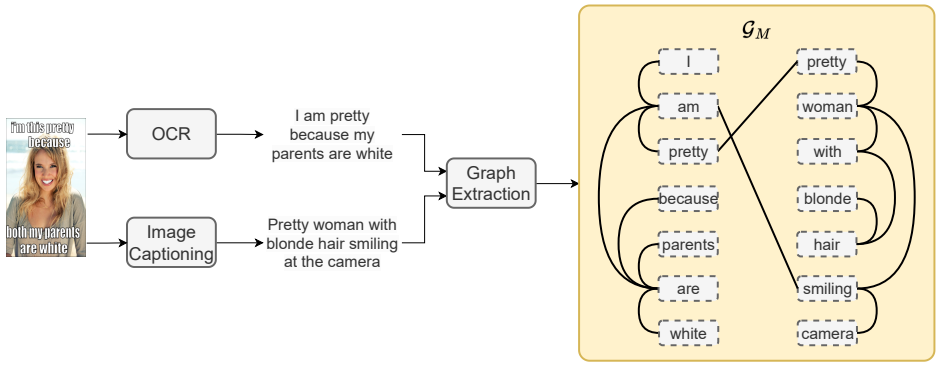


Figure 4.2. The workflow to build the meme graph $\mathscr{G}_M$

Graph Extraction block undertakes Part-Of-Speech (POS) tagging. This process identifies the set of nodes, denoted as $\mathscr{S}_M$, within $\mathscr{G}_M$, originating from both the caption and the embedded text. For instance, in the meme illustrated in Figure 4.2, we extract the nouns such as *woman*, *camera* (from the caption), and *parents*, *pretty* (from the embedded text).

Subsequently, we leverage dependency parsing to extract the set of relationships, designated as $\mathscr{P}_M$, among the aforementioned nodes. More specifically, we utilize the parsing tree to elucidate the relationships between words within the input text, be it the meme's caption or embedded text. This parsing tree delineates the connections based on the syntactic structure of the input sentence, including subject-verb or adjective-noun relationships. Since our primary interest lies in general relationships between words, rather than specific grammatical dependencies, we opt to replace the dependency typologies (such as subject modifier or coordination conjunction) with a more generalized relation, denoted as *relatedTo*. This simplification streamlines the representation of relationships within the *meme graph*, while preserving the overall dependencies.

Lastly, we merge the dependency trees of the caption and embedded text by establishing connections between their root nodes and common words. For example, this involves creating edges linking terms like "pretty" or terms like "am" and "smiling" in Figure 4.2. The outcome of this process yields the final meme graph, denoted as $\mathscr{G}_M$. It is important to note that this graph is not a strict dependency tree; rather, it represents the soft

connections between the caption and the embedded text.

In summary, through the utilization of dependency parsing and the simplification of relationships into a general *relatedTo* category, we conjecture that $\mathscr{G}_M$ offers an effective means of gaining a more comprehensive understanding of the concepts embedded within meme content.

Knowledge enrichment To enrich the information contained in the meme graph $\mathscr{G}_M$, we utilize ConceptNet [259] as an external knowledge base to construct the meme's *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$.

Our choice of ConceptNet over other factual knowledge bases, such as WikiData [289], is motivated by the belief that ConceptNet's repository of common-sense knowledge may be more adept at capturing the true meaning conveyed by memes. For instance, while WikiData may factually relate the term "black" to "color", ConceptNet's inclusion of common-sense knowledge recognizes that "black" can also be perceived as an offensive term linked to words like "negro" and "racist". This nuanced understanding proves invaluable when analyzing memes, given that their interpretation often hinges on context and connotations.

Figure 4.3 provides an illustration of a subset of knowledge extracted from the embedded text of the meme shown in Figure 4.2. It also underscores the recursive nature of the knowledge retrieval process. For example, our query to ConceptNet regarding the term *pretty* reveals its associations with terms like *putty* and *dolly*, synonymy with *lovely*, and an attribute of being *flower*. Furthermore, the term *dolly* triggers additional information retrieval (i.e., *truck* and *frivolous*). The depth of recursion, denoted as the *recursion depth*, is a parameter governing the knowledge extraction process, determining the number of nested ConceptNet queries performed. A higher *recursion depth* entails a greater infusion of external knowledge into the final $\widetilde{\mathscr{G}}_M$.

Algorithm 1 formalises the process of knowledge enrichment. It takes the initial meme graph $\mathscr{G}_M$ and the recursion depth l as inputs. The algorithm initializes $\widetilde{\mathscr{G}}_M$ to mirror $\mathscr{G}_M$ (line 2). For each node s_i in $\mathscr{G}_M$ with a Part-Of-Speech (POS) tag of noun, verb, or adjective (lines 3-4), ConceptNet is queried to retrieve a set of facts related to s_i (line 5). The retrieved nodes and edges, represented as $k_{entities}$ and $k_{relationships}$, respectively, are then iteratively incorporated into $\widetilde{\mathscr{G}}_M$ (lines 6-7). For

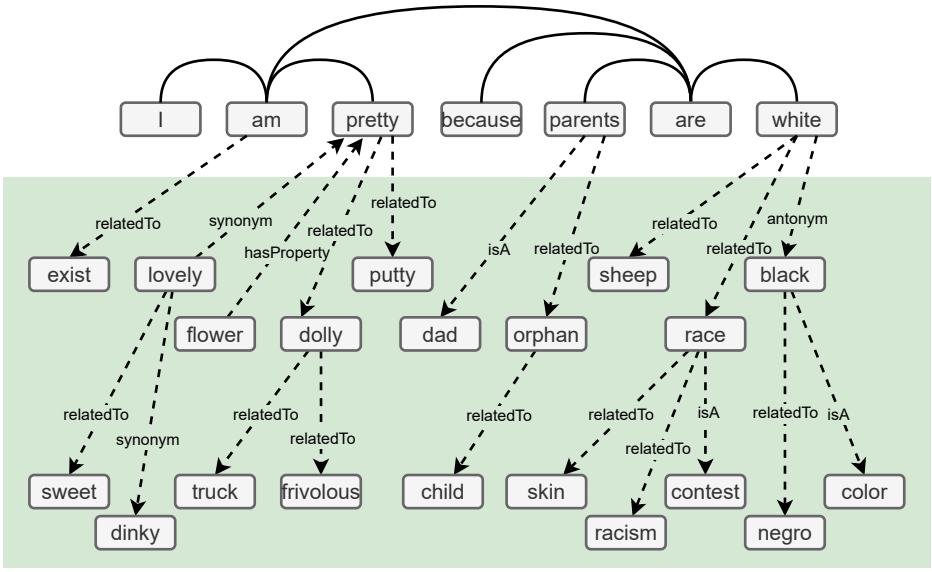


Figure 4.3. Knowledge enrichment: common-sense knowledge retrieved from ConceptNet related to the embedded text of the meme in Figure 4.2.

example, the green region in Figure 4.3 contains a subset of nodes and edges that enrich the caption of the meme in Figure 4.2.

Furthermore, the *GetConceptNet* function, as detailed in Algorithm 2, embodies the recursive querying mechanism. It accepts a general entity e to be queried and the recursion depth l. For instance, Figure 4.3 illustrates knowledge retrieved from the embedded text of the meme in Figure 4.2 at two levels of recursion ($l = 2$). The algorithm initializes the *entities* $\mathscr{S}_K$ and *relationships* $\mathscr{E}_K$ as empty sets (lines 2-3). If the recursion depth l has not been reached (lines 4-6), a set of *triples* related to e is retrieved (line 7). Each triple consists of a subject s and an object o, connected by the predicate p (e.g., the triple *white, relatedTo, race* in Figure 4.3). The algorithm then recursively constructs sub-graphs for s and o within each triple (lines 8-16). The (sub-)entities $sub\_\mathscr{S}_K$ and (sub-)relationships $sub\_\mathscr{E}_K$ returned from each recursive call are incorporated into the $\mathscr{S}_K$ and $\mathscr{E}_K$ sets, respectively. It is important to note that the meaning of the retrieved predicates encompasses a wide range of relationships (e.g., *relatedTo*, *isA*, *isAntonym*), contingent on ConceptNet's design.

Algorithm 1 $\widetilde{\mathscr{G}}_M$ Generation

procedure KnowledgeEnrichment($\mathscr{G}_M$, depth l)
  initialize $\widetilde{\mathscr{G}}_M = \mathscr{G}_M$
  for each node $s_i \in \mathscr{G}_M$ **do**
    if $pos\_tag(s_i) \in \{noun, adj, verb\}$ **then**
      $k\_entities$, $k\_relationships \leftarrow GetConceptNet(s_i, l)$
      add $k\_entities$ to $\widetilde{\mathscr{G}}_M$
      add $k\_relationships$ to $\widetilde{\mathscr{G}}_M$
  return $\widetilde{\mathscr{G}}_M$

Algorithm 2 Querying ConceptNet

procedure GetConceptNet(entity e, depth l)
  initialize $\mathscr{S}_K$ as empty entities' set
  initialize $\mathscr{E}_K$ as empty relationships' set
  if $l = 0$ **then**
    return $\mathscr{S}_K$, $\mathscr{E}_K$
  $triples \leftarrow query\_conceptnet\_api(e)$
  for each triple $= (s, p, o) \in triples$ **do**
    $\mathscr{S}_K \leftarrow o \cup \mathscr{S}_K$
    $\mathscr{S}_K \leftarrow s \cup \mathscr{S}_K$
    $\mathscr{E}_K \leftarrow (s, p, o) \cup \mathscr{E}_K$
    if $l > 1$ **then**
      $sub\_\mathscr{S}_K, sub\_\mathscr{E}_K \leftarrow$ GetConceptNet(o, $l - 1$)
      $sub\_\mathscr{S}_K, sub\_\mathscr{E}_K \leftarrow$ GetConceptNet(s, $l - 1$)
      $\mathscr{S}_K \leftarrow sub\_\mathscr{S}_K \cup \mathscr{S}_K$
      $\mathscr{E}_K \leftarrow sub\_\mathscr{E}_K \cup \mathscr{E}_K$
  return $\mathscr{S}_K$, $\mathscr{E}_K$

Observing Figure 4.3, we notice the inclusion of useful entities such as *dolly*, *racism*, and *negro*, which can aid in classifying the meme as hateful or not. However, the presence of polysemic words like *race* and *dolly* leads to unrelated terms such as *truck* and *contest*. Ultimately, the final knowledge-enriched information network $\widetilde{\mathscr{G}}_M$ emerges as a semantic

attributed graph, housing potentially informative concepts for meme comprehension, interconnected by edges that depict their relationships.

4.4.2 Knowledge Embedding Representation (KERM)

After obtaining the meme's *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$, the next objective is to represent this heterogeneous knowledge using graph embedding techniques. First, we consider two distinct node types within the graph: (i) Nodes derived from the meme's caption and embedded text, which are inherently part of the original meme graph; (ii) Nodes acquired from ConceptNet during the enrichment process. Then, our approach involves node embedding, where we aim to generate low-dimensional feature vectors for each node in $\widetilde{\mathscr{G}}_M$. These embeddings should encapsulate both the node's semantic meaning and the structural context provided by its neighboring nodes. To achieve this, we employ a two-step process: we begin by embedding the words associated with each node using BERT [52]. This step helps capture the semantic information of the node. Subsequently, we leverage an unsupervised random walk-based algorithm called HIN2Vec [70] to account for the local connectivity patterns within the meme graph and the various types of relationships that exist among nodes. This ensures that the structural context is considered when generating the embeddings.

It is important to note that our choice to perform node embedding, rather than embedding the entire graph into a single low-dimensional vector, is contingent upon the heterogeneity of knowledge within $\widetilde{\mathscr{G}}_M$. When embedding the entire graph as a whole, we risk losing the ability to distinguish between valuable and extraneous information. In contrast, by obtaining individual embeddings for each node, we retain the granularity required to identify useful information in the subsequent stage of our framework, namely the *Knowledge-Augmented Classification Module*.

4.4.3 Knowledge-Augmented Classification (KACM)

Our focus now turns to the integration of the knowledge-enriched information network into the classification process. The architecture of the KACM, illustrated in Figure 4.4, draws inspiration from memory-augmented neural networks [298, 265]. The KACM is designed to harness

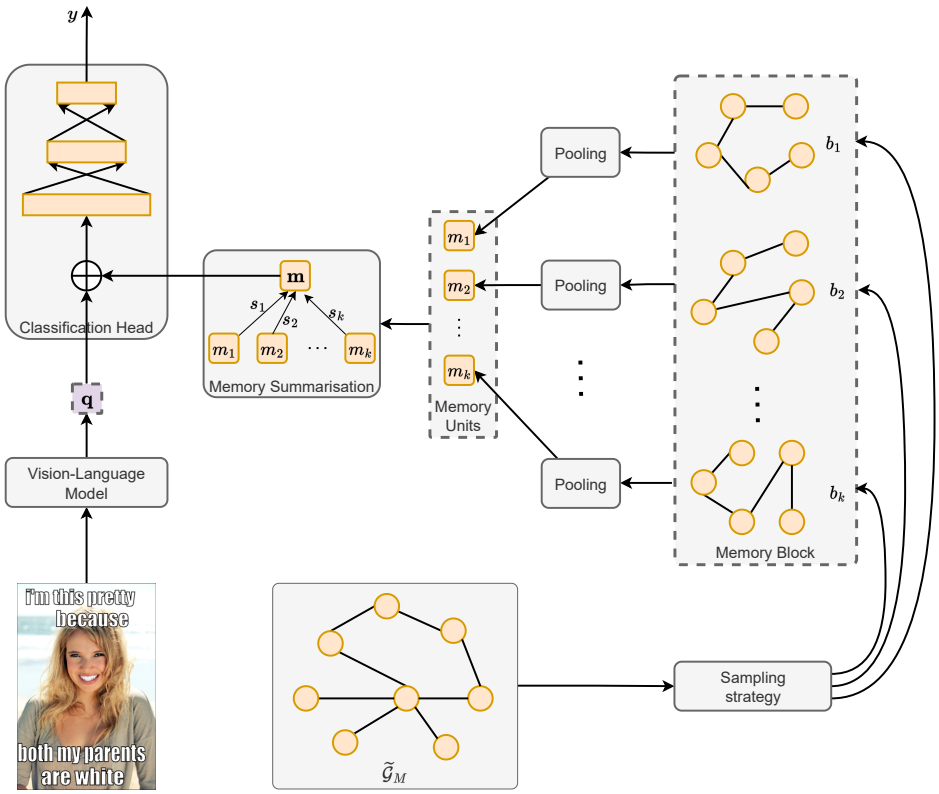


Figure 4.4. The architecture of the KACM module: the *vision-language model* encodes the input meme into a vector representation **q**. The *sampling* and *pooling* components provide the memory buckets b_i and their vector representations m_i, respectively. Next, the *memory summarisation* component dynamically learns the most informative knowledge context **m** for the hateful classification. Finally, the *classification head* performs the hateful classification based on the merged representations of the meme and summarised memory.

the wealth of knowledge stored in $\widetilde{\mathscr{G}}_M$ by incorporating it into a *memory block* accessible during the classification process. This memory component consists of several slots, referred to as *buckets*, each of which stores a subset of the knowledge contained in $\widetilde{\mathscr{G}}_M$. The workflow of the system, considering the input meme and the memory component, unfolds as follows:

(i) The Vision-Language model encodes the input meme into a vector representation $q \in \mathscr{R}^N$.

(ii) The pooling module independently encodes each bucket b_i into vector representations $m_i \in \mathscr{R}^M$.

(iii) The memory summarization module dynamically learns the most relevant context for the hateful classification task and fuses the memory units into a final vector representation.

(iv) The classification head performs the ultimate hateful classification based on the merged representations of the meme q and the (summarized) memory $\mathbf{m}$.

The following paragraphs provide a detailed explanation of each component within the KACM.

Vision-Language Model The integration of multiple data modalities, such as text and image, remains a challenging problem in machine learning research. This challenge has spurred the development of novel tasks in the realm of multimodal learning, including visual question answering (VQA) [260] and topic learning [139]. In response to these demands, researchers have designed advanced visual-language models capable of capturing intricate relationships between text and images [137, 112, 150].

In our approach, we leverage the Visual Transformer (ViT) architecture, which facilitates the simultaneous encoding of both image and text components of a meme into a shared representation. Specifically, we employ the ConcatBERT model [254], which harnesses ResNet-152 to extract image features and BERT to extract text features. To capture the interactions between these different modalities and learn a cohesive joint representation, we incorporate a cross-modal attention layer [137]. Importantly, our choice of model is motivated by the promising results similar

architectures have demonstrated in various disinformation mining tasks, including the detection of fake news in news articles [294] and social media posts [106].

While we acknowledge that newer methods for integrating text and image modalities have emerged, the primary contribution of our work lies in how we incorporate external knowledge into the classification process, rather than solely relying on the internal content of memes. Furthermore, the selection of the Visual-Language model is a parameter within the KACM architecture, and we plan to assess its impact through experiments.

Sampling strategy As previously mentioned, the knowledge network $\widetilde{\mathscr{G}}_M$ can be rich in information. To manage this wealth of information, we introduce the concept of a "bucket" as the fundamental unit of knowledge that may or may not be pertinent to the hateful classification task. We implement this concept by employing a random walk approach on $\widetilde{\mathscr{G}}_M$, which generates a sequence of nodes forming the bucket. This sequence is denoted as $b_i = \{s_1, s_2, \cdots, s_l | s_i \in \widetilde{\mathscr{G}}_M\}$, where l represents the length of the random walk. In this context, we opt for the standard random walk with restart (RWR) algorithm [276], which has demonstrated its effectiveness in various classification tasks [85, 324, 323].

Given the potentially extensive amount of information to be processed, we incorporate a *sampling strategy* that selects a predefined number of k buckets, or random walks, from $\widetilde{\mathscr{G}}_M$. These selected buckets serve as memory slots during the subsequent classification process. This strategy allows us to focus on a manageable subset of the available knowledge, reducing computational complexity while maintaining the potential to capture valuable information.

Pooling The *pooling component* assumes a pivotal role in the creation of a memory unit, serving as a concise representation of a knowledge *bucket*. Specifically, we consider the embeddings of the nodes within the bucket and apply the mean pooling operation to consolidate them into a single vector. This operation yields the memory unit $m_i \in \mathscr{R}^n$, which encapsulates the essential knowledge element or bucket and is subsequently integrated into the classification process. We posit that the mean pooling operation pre-

serves a significant amount of information, as the bucket size is relatively modest compared to the overall scale of $\widetilde{\mathscr{G}_M}$.

Memory Summarisation The memory units encompass distinct buckets, each carrying varying degrees of significance in the classification process. As illustrated in Figure 4.3, the acquired knowledge may also include superfluous and noisy information that needs to be sifted through before making a final determination regarding the meme's hatefulness. However, we lack prior knowledge regarding which memory unit(s) should be considered when processing a given input meme. Consequently, we employ an attention mechanism to dynamically discern the most relevant context, namely the most pertinent units, for the purpose of hateful classification.

Formally, we consider the group of memory units $\{m_1, m_2, \cdots m_k | m_i \in \mathscr{R}^n\}$ and leverage the soft attention mechanism proposed in [311]:

$$u_{it} = tanh(W_w m_{it} + b_w) \tag{4.1}$$

$$\alpha_{it} = \frac{\exp(u_{it} u_w)}{\sum_{k=1}^{T_x} \exp(u_{ik} u_w)} \tag{4.2}$$

$$s_i = \sum_t \alpha_{it} m_{it} \tag{4.3}$$

That is, the i-th memory unit m_{it} is projected to the vector u_{it} through a one-layer MLP. Subsequently, α_{it} is computed representing the normalized similarity between u_{it} and the (jointly learnt) context vector u_w. Next, the i-th attention score s_i is computed with the dot product of α_i and m_i and represents the contextual importance of the i-th memory unit. Finally, the memory summary $\mathbf{m} \in \mathscr{R}^n$ is obtained as a linear combination of the attention scores and the memory units:

$$\mathbf{m} = \sum_{i=1}^{k} s_i m_i \tag{4.4}$$

Notably, the MLP weights W_w, b_w and the context vector u_w are dynamically learnt during the training process, allowing the system to automatically recognize which knowledge buckets are informative for the hate-

ful classification. In essence, this dynamic learning capability empowers the system to filter out irrelevant or unhelpful information stored within $\widetilde{\mathcal{G}}_M$.

Classification Head The pivotal role of the classification head lies in accomplishing the definitive hateful classification task. Specifically, this process commences by concatenating the embeddings of the input meme denoted as $\mathbf{q}$ and the memory summary $\mathbf{m}$. Subsequently, this concatenated feature representation traverses through a succession of fully connected layers, ultimately culminating in the application of a final softmax layer. These fully connected layers serve the crucial purpose of nonlinearly mapping the input features into a higher-dimensional space. This transformation empowers the model to discern intricate decision boundaries, thereby enhancing its capacity to tackle complex classification tasks. Ultimately, the final softmax layer plays the role of normalizing the output probabilities across all potential classes, thereby producing the anticipated probability distribution concerning the potential classes, such as hateful or non-hateful, especially in binary classification settings.

4.5 Experimental Evaluation

4.5.1 Datasets & Metrics

We conduct our experiments using two widely recognized benchmark datasets that are commonly utilized in contemporary literature: the Facebook Hateful Meme dataset [114] and the Multimedia Automatic Misogyny Identification (MAMI) dataset [68].

The Facebook Hateful Meme dataset [114] was employed in the homonym challenge, marking its pioneering effort in the realm of hateful meme detection. Notably, this dataset was designed to detect hate speech directed towards protected categories, including race and gender. The dataset's annotations were meticulously carried out based on specific guidelines, ensuring both consistency and precision in labeling, with annotations curated by human annotators. The task framed by this dataset is binary classification, and our focus primarily centered on the dataset subset utilized during the challenge's second phase [113]. This subset comprises

8,500, 540, and 2,000 memes for training, validation, and testing, respectively. In total, this subset consists of 7,071 non-hateful memes out of 11,040 and 3,969 hateful memes out of 11,040. To align with the official challenge protocol, we adopted the Area Under the ROC Curve (AUC) as our chosen metric for evaluating classification performance.

The MAMI dataset [68] was employed in task-5 of the SemEval 2022 conference and was explicitly designed to address misogynistic content present on social media platforms. This dataset comprises 11,000 memes collected from Twitter and Reddit, with annotations generated through crowdsourcing platforms using two distinct annotation schemes. The first scheme involves binary classification, distinguishing between misogynous and non-misogynous memes. The second scheme employs a 5-class approach to discern various types of misogyny, including shaming, stereotypes, objectification, and violence, in addition to general misogyny. However, the multi-class scheme is somewhat constrained by the subjectivity observed in annotations, indicated by low inter-annotator agreement and the multi-label nature of the data [68]. Therefore, we opt to focus on the binary classification scheme within the dataset, which features a more balanced distribution with 5,500 memes per class. Finally, our choice of metric for assessing classification performance in this case is the macro average F1-score, consistent with the SemEval task's evaluation protocol.

4.5.2 Experimental Protocol

Our experimental protocol encompasses the following key objectives:

- *Performance Evaluation*: In this phase, we aim to comprehensively assess KERMIT's performance by benchmarking it against various multimodal baseline models.
- *Knowledge Contribution*: This experiment takes a deep dive into the impact of integrating external knowledge into the classification process, shedding light on how such knowledge enhances or influences the model's performance.
- *Ablation Study*: We conduct a systematic evaluation to discern the individual efficacy of each component within KERMIT's architectural framework.

In the context of the first experiment, our primary aim is to maximize performance on both datasets. To achieve this, we consider the inherent task hierarchy within the datasets. Specifically, while a misogynistic meme is inherently offensive, not all offensive memes necessarily exhibit misogyny. Therefore, when evaluating performance on the Hateful Meme dataset, we strategically choose to augment the training set of the Hateful Meme dataset with (training) data from the MAMI dataset.

On the other hand, our focus shifts towards a meticulous examination of the contributions of various knowledge types and an in-depth analysis of different components within KERMIT. To ensure that any observed differences in performance are primarily attributed to architectural changes or knowledge incorporation, rather than the collective effect of modifications and an expanded training dataset, we maintain a clear separation between the training sets.

Experimental Setup

All experiments have been conducted on Google Colab, utilizing a computational setup equipped with a single core hyper-threaded Xeon Processor running at 2.2 GHz, 12 GB of RAM, and a Tesla K80 GPU.

To construct the meme's knowledge-enriched information network, $\widetilde{\mathscr{G}}_M$, we harness a suite of tools and libraries. For image captioning, we rely on the "ydshieh/vit-gpt2-coco-en" model available within the HuggingFace library[3]. Extracting text from the meme's image is accomplished using the easyOCR library[4]. Subsequently, both the embedded text and the caption undergo preprocessing steps to eliminate punctuation and stopwords. The Spacy dependency parser[5] is employed to extract the dependency tree structures from the meme's caption and embedded text. Lastly, for retrieving pertinent information from ConceptNet, we utilize the official Python package[6].

Moving to the KERM module, we employ the "bert-base-uncased" model available through the HuggingFace library for text embedding. In

[3] https://huggingface.co/ydshieh/vit-gpt2-coco-en
[4] https://pypi.org/project/easyocr/
[5] https://spacy.io/api/dependencyparser
[6] https://pypi.org/project/ConceptNet/

the case of the meme's knowledge-enriched information network, we utilize the HIN2Vec implementation as proposed by [306] for network embedding.

Finally, regarding theKACM module, we adopt the ConcatBERT model from the MMF library[7] as the foundational vision-language model. Subsequently, the classification head is implemented using the PyTorch library, configured as a feed-forward neural network. This network consists of four linear layers, one dropout layer, and employed the softmax activation function for the final binary classification task.

4.5.3 Results

Performance Evaluation

We have chosen several multimodal baselines, categorizing them into two groups: (i) multimodal models pretrained on a single modality, i.e., VilBERT [151], VisualBERT [136] and SEER [83]; (ii) multimodal models pretrained on a multimodal objective [126], i.e., ViLT [116], VisualBERT COCO, OSCAR [138], CLIP [210], Ernie-Vil [312]. Notably, the latter model explicitly uses some knowledge to improve the learning process, even if that knowledge is internal to the content rather than external.

It is worth to note that we have deliberately omitted ensemble learning approaches, which have demonstrated state-of-the-art performance on the evaluated datasets.

Table 4.1 presents a detailed comparison between KERMIT and the aforementioned baseline models. The results reveal that models leveraging multimodal pre-training generally outperform those relying on unimodal pre-training. This underscores the significance of considering the intricate interactions between text and images when comprehending harmful memes. These findings align with prior research [292] in multimodal machine learning, emphasizing the need to account for cross-modal interactions to enhance performance in downstream tasks.

Remarkably, KERMIT and Ernie-ViL emerge as the top-performing models. This suggests that the incorporation of external knowledge can indeed yield improvements in classification outcomes. However, KERMIT exhibits a notable advantage, surpassing Ernie-ViL by 4.3% on the Hateful Meme dataset and 4.9% on the MAMI dataset. These results under-

[7] https://github.com/facebookresearch/mmf

Table 4.1. Comparison with state-of-the-art baselines

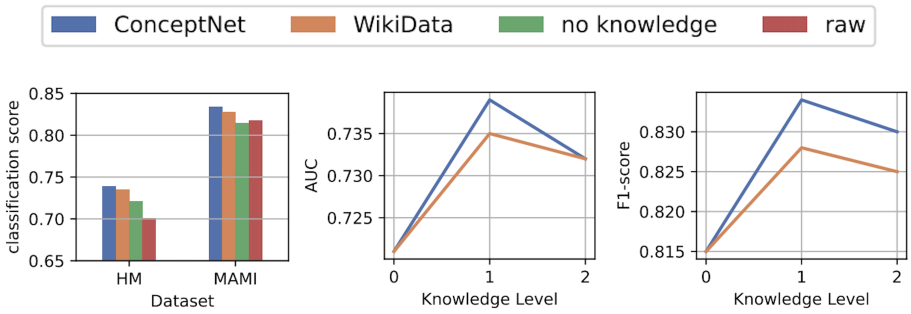
Category	Model	Dataset	
		Hateful Meme	MAMI
Multimodal (Unimodal pre-training)	VilBERT [151]	0.734	0.725
	VisualBERT [136]	0.732	0.723
	SEER [83]	0.708	0.718
Multimodal (Multimodal pre-training)	ViLT [116]	0.725	0.744
	VisualBERT COCO [136]	0.752	0.742
	OSCAR [138]	0.793	0.684
	CLIP [210]	0.803	0.765
	Ernie-Vil [312]	<u>0.816</u>	<u>0.793</u>
	KERMIT (Ours)	**0.853**	**0.834**
	Δ_{KERMIT} (%)	4.3%	4.9%

score the substantial benefits of explicitly modeling and integrating external knowledge into the classification process.

Overall, our experimental findings underscore KERMIT's promise as an effective approach for classifying harmful memes, irrespective of the specific typology of harm, such as racism, hate, or misogyny. Specifically, the incorporation of external knowledge into the model architecture enables KERMIT to outperform several models on both benchmark datasets, highlighting that incorporating additional knowledge could be a valuable direction for future research in multimodal disinformation detection.

Knowledge Contribution

The central hypothesis of this study revolves around the idea that comprehending harmful memes necessitates a backdrop of background knowledge that encapsulates the meme's context, including cultural references and societal issues. Therefore, our first investigative step is to assess the impact of integrating such knowledge into the classification process. To this end, we explore four distinct scenarios: (i) *no knowledge*, i.e., this scenario represents the absence of any external knowledge. Here, the evaluation solely involves the vision-language model and the classification head within the KACM module; (ii) *raw knowledge*, i.e., in this scenario, raw knowledge is incorporated, where the meme's caption and embedded text



(a) Knowledge Structure **(b)** Hateful Meme Dataset (HM) **(c)** Misogyny Dataset (MAMI)

Figure 4.5. The contribution of external knowledge: (a) comparison between structured knowledge from ConceptNet, structured knoweldge from WikiData, unstructured knowledge and no knowledge; (b), (c) performance by varying the recursion depth on Hateful Meme and MAMI datsets, respectively.

are amalgamated with the unprocessed terms extracted from ConceptNet. This combined data is then fed to both the vision-language model and the classification head; (iii) *ConceptNet*, i.e., this scenario leverages ConceptNet [259] to construct the meme's *knowledge-enriched information network* ($\widetilde{\mathscr{G}}_M$), as detailed in Section 4.4.1; (iv) *Wikidata*, i.e., similar to the previous scenario, this setting utilizes WikiData [289] instead of ConceptNet to build $\widetilde{\mathscr{G}}_M$.

The histogram in Figure 4.5(a) illustrates that classification performance notably improves under both the ConceptNet and WikiData settings when compared to the *no knowledge* scenario. This result underscores our hypothesis that the incorporation of external knowledge enhances the classification process, regardless of the specific characteristics of the knowledge source or the entities it encompasses. Furthermore, it is worth highlighting that the ConceptNet setting exhibits slightly superior performance relative to the WikiData setting, suggesting that the common-sense knowledge embedded in ConceptNet may be more beneficial than the purely factual knowledge contained in WikiData.

However, we also observe that incorporating *raw knowledge* into the classification process yields counterproductive results, leading to a decline in performance. This outcome empirically validates the efficacy of the

knowledge modeling strategy described in Section 4.4.1. In essence, merely adding unstructured knowledge does not necessarily enhance performance. Instead, structuring the knowledge in a manner that facilitates the model's extraction of relevant background information for identifying harmful signals in memes is crucial.

To delve deeper into the advantages of incorporating external knowledge, we explore the "quantity" of knowledge required to enhance classification performance. Specifically, we investigate the recursive nature of our knowledge-enrichment strategy by varying the depth of recursion l when querying ConceptNet or WikiData.

Figures 4.5(b) and (c) depict the changes in classification performance for the Hateful Meme and Misogyny datasets, respectively. Interestingly, we find that the optimal classification performance is consistently achieved at a recursion depth of one, regardless of the dataset or knowledge base used. This outcome is likely due to the knowledge hierarchy, where the relevance of knowledge to the meme diminishes as the recursion depth increases. In essence, there is an increased risk of introducing irrelevant or noisy information into $\widetilde{\mathscr{G}}_M$ with higher recursion depths. Additionally, we emphasize that increasing the recursion depth significantly raises the computational challenge, as the number of API queries to ConceptNet or WikiData grows exponentially, resulting in unfeasible construction times for $\widetilde{\mathscr{G}}_M$.

Overall, our findings underscore the advantages of incorporating external knowledge into the classification process of harmful memes. Furthermore, the results highlight that KERMIT's "knowledge-enriched information network" effectively captures essential information about the input meme, demonstrating its potential in enhancing classification outcomes.

Ablation Study

After establishing the advantages of integrating external knowledge, our focus shifts towards evaluating various components within the KERMIT framework, i.e., the vision-language model in the KACM module, the node embedding algorithm in the KERM module, the knowledge injection strategy in the KACM module.

Table 4.2. Ablation study: performance by varying the vision-language model in the KACM module, the node embedding algorithm in the KERM module, the knowledge injection strategy in the KACM module.

Module	Parameter	Dataset	
		Hateful Meme	MAMI
Visual-Language Model	ConcatBERT	0.721	0.815
	+ $\widetilde{\mathscr{G}}_M$	**0.739**	**0.834**
	MMBT	0.718	0.815
	+ $\widetilde{\mathscr{G}}_M$	**0.729**	**0.817**
KERM	FeatherNode	0.708	0.806
	Hin2Vec	**0.739**	**0.834**
KACM	Concatenation	0.725	0.825
	Attention	**0.739**	**0.834**

Vision-language model We investigate the vision-language model integrated within the KACM module, recognizing its pivotal role in representing the meme's internal information and capturing the interplay between its visual and textual elements. In our pursuit to gauge the effectiveness of different vision-language models, we conducted a series of experiments employing ConcatBERT [254] and MMBT [112]. The results of these experiments, as presented in Table 4.2, reveal that both models exhibit relatively similar performance, with ConcatBERT demonstrating a slight edge in performance, particularly on the Hateful Meme dataset.

However, a striking revelation emerges from our analysis: the integration of the meme's *knowledge-enriched information network* $\widetilde{\mathscr{G}}_M$ confers a substantial boost to the classification performance for both ConcatBERT and MMBT. This finding underscores that $\widetilde{\mathscr{G}}_M$ encapsulates additional contextual information that extends beyond the pre-trained knowledge ingrained in any vision-language model.

This revelation accentuates the inherent synergy between these two components and emphasizes the importance of their collaborative integration within the KERMIT framework.

Graph embedding We investigate the impact of the node embedding algorithm used in the KERM module. Specifically, we compare the performance of HIN2Vec [70] and FeatherNode [224] algorithms, as shown in Table 4.2.

Our results unveil that HIN2Vec outperforms FeatherNode, potentially owing to its superior representational capacity. This distinction arises from the contrasting approaches adopted by these algorithms. FeatherNode treats $\widetilde{\mathscr{G}}_M$ as a homogeneous graph and overlooks the nuanced semantics embedded within relationships between nodes. Conversely, HIN2Vec is expressly designed for heterogeneous networks, capitalizing on the multifaceted nature of relationships among nodes. Consequently, HIN2Vec's embeddings account for critical factors, including node types (i.e., whether a node represents a meme's entity or a concept retrieved from ConceptNet) and relationship types, thereby facilitating a more comprehensive understanding of the knowledge-enriched network's structural intricacies.

Knowledge injection Our investigation delves into the evaluation of two pivotal design choices embedded within the KERMIT model: the configuration of knowledge buckets denoted as b_i, and the memory summarization strategy.

To scrutinize the former parameter, we meticulously examine the number of knowledge buckets and their individual sizes. In precise terms, we control these parameters via the manipulation of the number of random walks and the length of each random walk, respectively. The results of these experiments, presented in Figure 4.6, reveal that these parameters wield a relatively modest influence on the model's performance. Intriguingly, optimal outcomes for both datasets materialize when employing a configuration of 15 knowledge buckets. Furthermore, we observe a decreasing trend for the length of random walks. This result is possibly attributed to the system placing too much importance on a few highly connected nodes, such as those belonging to the meme's embedded text and caption. As increasing the length of random walks also increases computational costs, we choose to use 11 nodes for each random walk, which results in the best classification performance.

Turning our attention to the memory summarization strategy within the KACM module, we conduct a comparative analysis between the at-

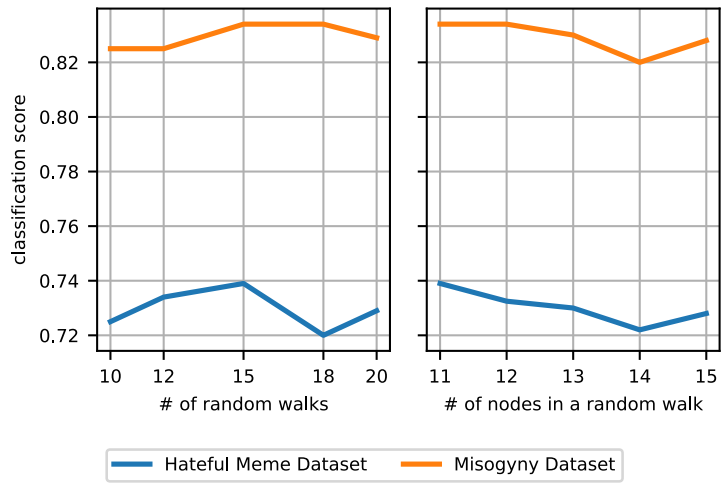


Figure 4.6. Ablation study: performance by varying the size of the knowledge buckets.

tention mechanism outlined in Section 4.4.3 and a more straightforward approach based on averaging the embeddings of each memory unit. The results, as presented in Table 4.2, unequivocally endorse the superiority of the attention-based method. This finding underscores our initial hypothesis that memory units do not possess uniform significance in the context of hateful classification. It further highlights the model's capacity to autonomously discern the most informative buckets without any external guidance or supervision.

4.6 Conclusions and Future Works

In this chapter, we introduced KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), a novel framework meticulously crafted for the purpose of knowledge-informed harmful meme detection. The fundamental essence of our approach revolves around the creation of a knowledge-enriched information network tailored explicitly for memes. This network harmoniously interweaves the intrinsic elements of a meme with pertinent external knowledge sourced from ConceptNet. No-

tably, KERMIT harnesses the power of dynamic learning, underpinned by memory-augmented neural networks and an attention mechanism, to astutely discern and exploit the most salient knowledge facets. These, in turn, enable precise and accurate classification of harmful memes.

Empirical validation of our proposed framework was meticulously carried out through a series of experiments conducted on two benchmark datasets: the Hateful Meme dataset and the Multimedia Automatic Misogyny Identification dataset. The results underscored KERMIT's prowess in sourcing and adeptly deploying contextual knowledge to bolster predictive performance. Moreover, when pitted against an array of multimodal baseline models, KERMIT consistently emerged as the torchbearer, achieving superior classification performance.

From a broader perspective, our study casts a spotlight on the profound efficacy of amalgamating external knowledge into the classification process, opening up a compelling vista for further exploration within the domain of harmful meme detection. These findings underscore the pivotal role played by artificial intelligence and knowledge discovery in the ongoing evolution of content moderation.

Our research lays a foundation for several promising avenues of future investigation. First, we envisage a foray into the realm of multimodal knowledge bases, notably the utilization of visual knowledge bases such as VisualGenome [123]. This extension would enable us to harness not only textual but also visual knowledge, potentially augmenting the discriminatory capabilities of KERMIT in identifying harmful memes. Second, we aspire to leverage the meme's *knowledge-enriched information network* as a vehicle for interpretability. In this context, we anticipate that the extracted knowledge can provide crucial insights into the rationale behind the classification of a particular meme as harmful or benign. Such interpretability could inject greater transparency into the classification process, aligning with the imperatives of content moderation. Finally, we remain acutely mindful of the ethical dimensions surrounding the use of KERMIT. Specifically, we are committed to exploring the ethical implications of our framework, particularly in regard to the potential amplification of biases or prejudices that may arise from the infusion of external knowledge sources.

Chapter 5

Knowledge Transfers across Disinformation Detection

5.1 Research Context and Contributions

In recent years, the rampant spread of misinformation and fake news across the internet and social media platforms carries profound societal, economic, and political repercussions. These consequences include election interference [16], polarization [290], and public health crises [41]. The origins of fake news, however, cannot be attributed solely to individuals' lack of access to accurate information or knowledge [57]. In the era of "post-truth" [181], psychological and sociological factors play a pivotal role in guiding individuals towards news that resonates with their emotions and beliefs. An illustrative example lies in the examination of Italian public emotions regarding the COVID-19 pandemic. It has been observed that while the initial wave of the pandemic saw a gradual decline in fear, the emotion of anger remained relatively constant but underwent a transformation, giving rise to diverse narratives, including polical countermeasure and vaccine hesitancy [74]. Additionally, the realm of fake news frequently exploits emotional appeals, moral emotions, and sensationalism to seize the attention of the masses [286] and expand its reach to a broader audience [200, 124].

Consequently, contemporary computational approaches [329, 240] aimed at disinformation detection are adopting a holistic approach by rec-

ognizing the interplay between various aspects of disinformation. For instance, understanding the stances expressed towards a news article can aid in debunking a rumor, while verifying the veracity of the news can inform the assessment of stances in associated posts [158]. Similarly, different topics exhibit distinct credibility distributions, suggesting that knowledge of the news topic could enhance predictions of its authenticity [140]. To operationalize these intuitions, Multi-Task Learning (MTL) has emerged as a promising solution. By simultaneously training on multiple tasks, such as stance detection and rumor verification [117] or emotion recognition and fake news detection [124], these approaches aim to learn a shared, more informative representation that enhances overall performance.

However, the process of information exchange between tasks is not always straightforward, and the application of MTL strategies can introduce features that lead to misguided predictions [303]. In other words, the incorporation of multiple tasks does not guarantee performance improvement (*positive transfer*) but may result in performance degradation (*negative transfer*) due to feature interference and limited model capacity [134].

In the domain of disinformation detection, mitigating the risk of *negative transfer* between tasks is addressed through a model-centric approach. This approach involves designing intricate (multi-task) architectures that facilitate the model's learning while minimizing the influence of irrelevant shared information [302, 280, 140]. A limited set of disinformation-related tasks, such as stance detection and rumor verification [302] or topic detection and fake news detection [140], is selected for analysis. Classification performance is subsequently evaluated in both single-task and multi-task settings to gain insights into the presence of positive and negative transfers. However, the underlying mechanisms governing the beneficial or detrimental information sharing in these contexts remain unclear.

Given these premises, we undertake a direct investigation into the phenomena of positive and negative transfers, aiming to uncover the distinctions and commonalities between models trained in single-task and multi-task scenarios. Our focus centers on a comprehensive array of disinformation-related tasks: Sentiment Analysis (SA), Fake News Detection (FND), Stance Detection (SD), and Topic Detection (TD). We formulate the following research questions (RQs):

RQ1 What advantages does multi-task learning offer compared to single-

task learning?

RQ2 How does the knowledge acquired by multi-task models differ from that of their single-task counterparts? Do the explanations provided by multi-task models reveal these differences?

To address these questions, we propose a versatile multi-task framework, built upon the architectural foundations of MT-DNN [144]. This adaptable framework stands capable of accommodating a variable number of disinformation-related tasks, thereby enabling a systematic exploration of the presence of positive and negative transfers across every combination of the aforementioned disinformation-related tasks. Additionally, our proposed system harnesses feature importance explanations proffered by both single-task and multi-task models. These explanations serve a dual purpose: they unveil shared knowledge within the models and, perhaps more critically, cast a revealing light upon the intricate dynamics governing positive and negative transfers. This study represents the first known attempt to investigate MTL within the context of multiple disinformation-related tasks and to explore the dynamics of positive and negative transfers through the lens of models' explanations, thus extending the examination beyond predictive performance metrics alone.

Our findings bring to light instances of *positive transfer* in several task combinations. Notably, we observe that the proposed system leads to substantial performance enhancements in SA, FND, and TD tasks, with improvements of up to 3.26%, 6.57%, and 0.62%, respectively, when comparing the performance of single-task models to their optimal multi-task counterparts. However, it is noteworthy that we do not identify any statistically significant improvement in the SD task. Furthermore, the occurrence of positive or negative transfers cannot be solely attributed to the number of training tasks or the effects of data augmentation, nor is it solely contingent on the similarity between tasks. Upon comparing the explanations provided by single-task and multi-task models, we discern that *positive transfer* does not revolutionize the fundamental knowledge of any model. Instead, it refines what the model could already learn in single-task settings by incorporating additional patterns gleaned from other tasks. Conversely, *negative transfer* significantly diminishes the model's knowledge, to the extent that the explanations furnished by multi-task models

are akin to random perturbations of the explanations generated by their single-task counterparts.

Overall, our research illuminates the intricate interplay among various disinformation-related tasks, offering valuable insights for the design of more effective (multi-task) strategies to combat disinformation within dynamic information landscapes.

5.2 Related Works

5.2.1 Multi-task learning

MTL is a valuable technique used to address data sparsity by capitalizing on shared characteristics among diverse datasets [31]. One variant of MTL, known as the private shared model, combines task-specific and shared representations [45]. Within this framework, the shared component captures common information across tasks, while each task retains its private component for task-specific knowledge. In the realm of disinformation detection, prior work [302, 280, 140] has adopted the private shared model due to the observed positive interactions between disinformation-related tasks, such as stance detection and rumor verification [117], or emotion recognition and fake news detection [124].

Despite its advantages, MTL can encounter performance degradation, referred to as *negative transfer*, when datasets exhibit distinct structures or objectives [322], or when tasks lack sufficient relatedness [22, 21, 164]. Previous research [302, 158] in multi-task disinformation mining has attempted to mitigate this issue by devising complex architectures capable of dynamically learning which features to share. Nevertheless, the underlying mechanisms and determinants of positive and negative transfers remain unclear.

Given these premises, we take a significant step toward elucidating this phenomenon by investigating the contributing factors across a comprehensive array of disinformation-related tasks. Our approach involves a comparative analysis of the knowledge acquired by models trained in both single-task and multi-task settings, thereby shedding light on the tangible benefits of *positive transfer* and the detrimental consequences of *negative transfer*.

5.2.2 Explainable Disinformation Detection

Modern predictive models, such as deep neural networks, excel at learning intricate patterns from extensive datasets but often lack transparency in their decision-making processes [274]. Fortunately, recent advancements in explainable artificial intelligence have opened doors to comprehending the inner workings of these models [227]. One notably potent approach within this domain is the use of *local* explanations, particularly those reliant on *feature importance*, which offer valuable insights into the factors influencing model predictions. These explanations have proven invaluable for domain experts, aiding in the identification of biases embedded within predictive models [24, 43, 146]. In the context of disinformation mining, prior research [2, 120] has routinely integrated these interpretability techniques to decipher the reasoning behind specific predictions. This often involves highlighting critical terms within fabricated news articles [246], emphasizing pivotal elements in the news propagation process [152], or unveiling malicious spreading behaviors [128].

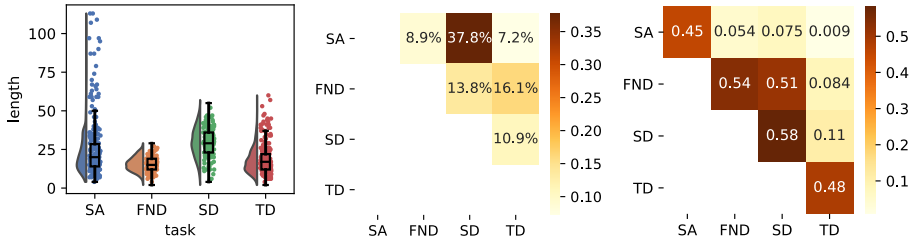
However, the utility of these explanations in the context of multiple disinformation-related tasks remains unexplored. Our work pioneers this domain by leveraging feature importance explanations to elucidate the mechanisms governing the emergence of both positive and negative transfers between these tasks. This research not only advances the field of explainable disinformation analysis but also contributes to a comprehensive understanding of how diverse tasks can collectively enhance efforts to combat disinformation.

5.3 Material and methods

5.3.1 Datasets

We employ four publicly available textual datasets in our study: the EyeMovement Database [170] for the SA task, RumorEval [82] for the FND task, PHEME [335] for the SD task, and LIAR [291] for the TD task. Notably, SA, FND, and TD tasks are conventional text classification tasks, whereas SD involves pairwise text classification to predict a target text's stance in relation to a given statement.

The SA dataset involves binary sentiment classification of 979 textual



(a) sentence length **(b)** vocabulary overlap **(c)** semantic similarity

Figure 5.1. Datasets' statistics

statements, distinguishing between *positive* and *negative* sentiment. In contrast, the FND dataset categorizes 2,537 claims as *false*, *real*, or *unverified*. SD is a pairwise-classification task with 10,658 instances, classifying a target text's stance as *support*, *refute*, *query*, or *comment* on a particular statement. For the TD dataset, we select the top-5 most common topics: *healthcare*, *taxes*, *election*, *immigration*, and *education*, resulting in 1,743 instances. SA and TD tasks exhibit balanced classification problems, while the FND and SD tasks are more unbalanced, with a higher number of true claims and instances supporting the statement.

Figure 5.1 shows additional statistics. In particular, Figure 5.1a illustrates the sentence length distributions for each task. All tasks predominantly feature relatively short texts spanning 20 to 30 words, with the SA dataset containing some outliers with instances exceeding 50 words. In addition, Figure 5.1b displays the vocabulary overlap matrix. Notably, SA and SD exhibit the highest overlap, while other combinations show less than 20% word overlap, likely due to diverse data sources contributing to these datasets.

Finally, we compute semantic similarity between tasks using sentence embeddings generated with SentenceBERT [217]. This method involves calculating cosine similarity scores between each sentence in each dataset and all examples in each other dataset, then averaging the scores to obtain similarity scores. Figure 5.1c shows the semantic similarity matrix. TD demonstrates the most heterogeneous dataset, as evidenced by the low self-similarity score, likely due to encompassing multiple unrelated topics,

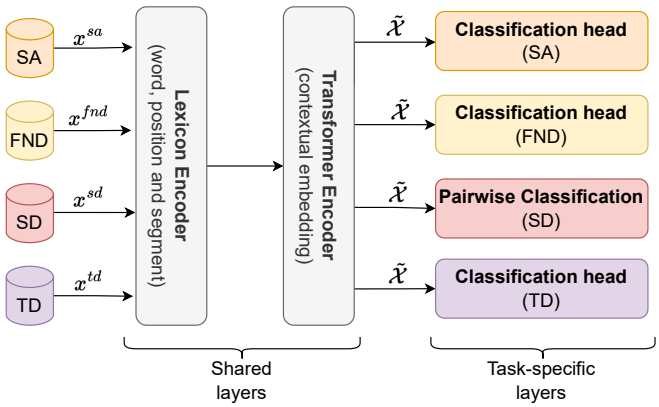


Figure 5.2. Overview of the MT-DNN architecture applied to our study.

while FND and SD display a similar degree of semantic similarity due to overlapping instances.

5.3.2 Methodology

Our methodology comprises two essential components: the utilization of the MT-DNN framework [144] to train models for various task combinations and the incorporation of the SHAP framework [155] to capture the knowledge within these models through local explanations. It is noteworthy that our methodology is adaptable and can be applied seamlessly to different multi-task and explanations frameworks.

Multi-task Component We employ the Multi-Task Deep Neural Network (MT-DNN) introduced by Liu et al. [144], and its architecture is illustrated in Figure 5.2. For task $\mathscr{T}$, the textual input is represented as a sequence of words, denoted as $x^{\mathscr{T}} = \{[CLS], x_1, ..., x_m\}$, where m is the sentence length. The *Lexicon Encoder* maps each token x_i to its corresponding input embedding vector, which is derived by summing the word, segment, and positional embeddings. Subsequently, the pre-trained *Transformer Encoder* transforms the inputs into contextual vectors $\tilde{\mathscr{X}}$, creating a shared representation across all tasks. In our experiments, we explore the influence of different pre-trained models as the backbone of this multi-

task framework. Finally, the *Task-specific layers* function as classification models that use the embeddings of the [CLS] token. Specifically, the classification heads for the SA, FND, and TD tasks are simple feed-forward networks. In contrast, for the SD task, the pairwise classification layer utilizes the [CLS] embeddings of the target text and statement, employing multi-step reasoning for classification [145].

Explanation Component Our goal is to extract the knowledge acquired by models under both single-task and multi-task scenarios. To achieve this, we employ local explanations based on feature-importance scores, utilizing the SHAP framework [155]. It is worth emphasizing that we chose SHAP due to its model-agnostic nature, allowing it to be compatible with any predictive model. In our context, these explanations provide a ranked list of words, reflecting their contribution to the model's prediction for a given instance (see Figure 5.6 for examples). By analyzing these explanations for all instances in a specific task, we can characterize the knowledge acquired by the model for that task.

Formally, the local explanation for a target instance $x^{\mathscr{T}}$ and a classification model $\mathscr{M}_i(\cdot)$ can be represented as a ranked list

$$e_i(x^{\mathscr{T}}) = \{(w_1, p_1), (w_2, p_2), \cdots, (w_m, p_m)\},$$

where w_j represents a word in $x^{\mathscr{T}}$ and $p_j \in \mathscr{R}$ indicates its importance for the classification outcome. Notably, the words are ordered based on their importance scores, with higher scores indicating greater importance.

Consequently, our primary objective of evaluating models' knowledge in single-task and multi-task scenarios can be rephrased as a comparison of SHAP explanations for these models. Specifically, for two models $\mathscr{M}_h$ and $\mathscr{M}_k$, we compare the explanations $e_h(x^{\mathscr{T}})$ and $e_k(x^{\mathscr{T}})$ for all instances $x^{\mathscr{T}}$ belonging to task $\mathscr{T}$. Depending on the evaluation setup and metrics, we compare explanations provided by single-task models to their corresponding multi-task counterparts, thereby revealing the effects of positive and negative transfers.

5.4 Results

5.4.1 Implementation Details

All experiments have been performed on Google Colab equipped with one single core hyper threaded Xeon Processor @2.2Ghz, 12 GB of RAM and a Tesla T4 GPU.

For the MT-DNN framework, we have employed the official implementation[1]. The model is trained for 50 epochs with a batch size of 16. To establish optimal hyperparameters, we have followed a specific strategy: initially, we have conducted preliminary experiments, including a random search to explore various hyperparameters such as learning rate, weight decay, and optimizer settings. Subsequently, we perform a greedy search around default values using a subset of the training data, following the common practice in the literature [309, 175]. Ultimately, we have settled on the following hyperparameters: an *Adamax optimizer* with a *learning rate* of 5e-5, *weight decay* of 0.01, and an *adam epsilon* of 1e-7. To prevent gradient explosion, we set the *grad clipping* parameter to 1.0, and for regularization purposes, we used a *dropout* rate of 0.3.

In our experiments, we consider three pre-trained backbone networks: *bert-base-uncased* for BERT [52], *roberta-base* for RoBERTa [333], and *distilbert-base-uncased* for DistilBERT [229].

For the SHAP method, we have employed the official implementation[2] and utilized KernelSHAP with default hyperparameters. Specifically, when generating explanations for each model, we construct the explainer using the entire training set of the dataset for which we sought explanations. Through empirical analysis, we have observed consistent stability in explanations across different iterations. Although the absolute values of feature importance may exhibit variation due to SHAP's sampling strategy, the overall order of feature importance remains invariant.

[1] https://github.com/namisan/mt-dnn
[2] https://github.com/slundberg/shap

5.4.2 Benefits of multi-task learning (RQ1)

Experimental Protocol

To assess the impact of multi-task learning, we employ the proposed framework to explore all possible task combinations within our set of four target tasks. In detail, we conduct Montecarlo cross-validation, utilizing five different random seeds to partition the datasets into training (60%), validation (20%), and testing (20%) sets. In the following experiments, our chosen classification metrics are accuracy and (macro) F1-score, and we present the averaged performance results across various seeds. To validate the observed improvements, we employ statistical t-tests.

Our experimental design comprises three primary investigations. Firstly, we scrutinize the presence of both positive and negative transfer effects as we vary the number of concurrently trained tasks. Secondly, we introduce a pairwise training approach to delve deeper into these transfer dynamics, considering all possible combinations of tasks. Lastly, we compare our MT-DNN models against three distinct sets of baselines, encompassing both single-task and multi-task models.

Positive and negative transfers

Table 5.1 provides a comprehensive overview of the optimal multi-task configurations for each task, with variations in the number of training tasks. Our findings consistently underscore the efficacy of multi-task learning, as denoted by notable F1-score improvements of 3.26%, 6.57%, 3.01%, and 0.62% across the SA, FND, SD, and TD tasks, respectively, when compared to single-task models. However, it is imperative to note that the improvement in the SD task does not attain statistical significance (p-value > 0.05), suggesting that the single-task model performs equivalently to the best-performing multi-task model.

Remarkably, Table 5.1 also underscores the critical role of task selection, with certain task combinations outperforming both single-task models and alternative multi-task setups. For instance, in the FND task, simultaneous training with both the SD and SA tasks yields statistically superior results in terms of F1-score compared to exclusive training on the SD task. Nevertheless, it is noteworthy that including all tasks does not lead to the highest overall performance. Instead, optimal performance is predomi-

Table 5.1. Performance by varying the number of training tasks. We report the average across the folds of the cross-validation. (bold indicates the best result, underline the first runner up, * indicates statistical significance, at p = .05, with the single-task model, ** indicates statistical significance, at $p = .05$, between the best and runner up models.)

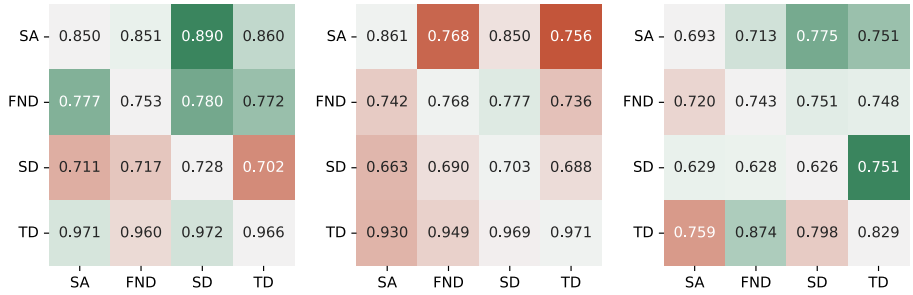
Task	Configuration	Accuracy	F1-score
SA	SA	0.859	0.861
	SA + SD	**0.890***	**0.890****
	SA + SD + TD	<u>0.889*</u>	<u>0.882*</u>
	All	0.848	0.851
FND	FND	0.762	0.768
	FND + SD	<u>0.822*</u>	<u>0.815*</u>
	FND + SD + SA	**0.826****	**0.822****
	All	0.772*	0.772*
SD	SD	**0.766**	<u>0.728</u>
	SD + TD	<u>0.749</u>	**0.751**
	SD + TD + SA	0.720	0.708
	All	0.738	0.712
TD	TD	0.971	0.971
	TD + SA	<u>0.972</u>	<u>0.972</u>
	TD + SA + FND	**0.977****	**0.977****
	All	0.960	0.961

nantly achieved with two tasks (for SA) or three tasks (for FND and TD) in combination. This underscores that the dynamics of positive or negative transfers cannot be solely attributed to data augmentation effects but instead hinge on intricate relationships between the tasks. Moreover, this complexity extends beyond mere task similarity. For instance, multi-task training involving SA, FND, and TD yields significant benefits for the TD task, despite its dissimilarity from these datasets (as illustrated in Figure 5.1). In contrast, the SD task fails to capitalize on its combination with the FND task, despite their higher degree of apparent similarity.

Pairwise training

To delve deeper into the effects of multi-task learning, we adopt a pairwise training strategy [261]. Figure 5.3 presents the F1-scores obtained when the row-indexed dataset is incorporated into multi-task settings with the column-indexed dataset; single-task performance is reported on the diagonal. This experiment considers various backbone networks for the MT-DNN's shared layer, including *roBERTa* [333], *BERT* [52], and *distilBERT* [229]. The results unveil a notable asymmetry in the knowledge transfer process. For example, when concurrently training the SD and FND tasks, we consistently observe a decline in SD's performance compared to the single-task model. In contrast, this same combination consistently improves FND's performance.

Furthermore, it is evident that the occurrence of positive and negative transfers is significantly influenced by the choice of the pre-trained backbone. Specifically, the *roBERTa* model tends to exhibit positive transfer effects, while the *BERT* backbone seldom benefits from MTL. For instance, when the SA task is trained alongside the SD task, we observe a 3.26% performance improvement with *roBERTa* on the former task, whereas the same combination leads to a 12.2% performance degradation with *BERT*. This divergence may be attributed to variations in the pre-training objectives, which impact how the models encode and represent information. Notably, previous studies [96] have demonstrated that *roBERTa*'s training on a denoising objective generates more robust representations compared to *BERT*. Conversely, *BERT*'s masked language modeling objective focuses on local context, encouraging the learning of more task-specific representations [216, 165]. Lastly, models trained with the *distilBERT* backbone



(a) roBERTa **(b)** BERT **(c)** distilBERT

Figure 5.3. Results (in terms of F1-score) of the pairwise training by varying the MT-DNN's backbone. Single-task results are reported on the diagonal and pair-wise multi-task results obtained on the row-indexed dataset are reported when it is used in a multi-task setting with the column-indexed dataset.

exhibit worse performance than *BERT* and *roBERTa*, but intriguingly, a higher number of combinations lead to *positive transfer*. This result likely stems from the distillation process employed in *distilBERT*, which enhances the model's generalizability across different tasks [229].

Comparison with baselines

To further validate the performance of our proposed system and assess the impact of multi-task learning, we compare our approach with two multi-task baselines: (i) AdversarialMTL [143], which employs adversarial training to mitigate interference between shared and private latent feature spaces, and (ii) MaChAmp [282], sharing the same MT-DNN backbone but using a distinct fine-tuning strategy based on customized task weighting. We also introduce the foundational model GPT3.5 [27] into our evaluation, focusing particularly on its performance in zero-shot settings [208]. This choice is influenced by recent findings indicating that such models have demonstrated human-level performance in various text annotation tasks [77, 334], including those relevant to disinformation, such as toxic content classification [135].

Table 5.2 shows the F1-scores for each task under both single-task and multi-task settings. Our approach consistently achieves the best per-

Table 5.2. Comparison, in terms of F1-score, with baselines, under both single-task (ST) and multi-task (MT) settings. GPT3.5 is configured under 0-shot settings. (bold indicates the best result, underline the first runner up)

Method	Configuration	SA	FND	SD	TD
GPT3.5	0-shot	0.844	0.312	0.156	0.870
AdverMTL	ST	0.627	0.487	0.195	0.305
	MT	0.648	0.527	0.190	0.320
MaChAmp	ST	0.859	0.814	0.682	0.960
	MT	<u>0.879</u>	**0.834**	<u>0.729</u>	0.937
Ours	ST	0.861	0.768	0.728	<u>0.971</u>
	MT	**0.890**	<u>0.822</u>	**0.751**	**0.977**

formance across tasks, with the exception of the FND task, where it ranks as the top performer, closely behind MaChAmp. This highlights the effectiveness of our method in optimizing performance across diverse disinformation-related tasks. Furthermore, it is noteworthy that both MaChAmp and our approach significantly outperform AdversarialMTL in all tasks. This performance gap can be attributed to AdversarialMTL's practice of training sequence models from scratch, without leveraging pre-trained language models.

Furthermore, our evaluation sheds light on the specific challenges faced by GPT3.5. In tasks related to factuality, the model exhibits notably poor performance, often categorizing news articles as unverified in the FND task and misinterpreting input sentences in the SD task. However, GPT3.5 performs reasonably well in general text tasks, such as the SA and TD tasks, although it still falls short of multi-task models. While acknowledging that the comparison may not be entirely fair due to the additional fine-tuning required for multi-task models, these findings serve as a cautionary note against the indiscriminate use of foundation models in disinformation detection tasks.

Overall, our experiments consistently highlight the merits of MTL, as most multi-task models outperform their single-task counterparts across the majority of tasks and baselines. This underscores the value of a multi-

task approach in gaining a comprehensive understanding of various aspects of information.

Findings & Remarks Addressing *RQ1*, we observed that multi-task learning generally enhances task performance in the context of disinformation detection. However, we found that the underlying transfer dynamics driving these improvements are intricate and extend beyond conventional performance evaluation. Indeed, these dynamics are not solely determined by the number of training task or datasets' similarity as well as they can be asymmetric as one task can benefit the others but not vice-versa. Overall, this highlights that the dynamics of positive and negative transfers manifest as a complex interplay of factors that defy simplistic characterization but rather require a more in-depth analysis of what the models actually learn during the (multi-task) training process.

5.4.3 Single- and multi-task explanations (RQ2)

Experimental protocol

We now shift our focus towards investigating the similarities and disparities between single-task and multi-task models by leveraging SHAP explanations for all tasks and models.

To facilitate our objective and make this investigation feasible, we consider three distinct models for each task $\mathscr{T}$: $\mathscr{M}_{st}^{\mathscr{T}}$, $\mathscr{M}_{pos}^{\mathscr{T}}$, and $\mathscr{M}_{neg}^{\mathscr{T}}$. These models represent the single-task model, the multi-task model with the highest degree of *positive transfer*, and the multi-task model with the highest degree of *negative transfer*, respectively. Specifically, for a target instance $x^{\mathscr{T}}$, we extract the explanations $e_{st}(x^{\mathscr{T}})$, $e_{pos}(x^{\mathscr{T}})$, and $e_{neg}(x^{\mathscr{T}})$ from $\mathscr{M}_{st}^{\mathscr{T}}$, $\mathscr{M}_{pos}^{\mathscr{T}}$, and $\mathscr{M}_{neg}^{\mathscr{T}}$, respectively. As detailed in Section 5.3.2, $e_i(x^{\mathscr{T}})$ represents a ranked list of words ordered by their importance in determining the classification outcome of the i-th model. Hereafter, we omit the $x^{\mathscr{T}}$ dependence to streamline the notation.

Our experimental protocol comprises two main experiments. Firstly, we assess the similarity of explanations to ascertain the extent to which multi-task learning models diverge from their single-task counterparts. Secondly, we compare the rankings of explanations in terms of hypothesis verification to investigate whether the observed differences are attributable

to the treatment (i.e., multi-task training) rather than random variation stemming from unobserved variables. Finally, we conduct a qualitative analysis to illustrate the distinctions in explanations across the models $\mathscr{M}_{st}^{\mathscr{T}}$, $\mathscr{M}_{pos}^{\mathscr{T}}$, and $\mathscr{M}_{neg}^{\mathscr{T}}$.

Explanations' similarity

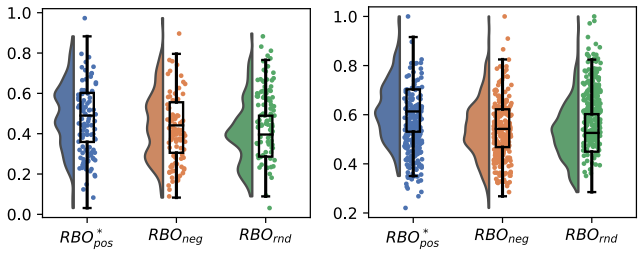
Experimental Setup We gauge the distinctions between SHAP explanations using both the Kendall correlation coefficient τ [3] and the *RBO* metric [297]. It is noteworthy that these chosen metrics do not rely on the absolute values of feature importance but solely on their relative order. This characteristic bolsters the robustness of our analysis, mitigating the effects of randomness introduced by SHAP's sampling technique.

In particular, $\tau : e_i, e_j \rightarrow [-1, 1]$ quantifies the correlation between the rankings of explanations derived from models i and j for the same target instance $x^{\mathscr{T}}$. Similarly, $RBO : e_i, e_j, p \rightarrow [0, 1]$ measures the similarity between the two rankings but employs the parameter p to control the emphasis placed on higher positions in the rankings. We set $p = 0.9$ to attribute 86% of the contribution in the final score to the top-10 features in the explanation [297].

Concretely, we consider the explanations from $\mathscr{M}_{st}^{\mathscr{T}}$ as the reference and evaluate their similarity to the explanations extracted from $\mathscr{M}_{pos}^{\mathscr{T}}$ and $\mathscr{M}_{neg}^{\mathscr{T}}$. Additionally, we introduce a null model $\mathscr{M}_{rnd}$, generating explanations by randomly shuffling the reference explanation. Consequently, for each instance $x^{\mathscr{T}}$ belonging to task $\mathscr{T}$, we compute $RBO(e_{st}, e_{pos})$, $RBO(e_{st}, e_{neg})$, $RBO(e_{st}, e_{rnd})$, and similarly for the τ metric. Given that the reference explanations remain consistent, we omit the dependence on e_{st} to simplify the notation.

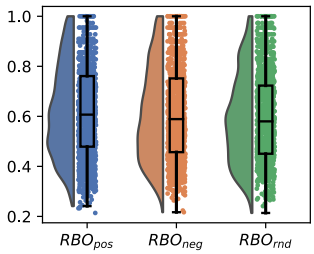
Finally, we employ the Kolmogorov-Smirnov test to compare the distributions of RBO_{pos}, RBO_{neg}, and RBO_{rnd} (resp. τ_{pos}, τ_{neg}, and τ_{rnd}) calculated for all instances within task $\mathscr{T}$.

Results Figure 5.4 shows the distributions of RBO_{pos}, RBO_{neg}, and RBO_{rnd} for each task. A compelling trend emerges as we consistently observe that, on average, RBO_{pos} surpasses RBO_{neg} for all tasks. For instance, in the context of the FND task, the average values of RBO_{pos} and RBO_{neg} stand at 0.613 and 0.542, respectively. This observation im-

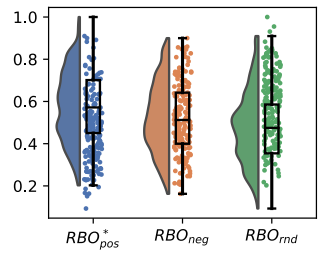


(a) Sentiment Analysis

(b) Fake News Detection



(c) Stance Detection

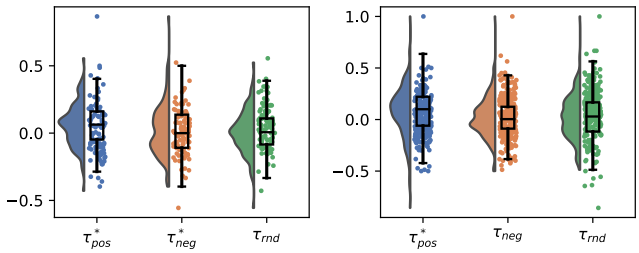


(d) Topic Detection

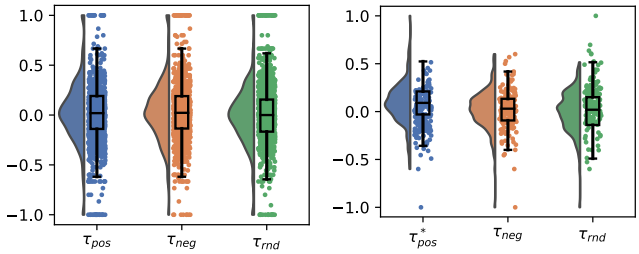
Figure 5.4. The distribution of RBO_{pos}, RBO_{neg}, and RBO_{rnd} for each task. (* indicates statistical difference, at $p = .05$, with respect to RBO_{neg})

plies that the explanations furnished by $\mathscr{M}_{pos}$ exhibit a higher degree of similarity with those from $\mathscr{M}_{st}$ compared to the explanations generated by $\mathscr{M}_{neg}$. This trend is likely attributed to *positive transfer*, which does not fundamentally transform the model's knowledge but instead hones what the model could already discern in single-task scenarios, incorporating additional patterns derived from other tasks.

Furthermore, we delve into the entire distribution rather than focusing solely on average values. We discern that, except for the SD task, the distribution of RBO_{pos} consistently differs from that of RBO_{neg} (p-value$> .05$). This outcome substantiates our prior hypothesis that multi-task *positive transfer* serves to refine the model's knowledge rather than undergoing a revolutionary transformation. It is noteworthy to mention that the SD task exception aligns with our previous discovery that multi-task learning does not enhance this specific task, as the performance of single-task and multi-



(a) Sentiment Analysis **(b)** Fake News Detection



(c) Stance Detection **(d)** Topic Detection

Figure 5.5. The distribution of τ_{pos}, τ_{neg}, and τ_{rnd} for each task. (* indicates statistical difference, at $p = .05$, with respect to τ_{neg})

task models remains comparable. In contrast, the distribution of RBO_{neg} never deviates from the distribution of RBO_{rnd} (p-value $> .05$), signifying that *negative transfer* diminishes the model's knowledge to such an extent that the explanations provided by multi-task models can be regarded as random perturbations of the explanations produced by single-task models.

Lastly, Figure 5.5 illustrates the distributions of τ_{pos}, τ_{neg}, and τ_{rnd} for each task. In essence, this reaffirms the insights gleaned from the RBO analysis, as τ_{pos} is always greater than τ_{neg} across all tasks. Additionally, τ_{neg} is statistically equivalent to τ_{rnd} for all tasks, underscoring the influence of *negative transfer* in rendering the explanations provided by multi-task models akin to random perturbations of those generated by single-task models.

Explanations' hypothesis test

Experimental Setup We perform a revised version of the test proposed by [230] to quantify the difference between two performed rankings (a.k.a. treatments) in terms of hypothesis verification. Specifically, for each task $\mathscr{T}$, we compare the explanation rankings e_{pos} and e_{neg} generated by $\mathscr{M}_{pos}^{\mathscr{T}}$ and $\mathscr{M}_{neg}^{\mathscr{T}}$, respectively, with the reference explanation rankings e_{st} provided by $\mathscr{M}_{st}^{\mathscr{T}}$.

In particular, we consider the explanation provided by the single-task model $e_{st}(x^{\mathscr{T}})$ as the ground-truth ordering, i.e., $L^t \equiv e_{st}(x^{\mathscr{T}}) = \{w_1, w_2, \cdots, w_n\}$, n being the number of words of $x^{\mathscr{T}}$. Each word in the ordering has also associated a measure of relevance $S(w_i, e_{st}(x^{\mathscr{T}}))$ (hereafter $S(w_i)$) dependent on its position in L^t. Formally, $S(w_i) \in [0, 1]$ and is such that $\forall i \in [1, (n-1)]$, we have $S(w_i) \geq S(w_{i+1})$. We set $S(w_i) = 1$ for the top-5 positions, $S(w_i) = 0.5$ from the fifth to the tenth positions, and $S(w_i) = 0.05$ for all other positions in the ranking.

Then, L^t is compared with an experimental ordering $L^d = \{w_{\pi_1}, \ldots, w_{\pi_n}\}$ where $\{\pi_1, \ldots, \pi_n\}$ is a permutation of $1, \ldots, n$. As a result, the displacement of w_i is defined as $d(w_i) = |i - \pi_i|$, while the relative displacement of L^d is computed as follows:

$$\Theta(x^{\mathscr{T}}) = \frac{\sum_i S(w_i) d(w_i)}{\Omega},$$

$\Omega = \lfloor \frac{n^2}{2} \rfloor$ being a normalization factor.

Under these settings, the null hypothesis H_0 is that all *relative displacements* measured for different instances are the same, and we test its acceptance with the F ratio [161]. In particular, we have $m = 2$ treatments, i.e., the training process to obtain $\mathscr{M}_{pos}$ and $\mathscr{M}_{neg}$, and $k = 50$ measurements of Θ_{pos} and Θ_{neg}. These measurements are derived by explaining 50 randomly selected instances from task $\mathscr{T}$. Let Θ_{ij} be the weighted displacement of the j-th experiment on the i-th treatment. We define $\mu_i = \frac{1}{k} \sum_{j=1}^{n} \Theta_{ij}$ as the average result for the i-th treatment, $\mu = \frac{1}{m} \sum_{i=1}^{m} \mu_i$ as the total average, $\sigma^2 = \frac{k}{m-1} \sum_{i=1}^{m} (\mu_i - \mu)^2$ as the variance *between* treatments and $\sigma_W^2 = \frac{1}{m(k-1)} \sum_{i=1}^{m} \sum_{j=1}^{k} (w_{ij} - \mu_i)^2$ as the variance *within* treatments. Then, the F ratio is computed as $F = \frac{\sigma^2}{\sigma_W^2}$.

Table 5.3. Hypothesis test: weighted displacements average (μ) and variance σ^2. (* indicates statistical validity, at $p = .05$)

Task	Treatment	μ	σ^2	F
SA	pos	0.194	0.009	9.97*
	neg	0.257	0.015	
FND	pos	0.220	0.008	23.3*
	neg	0.317	0.009	
SD	pos	0.381	0.027	0.260
	neg	0.391	0.027	
TD	pos	0.201	0.005	10.8*
	neg	0.268	0.009	

Results Table 5.3 presents compelling evidence as it consistently reveals the rejection of the null hypothesis for all tasks, except in the case of the SD task. This unequivocally suggests that the observed disparities in the explanations can largely be ascribed to the specific treatments applied, i.e., the multi-task training process.

Moreover, the rejection in favor of $\mathscr{M}_{pos}$, represented as $\mu_{pos} < \mu_{neg}$, lends further support to our earlier deduction. It corroborates the notion that the explanations stemming from *positive transfer* showcase a higher degree of resemblance to single-task explanations in comparison to those originating from *negative transfer*.

Explanations' qualitative analysis

We embark on a qualitative comparison of the explanations furnished by $\mathscr{M}_{st}$, $\mathscr{M}_{pos}$, and $\mathscr{M}_{neg}$ across diverse tasks. In a comprehensive overview, it becomes evident that MTL exerts a regulatory influence on feature importance. Multi-task models tend to distribute importance across a wider spectrum of words when contrasted with their single-task counterparts. This phenomenon can be attributed to the nature of *positive transfer*, which redistributes importance based on knowledge gained from

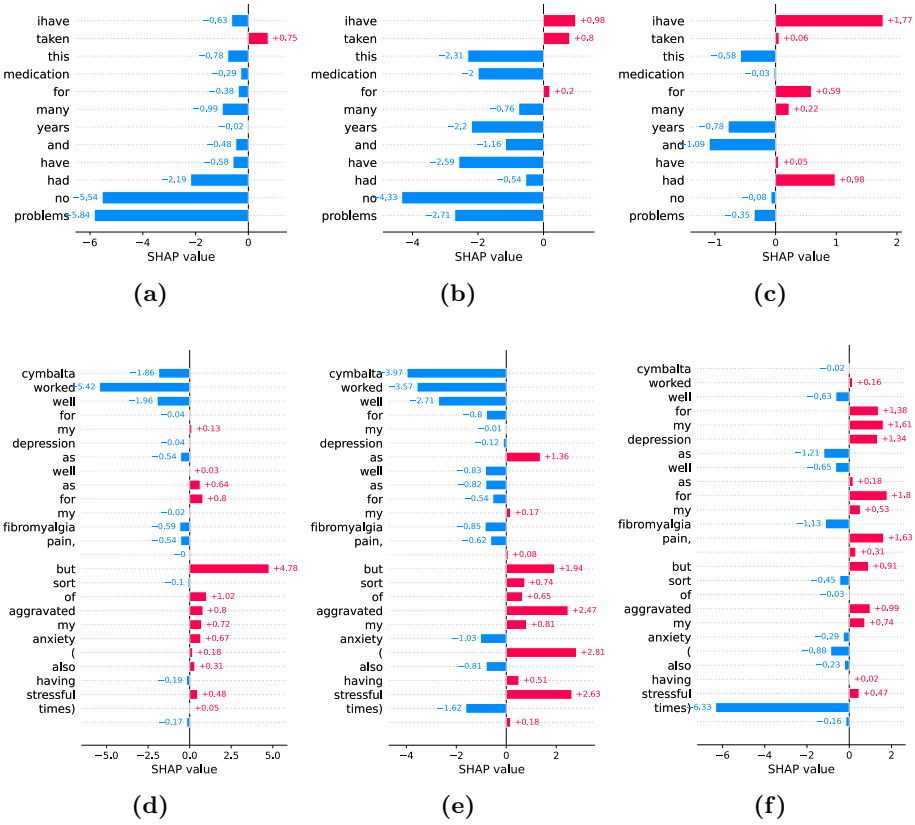


Figure 5.6. Qualitative analysis: explanations of $\mathscr{M}_{st}$ (left), $\mathscr{M}_{pos}$ (center), and $\mathscr{M}_{neg}$ (right) for two samples in the SA dataset. Blue (resp. red) bars refers to features positively (resp. negatively) contributing to the chosen class ("positive" sentiment).

other tasks, while *negative transfer* leads to a haphazard redistribution of feature importance.

Figure 5.6 offers a tangible representation of this phenomenon within the context of the SA task. In particular, Figures 5.6a, 5.6b, and 5.6c showcase the explanations produced by $\mathscr{M}_{st}$, $\mathscr{M}_{pos}$, and $\mathscr{M}_{neg}$ for the input instance "I have taken this medication for many years and have had no problems". Intriguingly, both $\mathscr{M}_{st}$ and $\mathscr{M}_{pos}$ correctly predict a "positive" sentiment for this input, whereas $\mathscr{M}_{neg}$ misclassifies it. A closer examination of these explanations reveals distinct patterns. $\mathscr{M}_{st}$ primarily focuses on the sub-sentence "had no problem", whereas $\mathscr{M}_{pos}$ assigns importance to the entire sentence. This reallocation of importance can be attributed to the influence of stance detection, as $\mathscr{M}_{pos}$ is trained on both the SA and SD tasks. Consequently, the explanation from $\mathscr{M}_{pos}$ encapsulates not only the sentiment but also the subject's stance towards the medication.

Additionally, Figures 5.6(d-e) showcase an additional instance with "positive" sentiment, i.e., *"Cymbalta worked well for my depression as well as for my fibromyalgia pain, but sort of aggravated my anxiety (also having stressful times)"*. In this scenario, $\mathscr{M}_{st}$ and $\mathscr{M}_{neg}$ misclassify the input, while $\mathscr{M}_{pos}$ correctly predicts the "positive" sentiment. Upon examining the explanations from $\mathscr{M}_{st}$ and $\mathscr{M}_{pos}$ (Figures 5.6d and 5.6e, respectively), a shared proficiency emerges in identifying the positive sentiment conveyed in the initial segment, *"Cymbalta worked well"*. Moreover, both models effectively comprehend the subsequent revelation about increased anxiety as a side effect of the medication. However, upon closer scrutiny, a subtle yet pivotal nuance surfaces. The inclusion of the phrase *"(also having stressful times)"* introduces a contextual layer wherein the escalation of anxiety is intricately linked to the user's psychological state - specifically, stress stemming from factors unrelated to the medication. Remarkably, $\mathscr{M}_{st}$ falters in recognizing the contextual significance of this phrase. It overlooks the weight carried by this segment of the text and fails to attribute it any substantial importance. In contrast, $\mathscr{M}_{pos}$ adeptly comprehends that the user's experience is not exclusively tied to the medication's impact; rather, it is substantially influenced by their emotional state. We conjecture that this capability may arise from $\mathscr{M}_{pos}$'s training on the SD task, enabling it to adeptly gauge the overall stance in support of the medication's usage while concurrently isolating external factors.

Finally, we corroborate our prior quantitative results by scrutinizing the explanations produced by $\mathscr{M}_{neg}$ (depicted in Figures 5.6c and 5.6f). These explanations starkly highlight the adverse consequences of *negative transfer*, vividly illustrating the disorderly reshuffling of feature importance in contrast to their single-task model counterparts.

Findings & Remarks In response to *RQ2*, we found that *positive transfer* does not fundamentally alter the knowledge base of any model but rather refines the model's existing capabilities by incorporating additional patterns derived from other tasks. In contrast, *negative transfer* has a detrimental effect on the model's knowledge, to the extent that the explanations provided by multi-task models are akin to random perturbations when compared to their single-task counterparts. Specifically, our qualitative findings suggest that *positive transfer* enriches the explanations generated by the models. These enriched explanations are not only suitable for the target task for which they are generated but also exhibit suitability for all other tasks involved in the multi-task training process.

5.5 Discussion

5.5.1 Contributions

In this chapter, we delved into the realm of disinformation detection, with a specific focus on sentiment analysis, fake news detection, stance detection, and topic detection tasks. We proposed a framework based on the MT-DNN framework [144] and harnessed SHAP explanations [155] to facilitate a meticulous comparative analysis of knowledge acquisition within single-task and multi-task paradigms.

Our investigation unfolded several noteworthy insights. Across the spectrum of disinformation detection tasks, the proposed (multi-task) system generally exhibited a propensity for augmenting task performance, with the notable exception of the stance detection task. However, these performance enhancements were underpinned by intricate transfer dynamics that transcended conventional performance evaluation metrics. In this intricate landscape, the concept of *positive transfer* revealed itself as a multifaceted phenomenon, influenced not merely by the number of train-

ing tasks or the degree of similarity between datasets. Instead, it exhibited intriguing asymmetry, where one task could bestow benefits upon others without commensurate reciprocity.

Moreover, our deep dive into the explanations generated by models trained under multi-task settings unveiled compelling insights into the mechanisms of knowledge transfer. Specifically, *positive transfer* emerged as a force of refinement, fine-tuning existing knowledge by incorporating additional patterns gleaned from other tasks. In stark contrast, *negative transfer* cast a shadow of disruption, severely undermining the integrity of models' knowledge, ultimately yielding explanations that bore a striking resemblance to randomness. These revelations underscore the multifaceted nature of multi-task learning in disinformation detection, challenging conventional wisdom and paving the way for more nuanced and insightful approaches in the domain.

5.5.2 Limitations

Our study is subject to several limitations that warrant acknowledgment. Firstly, despite the additional validation from various MTL baselines, it is important to recognize that our analyses may exhibit a certain degree of bias due to our choice of the SHAP explanation tool [155] and the MT-DNN architecture [144] as building blocks of our framework.

Secondly, our comparative examination of explanations predominantly relies on the absolute rankings of features and does not account for their weighted importance. Although we indirectly address this limitation by incorporating the RBO metric, which assigns higher significance to features with superior rankings, the inherent focus on rankings remains.

Thirdly, while our dataset collection encompasses diverse sources, encompassing both similar and dissimilar data, we are aware that our study may bear the imprint of these specific base datasets. This influence is a potential limitation to the generalizability of our findings.

Finally, we acknowledge that the domain of disinformation detection is multifaceted, encompassing a wide spectrum of challenges beyond the specific tasks scrutinized in our study. For instance, the detection of fake news spreaders represents a critical facet that remains unexplored within our current research scope. Our future endeavors aim to expand the purview of our investigation to encompass these vital tasks.

5.5.3 Conclusions and Future Works

Our study yields two pivotal insights. Firstly, multi-task learning demonstrates considerable potential in the domain of disinformation detection, as it generally enhances performance across a spectrum of sub-tasks. Nevertheless, it is essential to underscore that the degree of effectiveness in this approach can vary, with certain tasks reaping more substantial benefits than others.

Secondly, our findings emphasize the significance of holistic system optimization in designing effective multi-task systems for disinformation detection. This optimization should transcend the conventional focus on performance metrics and encompass a comprehensive evaluation of the intricate transfer dynamics at play within the system.

As we navigate the path forward in our research, we intend to extend the applicability of our findings to encompass a broader range of multi-task frameworks and explanation mechanisms. Furthermore, our future endeavors will delve into leveraging explanations to proactively identify task groups that stand to gain the most from multi-task settings. Finally, we aspire to harness the insights gleaned from our study to craft an explanation-driven learning mechanism that enhances the seamless transfer of information across disparate tasks, thereby advancing the field of disinformation detection.

Chapter 6

Towards Collaborative Moderation: Sharing Social Media Moderation Intervention

6.1 Research Context and Contributions

Social media platforms play a pivotal role in shaping the contemporary digital information landscape, providing users with the means to engage in discussions across a diverse spectrum of topics encompassing public health, information technology, and socio-political issues. Nonetheless, the open nature of these platforms, coupled with relatively lenient content moderation policies, poses a potential threat to the integrity of these digital ecosystems when harmful content, such as fake news, propaganda, and inappropriate or violent materials, is disseminated and proliferates throughout the online population.

Prominent social media platforms like Facebook and Twitter endeavor to preserve the integrity of their virtual realms by enforcing conduct guidelines and implementing a variety of moderation interventions that target both harmful content and the users responsible for its dissemination. These interventions encompass actions such as content flagging, demotion, or removal, as well as the temporary or permanent suspension of offending

users.

However, these moderation endeavors are often implemented in isolation, without sufficient regard for interventions carried out on other platforms regarding harmful content that has found its way into their domain. This unilateral approach presents risks, as content originating on a *source platform* can migrate to other *target platforms*, where it may garner attention within specific communities and reach a broader audience. For instance, the cross-platform dissemination of anti-vaccine content across platforms like YouTube and Twitter has led to extensive amplification and virality across multiple online spaces [79, 40]. Furthermore, recent research has revealed that moderation efforts on a source platform can inadvertently foster the proliferation of harmful content on target platforms [226]. For instance, the removal of anti-vaccine groups on Facebook has been found to boost engagement with anti-vaccine content on Twitter [172]. Similarly, when YouTube decided to demote conspiratorial content, certain Reddit communities actively promoted these demoted videos on their platform, leading to their viral spread and undermining YouTube's moderation strategy [28]. Additionally, from the users' perspective, accounts banned from platforms like Twitter or Reddit tend to exhibit increased levels of toxic behavior when they migrate to low-moderated spaces like Gab or Bitchute [10].

Overall, this underscores the necessity for cross-platform moderation strategies that acknowledge the interconnected nature of the digital information ecosystem. A mere content removal or user suspension on one platform may fall short in mitigating the dissemination of inappropriate content across various platforms. Therefore, fostering collaboration among social media platforms becomes not only advisable but also practically advantageous. Being aware of content deemed inappropriate on other platforms can inform moderation strategies and aid in the early detection of similarly harmful or related content. Recent studies have demonstrated how cross-platform strategies can be instrumental in moderating radical content [219, 61] or identifying inauthentic activities by tracking users' behavior across multiple platforms [10, 172, 40].

Our research [73] delves into a distinctive perspective, focusing on YouTube (YT) as the *source platform* and Twitter as the *target platform*. Specifically, we investigate the prevalence of moderated YT videos on Twit-

ter, i.e., videos initially shared on Twitter but subsequently removed from YouTube. Additionally, we characterize the Twitter users responsible for sharing these YT videos, whom we refer to as *YouTube mobilizers* [38], taking into account various dimensions, including their political affiliations and potential involvement with "fringe" platforms.

In particular, we aim to answer the following research questions (RQs):

RQ1 What is the prevalence, lifespan, and reach of moderated YT videos that are shared on Twitter?

RQ2 What are the characteristics of the mobilizers of moderated YT videos? And, are there any differences with the mobilizers of non-moderated YT videos?

RQ3 Do the mobilizers of moderated YT videos receive significant engagement from the Twitter population?

Leveraging a large-scale dataset related to the 2020 U.S. election [35], we observed that YouTube is the most shared mainstream social media platform on Twitter. Through the use of the YouTube API to retrieve video metadata, we identified that 24.7% of the videos shared in the election discourse were subject to moderation on YouTube. Furthermore, we found that these moderated videos exhibited notably higher levels of dissemination compared to non-moderated ones, and they received more shares than content from other mainstream and fringe social media platforms, such as Gab and 4chan.

An examination of Twitter users sharing YT videos revealed distinct patterns. Users sharing moderated videos predominantly engaged with YT content through retweets, while users sharing non-moderated videos actively disseminated YT content via their original tweets or replies. Intriguingly, more than half of the users in the former group had been suspended by Twitter. Surprisingly, there was a higher incidence of accounts involved in information operations, as identified by Twitter, within the latter group. Additionally, users sharing moderated YT videos predominantly expressed support for Trump and propagated claims of election fraud, whereas users sharing non-moderated videos collectively exhibited less extreme political leaning, encompassing supporters of both Biden and Republican representatives who did not endorse Trump's political campaign. Finally, we

observed that users sharing moderated and non-moderated YT videos tend to interact within their respective groups and display analogous interaction patterns, primarily through retweets. This phenomenon implies the formation of segmented communities akin to echo chambers.

Overall, our findings shed light on the intricate dynamics of cross-platform information diffusion and underscore the necessity for a more comprehensive approach to content moderation.

6.2 Related Work

6.2.1 Cross-platform moderation

In a concerted effort to safeguard the integrity of their respective digital domains, mainstream social media platforms employ a variety of intervention strategies. These strategies encompass actions targeted at both inappropriate content (comprising measures like flagging, demotion, or deletion) and the users responsible for its dissemination (encompassing temporary or permanent suspensions). Nevertheless, the effectiveness of these interventions is under increasing scrutiny, with researchers and policymakers advocating a proactive and comprehensive approach, as opposed to the prevailing isolated and reactive solutions [299, 55, 204]. Indeed, even if moderation interventions prove effective within the confines of an individual platform, their consequences may transcend platform boundaries, giving rise to unintended detrimental activities.

For instance, [10] demonstrated that following the suspension of radical communities on Reddit, users migrated to alternative platforms, where they became more active and shared content of a more toxic nature. A comparable pattern emerged after the de-platforming of Parler, as users shifted to other fringe social media platforms like Gab and Rumble [97]. Furthermore, [226] revealed that the antisocial behaviors of migrated users could have repercussions on mainstream platforms through interactions with non-radical users active across multiple platforms. In this context, [172] discovered that when Facebook banned certain anti-vaccine groups, the toxic content propagated by these groups resonated on Twitter. Additionally, during the 2020 U.S. election, videos removed from mainstream platforms were subsequently re-uploaded on the less-moderated BitChute

platform [262].

The above-mentioned studies underscore the imperative for proactive and collaborative moderation approaches to ensure the overall integrity of the digital information ecosystem. In this study, we delve into the potential advantages of social media platforms openly sharing information pertaining to their moderation interventions. This investigation takes shape through the examination of users posting moderated YouTube (YT) videos on Twitter.

6.2.2 Cross-platform spread of YouTube content

The dissemination of multimodal information across various platforms, incorporating images and videos, carries substantial implications due to the heightened appeal and perceived credibility of multimedia content compared to textual posts alone [90]. Particularly, the cross-posting of videos, including potentially harmful content, across multiple social media platforms constitutes a well-documented concern within the realm of scientific literature [299, 81]. To exemplify, a notable volume of dubious YouTube (YT) videos was shared on Twitter with the aim of sowing doubt concerning the COVID-19 vaccination campaign [203]. Additionally, [40] underscored that anti-vaccine YT videos shared on Twitter experienced heightened visibility and propagation on YouTube. Similarly, [192] identified YouTube as one of the primary channels utilized by the infamous "Disinformation Dozen" to disseminate Covid-19-related conspiracy theories on Twitter. Furthermore, [81] revealed the Internet Research Agency's (IRA) use of YouTube content in their 2016 Twitter propaganda campaign, and more recently, [299] reported the incorporation of YT content to support anti-White Helmet operations in 2020.

Collectively, the studies mentioned above illustrate that harmful YT content is not confined solely to the *source platform* but often thrives in other *target platforms*, particularly with the intention of influencing vulnerable and fringe communities. In this context, the actors responsible for mobilizing and disseminating such detrimental content can assume various identities, including bots, sockpuppets, influential figures, or even individuals susceptible to misinformation and conspiracy theories. This study delves into the characterization of these entities, focusing on Twitter users who share moderated YT videos, and investigates their commonalities and

disparities compared to users sharing non-moderated YT content.

6.3 Methodology

In this section, we describe the data used in the analysis and detail the methodology used to understand the prevalence of YT moderated content on Twitter (*RQ1*) and to characterize users who share moderated YT videos (*RQ2* and *RQ3*).

6.3.1 Data Collection

Our dataset comprises election-related tweets, gathered via Twitter's streaming API service, during the lead-up to the 2020 US election [35]. This specific data collection spans six months, ranging from June 2020 to December 2020, encompassing the latter part of the electoral campaign and the subsequent aftermath characterized by the widespread dissemination of misleading claims and election integrity conspiracies [267, 67]. Over this observation period, we collected a vast repository of more than 600 million tweets, originating from 7.5 million distinct users. Of particular interest are tweets featuring YouTube (YT) videos, constituting approximately 0.65% of the entire dataset, which amounts to 3.9 million messages. Notably, it is important to clarify that our consideration of YT videos does not encompass URLs to YT channels.

Figure 6.1 illustrates the prevalence of tweets and users sharing YT content, consistently surpassing those directing their attention to other mainstream social media platforms. This trend aligns with prior findings [38, 4]. Moreover, both categories exhibit an upward trajectory during the second half of 2020, likely due to the impending election on November 3rd, 2020.

A total of 527,000 YT videos were disseminated on Twitter by 830,000 users. Leveraging the YouTube API, we were able to retrieve diverse video metadata, including the crucial capability to identify whether a specific video had been removed from the platform. Intriguingly, among all YT videos shared, approximately 24.7% (130,000 out of 527,000) were subjected to moderation. Prior to the moderation intervention, these videos were shared on Twitter by 34.5% of users, equating to 287,000 out of

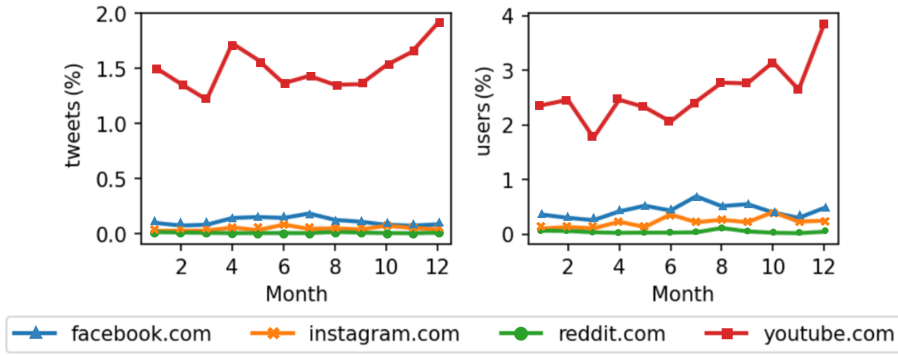


Figure 6.1. The monthly percentage of tweets (left) and users (right) who shared a link to a *mainstream* social media platforms in 2020

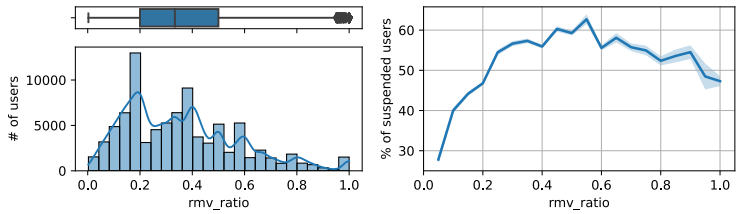
830,000. Furthermore, we collected YT video metadata for non-moderated videos, encompassing details such as video titles, descriptions, tags, and the channel responsible for publishing the video. It's worth noting that gathering metadata for moderated videos, including information on the intervention date and the specific reason(s) for moderation, is not feasible due to restrictions imposed by YouTube. However, a reasonable assumption can be made that a video remains accessible when shared in an original tweet, as this necessitates the user to include the video URL in the Twitter post.

6.3.2 Identifying mobilizers of moderated YouTube videos

To identify the mobilizers of moderated YT videos, we first consider the most active YT mobilizers, as our objective is to investigate the characteristics and behaviors of users who repeatedly (rather than occasionally) post YT content. For this reason, we consider users who shared at least 5 YT videos on Twitter, which results in a set of 113k Twitter users.

Subsequently, we subdivide this cohort of YT mobilizers into two distinct groups based on the quantity of shared *moderated* YT videos.

For each user denoted as u, we introduce the concept of the *ratio of moderated videos* (*rmv(u)*), defined as the proportion of moderated YT videos out of the total number of YT videos shared by user u during



(a) **(b)**

Figure 6.2. YT Mobilizers characteristics with (a) Distribution of the *rmv(u)*; (b) Percentage of the suspended users with respect to their *rmv(u)*. The shaded area is the 95% confidence interval.

our observation period. Using this metric, we define the group of users sharing non-moderated YT videos (**NMYT**) as individuals possessing an $rmv(u) = 0$. This classification results in a subset of 25,400 Twitter accounts. Then, by examining the distribution of the *ratio of moderated videos* (see Fig. 6.2a), we identify the users categorized as mobilizers of moderated YT videos (**MYT**). These are individuals whose $rmv(u)$ is equal to or exceeds 0.5, encompassing all users with an $rmv(u)$ higher than the 75th percentile of the distribution. This selection criterion allows us to focus specifically on users with a pronounced inclination toward sharing moderated videos, thereby excluding those who sporadically engage in disseminating moderated YT content.

To validate this selection criteria, we assess whether the accounts sharing moderated videos remain active on Twitter or have been subject to suspension. Figure 6.2b depicts the percentage of suspended users as a function of *rmv(u)*. It is noteworthy that the percentage of suspended users does not exhibit a significant increase beyond the point where users share more than 50% of moderated videos. Additionally, there exists a positive correlation between the probability of suspension by Twitter and the *rmv(u)* value, as indicated by a Spearman correlation coefficient of 0.451.

6.4 Case Study Results

6.4.1 Prevalence of moderated YouTube videos on Twitter (RQ1)

To answer RQ1, we perform an analysis of the consumption of YT videos on Twitter comparing moderated vs. nonmoderated YT videos. It is worth noting that moderated videos typically have a shorter duration of sharing on Twitter, averaging around 20 days, while non-moderated videos enjoy a more extended lifespan, with an average of approximately 50 days. This discrepancy can be attributed to YT's moderation interventions. To ensure a fair comparison between the two, we examine the number of tweets containing YT videos during the initial week after their first appearance on Twitter.

Figure 6.3a illustrates the distributions of the tweet count for both non-moderated and moderated videos. Notably, the distribution for moderated videos exhibits a right heavy-tail, indicating that when initially posted on Twitter, moderated videos stimulate a higher volume of sharing activity compared to non-moderated videos. This observation is corroborated by the results of a Mann–Whitney test (p-value< 0.01). This finding aligns with prior research [148], which investigated COVID-19-related YT videos and similarly noted that moderated videos tend to elicit more active engagement from viewers.

To gain further insights into the prevalence of moderated YT content on Twitter, we compare the interaction levels with content originating from other social media platforms. Specifically, Figures 6.3b and 6.3c show the volume of tweets and retweets involving moderated YT videos, URLs directing to *mainstream* online social networks, and URLs redirecting to fringe platforms [295].

We observe that the volume of tweets linking *moderated* YT videos alone is greater than the volume of tweets pointing to any other social media platform. According to [38], we find that *fringe* content supplied by *Parler* and *BitChute* is outnumbered by the content provided by mainstream platforms.

Findings and Remarks Addressing *RQ1*, we discovered that moderated YouTube videos were widely shared on Twitter during the run-up and

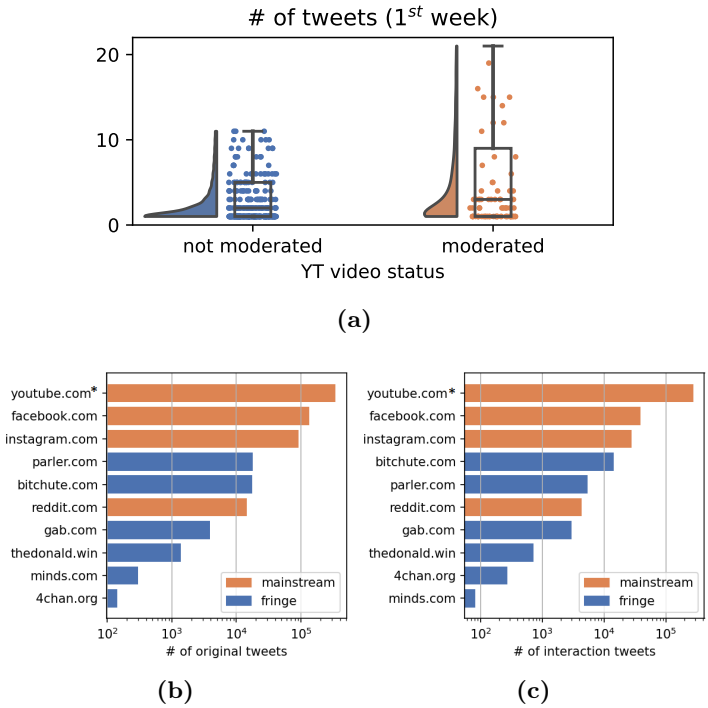


Figure 6.3. The prevalence of moderated YT videos with: (a) The distribution of the number of tweets sharing each video during the week after its first share; (b) Number of original tweets containing a link to each social media platform (Log-scale); (c) Number of retweets containing a link to each social media platform (Log-scale)

aftermath of the 2020 US election. In addition, moderated videos received a higher volume of interactions in the early days of their lifespan. Additionally, moderated YT content alone received more engagement than whole content from other *mainstream* platforms.

6.4.2 YouTube Mobilizers (RQ2)

To address RQ2, we characterize YT mobilizers that share moderated (MYT) and non-moderated videos (NMYT) and investigate whether these users show significantly different behaviors and characteristics across three

dimensions:

- *Cross-Posting Activity*: We delve into users' sharing behaviors on Twitter, specifically their engagement in cross-posting content from various sources, including both *mainstream* and *fringe* platforms.
- *Trustworthiness of the account*: We investigate whether MYT and NMYT users differ in terms of account verification status and the presence of automated bot accounts. We also examine their potential involvement in information operations.
- *User Interests*: Our analysis extends to the political leanings and areas of interest exhibited by mobilizers, both within the Twitter and YT platforms.

Cross-Posting

We consider all possible user activities on Twitter, including posting *original* tweets, engaging in *replies*, and sharing tweets via *retweets* or *quotes*.

Figure 6.4 reveals that while both MYT and NMYT users exhibit similar behaviors in posting *original* tweets, MYT users tend to engage more in retweeting and less in replying compared to NMYT users. This behavior discrepancy regarding retweets is further characterized by assessing the proportion of retweets that contain external web domain links, excluding YouTube. Notably, MYT users frequently retweet content with external links, with approximately 50% of their retweets containing URLs, compared to 28% for NMYT mobilizers. This suggests that NMYT users demonstrate more diversified activity on Twitter, whereas MYT users exhibit a tendency to passively consume and disseminate content, especially when it directs to external resources.

To delve deeper into understanding the two groups of mobilizers, we examine their interaction with YT videos on Twitter. Specifically, we categorize users as either *producers* or *consumers* of a YT video, depending on whether their initial share of the video is in the form of an original tweet or reply (producers) or a retweet (consumers). We then introduce the *prod_ratio* metric, representing the proportion of YT videos that a user *produces* out of the total videos they engage with. It is worth to note

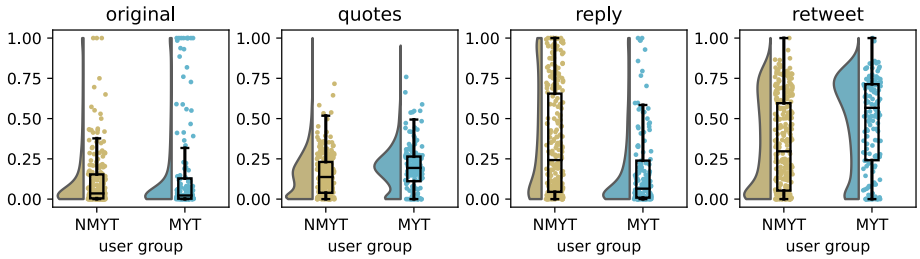


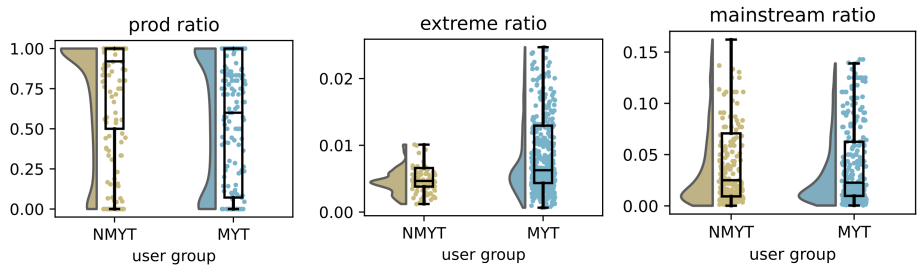
Figure 6.4. The distribution of original tweets, replies, retweets and quotes for NMYT and MYT mobilizers

that being a "video producer" does not imply that the user is the actual publisher of the video on YT.

Interestingly, we find that in both MYT and NMYT groups, mobilizers predominantly fall into either the *producers* category (*prod_ratio* > 80%) or the *consumers* category (*prod_ratio* < 20%). In the NMYT group, there are 15,761 (62.1%) producers and 4,439 (17.4%) consumers, while the MYT group consists of 4,630 (31.9%) producers and 3,742 (25.8%) consumers. Figure 6.5a illustrates the distribution of the *prod_ratio* for both groups, demonstrating that NMYT mobilizers primarily take on the role of producers. In contrast, the MYT group comprises an equal number of producer and consumer accounts. This supports the earlier finding, confirming that MYT users are more inclined to passively retweet content, while NMYT users actively engage through replies and quotes.

Furthermore, we investigate how mobilizers interact with content from various social media platforms, including *mainstream* platforms like Facebook, Instagram, and Reddit, as well as the *fringe* platforms highlighted in Fig. 6.3b. We introduce the *extreme ratio*, which represents the fraction of tweets that contain extreme URLs out of the total number of tweets shared by each user.

While Figure 6.5c shows similar distributions in terms of *mainstream ratio* for MYT and NMYT mobilizers (p-value > 0.01), with both groups sharing approximately 3.3% of *mainstream* content on average, the results differ for *fringe* platforms. Specifically, Figure 6.5b indicates that MYT users engage more frequently with *fringe* platforms compared to NMYT mobilizers, despite both groups having similar mean values (0.5%



(a) **(b)** **(c)**

Figure 6.5. The distribution of *prod_ratio*, *extreme ratio* and *mainstream ratio* for NMYT and MYT mobilizers

and 0.7% for NMYT and MYT mobilizers, respectively). This analysis suggests that while the two groups do not significantly differ in their interactions with *mainstream* platforms, MYT mobilizers share more content from less-moderated online spaces than NMYT users. This implies a degree of endorsement of extreme ideas promoted on *fringe* platforms [295].

Trustworthiness

Here, we delve into the characteristics and account statuses of the two mobilizer groups. Recognizing the influential roles played by political elites, bot accounts, and state-backed trolls in orchestrated campaigns and (mis)information operations [154, 60, 81, 304, 192], we seek to identify the entities responsible for propagating moderated or non-moderated YT videos on Twitter. To achieve this, we employ Botometer [307] and the Twitter API to assess whether our mobilizers consist of automated or verified accounts, respectively. As revealed in Figure 6.6, both groups comprise a meager number of verified accounts (268 in the NMYT group and 19 in the MYT group) and bot accounts (2,234 in the NMYT group and 586 in the MYT group).

Subsequently, we explore whether users in the two mobilizer groups have been suspended as a result of Twitter's moderation interventions. Indeed, during the 2020 US election, the platform made heightened efforts to safeguard the integrity of discussions, which included labeling suspicious

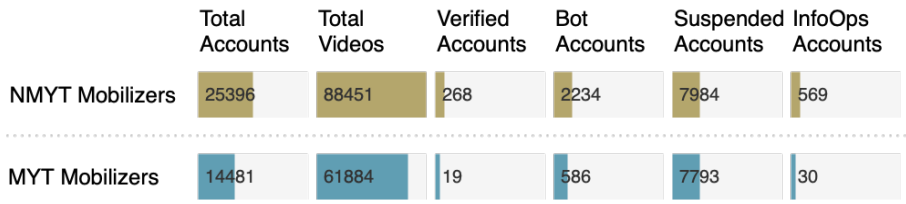


Figure 6.6. The number of accounts in each mobilizer group that were verified, bots or suspended. The columns are as follows: "Total Accounts" is the total number of accounts in each group. "Total Videos" is the number of unique YT videos shared by each group. "Verified Accounts" is the number of verified accounts in each group. "Bot Accounts" is the number of accounts labeled as a bot by the Botometer API in each group. "Suspended Accounts" is the number of accounts in each group that were later suspended by Twitter. "InfoOps Accounts" is the number of (suspended) accounts involved in information operation in each group

or misleading content and suspending accounts involved in information operations [153, 228]. In our analysis, we are keen on understanding the extent to which MYT and NMYT mobilizer accounts were affected by these measures. Our findings indicate that both groups encompass suspended accounts, although in varying proportions. As illustrated in Figure 6.6, approximately 53.8% of MYT mobilizers (totaling 7,793 accounts) have faced moderation by Twitter. In contrast, around 31.4% of NMYT users (equivalent to 7,984 accounts) have been suspended. Furthermore, we extend our inquiry to ascertain whether these suspended accounts were involved in state-backed information operations (InfoOps) on Twitter. As presented in Figure 6.6, we identify a minority of mobilizers, accounting for a total of 599 accounts, who have participated in these campaigns. Intriguingly, NMYT users exhibit a higher level of involvement compared to MYT users, with 2.2% of MYT mobilizers implicated, as opposed to just 0.2% of NMYT mobilizers.

Overall, this analysis suggests that even though MYT mobilizers contravened Twitter policies, they were not actively engaged in state-backed orchestrated campaigns during the election.

User Interests

We now shift our focus to the content disseminated by MYT and NMYT mobilizers. Given that our dataset pertains to the U.S. 2020 Presidential election, we anticipate a prominent presence of political topics, especially related to the candidates' electoral campaigns and the contentious discussions surrounding alleged election fraud. We also place emphasis on the political inclinations of the users under examination and their potential associations with conspiratorial and fringe ideologies.

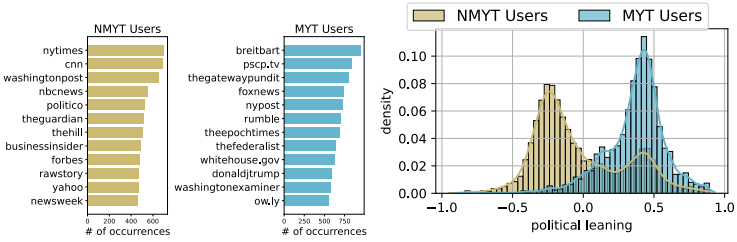
To probe the general interests of MYT and NMYT mobilizers, we undertake a comparative analysis of the hashtags present in their tweets and the descriptions of YT videos they share on Twitter. To this end, we employ SAGE [58] to identify the most distinctive hashtags and keywords in tweets and video descriptions, respectively. It is noteworthy that for MYT users, we only consider non-moderated videos, as the YouTube API does not provide metadata for moderated content. Table 6.1 presents the keywords and hashtags extracted by SAGE.

Our findings reveal that MYT users are supporters of former President Trump, as indicated by hashtags such as *#bestpresidentever45* and *#demonrats*. They also align with Trump's claims of voter fraud post-election, as evidenced by hashtags like *#krakenteam* and *#trumpwon*. In contrast, NMYT mobilizers express explicit disdain for Trump, with hashtags like *#trumpvirus* and *#traitorinchief*. However, their political orientation is not as unequivocal as that of MYT users, with *#gojoe* being the sole pro-Biden hashtag among NMYT's top-50 hashtags.

Overall, the keywords extracted from YT video descriptions correspond with the Twitter hashtags of both groups. However, these keywords do not convey positive or negative sentiments but generally refer to individuals or groups openly expressing their political preferences. For example, NMYT Mobilizers shared several videos related to the *Lincoln Project*, which includes an ad featuring *Barkhuff Dan*, who explicitly criticizes Trump. It is worth noting that the Lincoln Project is comprised of Republicans opposing Trump, supporting the idea that NMYT users are not exclusively Biden supporters. On the contrary, MYT mobilizers demonstrate a clear backing for Trump, with shared videos mentioning *Christina Bobb*, who was closely associated with Trump's legal team during the election result

Table 6.1. Most shared hashtags and YT video keywords by NMYT and MYT mobilizers

	NMYT Mobilizers	MYT Mobilizers
Twitter hashtags	#putinspuppet #trumpvirus #resignnowtrump #trumplies #traitorinchief #gojoe #trumpkillus #weirdotrump	#krakenteam #chinebitchbiden #demonrats #evidenceoffraud #bestpresidentever45 #arrestfauci #trumpwon #trumppatriots
YouTube keywords	barkhuff dan bernie sander lincoln rainbow loyalty cnn incompetence	rsbn bobulinski censored christina bobb fitton spoiled tucker carlson



(a) **(b)**

Figure 6.7. (a) The news outlet shared by each group of mobilizers (we omit the *.com* extension for brevity) ; (b) the distribution of the political leaning within the two groups of mobilizers

dispute, and *Tucker Carlson*, often likened to Trump's successor[1].

To further dissect users' political orientations, we utilize MediaBias-FactCheck's political leaning scores for various news outlets. Following previous research [67, 206], we gauge user political orientation by averaging the scores of the domains they share on Twitter during the observation period. Figures 6.7a and 6.7b display the top-10 domains shared by YT mobilizers and the political leaning distribution of MYT and NMYT users.

The former group encompasses several far-right users, with a median political leaning score of 0.47, who predominantly share news from *breitbart.com* and *thegatewaypundit.com*, known for propagating conspiracy theories and publishing extremely conservative content [127, 177]. In contrast, the latter group comprises users with less extreme and more liberal political leanings. However, the political leaning distribution of the NMYT group exhibits a bimodal pattern, with the larger mode at -0.27 and the smaller one at 0.41. A subset of these mobilizers demonstrates an extremely conservative ideology. This observation is further substantiated by examining the top-10 domains of MYT mobilizers, which include *foxnews.com* (frequently shared by the NMYT group) and *forbes.com*, a center-right news outlet.

Findings and Remarks In response to *RQ2*, we found that MYT users tend to passively retweet what they see on Twitter rather than actively posting original tweets or replies. In addition, they are usually suspended on Twitter but are not involved in information operations. Finally, when assessing the (political) interests, we found that MYT are far-right supporters and backed Trump during the 2020 US election, while the political leaning of NMYT users is less extreme and more diverse.

6.4.3 Engagement towards mobilizers of moderated YouTube videos (RQ3)

To address RQ3, we delve into the interaction dynamics exhibited by MYT and NMYT mobilizers, scrutinizing the intra- and intergroup retweets they engage in. We also consider the larger category of *other*

[1]https://www.theguardian.com/media/2020/jul/12/tucker-carlson-trump-fox-news-republicans

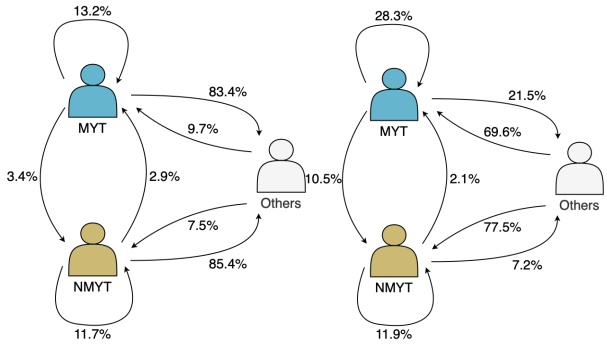
users, comprising 750,000 Twitter accounts that shared at least one YT video and do not fit into the YT mobilizer classification.

As the group sizes differ, a direct comparison of absolute numbers for intragroup and intergroup retweets is unwarranted. To address this, we normalize the interactions by source (i.e., the total number of retweets each group generates, see Figure 6.8a) and by target (i.e., the total number of retweets each group receives, see Figure 6.8b).

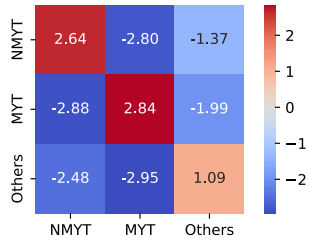
We observe that both MYT and NMYT mobilizers exhibit similar proportions of intragroup retweets, with figures of 13.2% and 11.7%, respectively, and intergroup retweets, with 3.4% and 2.9%, respectively. However, Figure 6.8b unveils a distinction in their engagement when it comes to the proportion of retweets they receive. Specifically, 28.3% of retweets received by MYT users originate from within their group, with only 2.1% coming from NMYT mobilizers. In contrast, NMYT mobilizers tend to retweet both groups at nearly the same rate, with 11.9% for NMYT to NMYT and 10.5% for MYT to NMYT. This suggests that MYT users are more inclined to retweet and be retweeted within their group than across groups. Additionally, users in the *others* category exhibit a nearly equal tendency to retweet MYT and NMYT mobilizers. This underscores that the level of user activity is not a distinguishing factor in characterizing interactions between NMYT users and the rest.

To bolster our findings, we compare the observed number of retweets with a null model that assumes interactions occur randomly. Specifically, we permute the users' group assignments (i.e., NMYT, MYT, and Others) and calculate the mean and standard deviation of intergroup interactions for 100 iterations. Subsequently, we compute z-scores to compare observed retweets with the expected number of retweets from the null model. Figure 6.8c illustrates that the observed retweet patterns among mobilizers align with the principle of homophily. In particular, both mobilizer groups exhibit a higher number of intragroup retweets and a lower number of intergroup retweets than expected by chance, with $z > 2.5$ and $z < -1.5$, respectively.

Findings and Remarks As for *RQ3*, we found that the MYT and NMYT groups exhibit strong group cohesion and are equally engaged by the Twitter audience. However, MYT users are not reciprocated by NMYT



(a) (b)



(c)

Figure 6.8. Interaction patterns enacted by NMYT and MYT accounts: (a) Proportion of interactions between YouTube mobilizers normalized by the source; (b) Proportion of interactions between YouTube mobilizers normalized by the destination; (c) Z-scores of observed retweets between YouTube mobilizers (p-value < 0.01)

users.

6.5 Discussion

6.5.1 Contributions

In this chapter, we studied the Twitter discussion around (video) content that is deemed harmful on YouTube. Leveraging an unprecedented large-scale dataset of 600M tweets shared by more than 7.5M users, we

discovered an unexpectedly high number of *moderated* YT videos shared on Twitter during the 2020 US election. Overall, moderated videos were shared more than nonmoderated ones and received far more attention than content from *fringe* social media platforms. Moving beyond previous work, we investigated the characteristics of the Twitter users responsible for sharing both moderated and non-moderated YT videos. On the one hand, we found that users sharing moderated content tend to passively retweet what they see on Twitter rather than actively posting original tweets or replies. On the other hand, the mobilizers of non-moderated videos actively share YT videos in their original tweets. Overall, most of the users were regular Twitter accounts rather than bots or state-sponsored actors, and, even if we did not find any involvement in information operations, Twitter suspended more than half of the moderated video mobilizers. Furthermore, we found that the mobilizers of moderated YT videos are far-right supporters and sustained Trump during the 2020 US election. By contrast, the political preference of the mobilizers of non-moderated YT videos is more diverse since users in this group range from Biden supporters to other Republican representatives who did not endorse Trump's political campaign. Finally, we studied the interactions between the mobilizers of moderated and non-moderated videos and discovered that both groups exhibit strong group cohesion and are engaged similarly to the general Twitter audience.

6.5.2 Limitations

There are a number of limitations to our study. First, neither Twitter nor YouTube provides any additional information on account suspension and video moderation, and the timing of their interventions is also unknown. Therefore, there is no guarantee that YT videos were still online when reshared through retweets on Twitter, but we can confidently assume they were not moderated yet when shared in an original tweet. Furthermore, we acknowledge that our analyzes, as in several previous works [148, 125], could be biased towards moderated YT content that includes not only videos that violate YouTube policies but also those removed by their publishers for any reason. Second, we overlooked the YT channels shared on Twitter to safeguard our analysis from Twitter users who just advertise their own (or others) YouTube channel [5]. However, this choice might prevent us from considering another potential source of harmful YT content

on Twitter. Third, the partition strategy to define the two groups of mobilizers is quite conservative, since we considered Twitter users who *never share* moderated videos and those who *mostly share* moderated videos.

6.5.3 Conclusions and Future Works

Our study has two major takeaways: first, moderated YT videos are widely shared on Twitter, and users who (passively) share those endorse extreme and conspiratorial ideas. From a broader perspective, we have shown how harmful content originating in a *source* platform significantly pollutes discussion on a *target* platform. Although more research is still needed, we conjecture that sharing information about the interventions taken would improve our understanding of cross-platform harmful content diffusion and benefit all entities within the information ecosystem. For instance, in the YouTube-Twitter cross-posting scenario considered in this paper, YouTube moderation activity can benefit both parties of the cooperation: on the one hand, Twitter has the opportunity to (early-)detect intra-platform harmful activities; on the other hand, YouTube can further improve its moderation based on the cross-platform signals tied with harmful YT content diffusion on Twitter.

Second, the mobilizers of the moderated YT videos appeared to be regular Twitter users who do not necessarily share content from *fringe* platforms. This suggests that cross-posting (harmful) cross-platform content is participatory [153] and research in this field should not only target bots and trolls but instead consider the role of online crowds and more complex social structures on different social media platforms.

Future work might build upon our findings to design algorithms to automatically identify or predict whether a YT video will be moderated based on the engagement it receives on Twitter, as well as to detect early signals of radicalization. In addition, we aim to investigate whether our results generalize to other topics beyond political elections or other highly-moderated social media (e.g. Facebook, Instagram).

Chapter 7

Epilogue

Disinformation, while not new, has become a pressing concern in the era of the Internet and social media. In this digital age, with easy access to social media channels and growing distrust in traditional media, misleading narratives find fertile ground to flourish.

To tackle this challenge, the field has seen a shift towards knowledge-informed disinformation mining. This approach understands disinformation as more than just false information - it considers the intricate interplay of human psychology, societal dynamics, and emotional triggers. Consequently, the research community has increasingly focused on comprehending the intricate role of each one of these aspects by studying (dis-)information propagation and users behavior on social media platforms.

A significant portion of these efforts has focused on rapidly identifying harmful content using various techniques. However, the vast amount of data on social media makes it challenging to develop a comprehensive automated system for detecting any problematic content. As a result, many platforms still depend on human fact-checkers and crowd-sourced efforts to identify and address harmful content.

In parallel, some studies has aimed to uncover the various players involved in spreading harmful content on digital platforms. These investigations have unveiled the pivotal roles played by both human and algorithmic factors, while shedding light on the deployment of malicious agents such as bots, cyborgs, and trolls, which are employed to manipulate and influence public opinion.

Overall, much remains unknown regarding the susceptibility of individuals, institutions, and broader society to the pernicious effects of online disinformation and its real-world repercussions. There seems to be a noticeable gap between crafting state-of-the-art detection systems and truly understanding the intricate dynamics of disinformation spread online. Recognizing this, the research community is increasingly championing interdisciplinary approaches, seeking to tackle the issue from a multitude of angles.

7.1 Summary of the Contributions

In drawing the curtains on this research journey, it becomes essential to reiterate and encapsulate the contributions made through this thesis. Throughout this exploration, the prevailing theme has been the holistic understanding and management of the challenge of disinformation, especially in the age of digitization. By adeptly fusing insights from artificial intelligence with knowledge-driven techniques, this research has illuminated how diverse forms of contextual knowledge can enrich our comprehension of online disinformation and bolster the creation of more robust detection and counteraction mechanisms.

A significant hallmark of this research was the exploration of leveraging pre-verified information to enhance the fact-checking process. Recognizing the exponential growth of online content, the inherent challenges of comprehensive verification, and the recurrent nature of specific falsehoods, we proposed an innovative AI-based multimodal information retrieval system. This system not only achieved state-of-the-art performance on benchmark datasets but also demonstrated its efficacy in real-world scenarios, exemplified by a case-study on the Ukraine-Russia conflict. This underscored the merits of not only identifying but also preemptively curating factual content in a sea of information.

The dynamic evolution of online content, underscored by the rising prominence of Internet memes and short videos as mediums of expression, constituted another primary focus of this research. In response to the challenge of detecting harmful memes, we introduced KERMIT, an innovative approach that seamlessly fuses meme content with real-world knowledge, represented through a dynamic knowledge graph. KERMIT consistently

exhibited robust efficacy in discerning harmful memes and assessing the logical coherence within each meme.

Pivoting to the deeper cognitive and emotional layers of disinformation, the research ventured into the world of sentiment, stance, and topic detection. Through the avant-garde multi-task learning framework, the interlaced dynamics of various facets of disinformation were unraveled. These insights transcended mere computational triumphs; they illuminated the nuanced tapestry of disinformation, where emotions, stances, and topics converge to influence perceptions. Furthermore, they set the foundation for the development of more proficient disinformation detection systems.

Lastly, in recognizing that the challenge of disinformation is not confined to silos but is an intricate web across platforms, the research made strides in the realm of cross-platform moderation. The insights gleaned from Twitter discussions linked to harmful YouTube content brought to light the symbiotic nature of online harm, emphasizing the pressing need for a unified front in moderation strategies.

To conclude, this thesis, while firmly rooted in academic rigor, serves as a lighthouse for the broader digital realm. It showcases that the battle against disinformation is not just about advanced algorithms but also about understanding human behavior, community culture, and the interplay of various digital platforms. The findings and methodologies elucidated within provide a roadmap not only for future researchers but also for technologists, platform designers, and policymakers, accentuating the collaborative ethos required to ensure the integrity of our digital information landscapes.

7.2 Outlook

The findings presented in this dissertation underscore the advantages of incorporating external and contextual knowledge in the field of disinformation mining. However, while these contributions address challenges that are somewhat distinct from mainstream disinformation issues, they do not advocate for a structured approach that standardizes where, which, and how such knowledge can be most effectively utilized. Our analysis suggests that assimilating cultural and common-sense knowledge enhances the nuanced understanding of deleterious content, facilitating its detection.

Furthermore, insights into emotional appeals and user susceptibilities can effectively neutralize disinformation campaigns. Nevertheless, we concede the inherent limitations of exclusively relying on static knowledge sources. The adaptability of these methodologies in environments with fluid and dynamically evolving knowledge remains an area for further exploration.

From the perspective of characterizing disinformation on social media platforms, our study primarily examined mainstream platforms like Twitter, Facebook, and YouTube, reflecting the focus of much existing research. Recent academic endeavors are diversifying, probing niche platforms like Bitchute and Parler, as well as decentralized ones like Mastodon and Minds. One of the overarching challenges impeding progress in this direction pertains to data accessibility, with specific impediments arising from limited visibility into moderation actions. This concern is magnifying, as even mainstream platforms are increasingly restricting data access, exemplified by Twitter's new, expensive API usage plans.

In conclusion, this thesis fervently champions the responsible application of artificial intelligence tools as powerful weapons to combat disinformation. However, emerging developments, particularly in foundational models, spotlight the potential hazards associated with such tools. Challenges encompass content veracity, the fabrication of convincing counterfeit profiles, systemic biases, and the synthesis of misleading multimedia that targets public figures. Projecting forward, we foresee a surge in research efforts addressing these concerns.

Bibliography

[1] Athira A.B., S.D. Madhu Kumar, and Anu Mary Chacko. A systematic survey on explainable ai applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122:106087, 2023.

[2] Athira A.B., S.D. Madhu Kumar, and Anu Mary Chacko. A systematic survey on explainable ai applied to fake news detection. *Eng. Appl. Artif. Intell.*, 122(C), jun 2023.

[3] Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510, 2007.

[4] Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. Voterfraud2020: a multi-modal dataset of election fraud claims on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):901–912, May 2021.

[5] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. Who watches (and shares) what on youtube? and when? using twitter to understand youtube viewership. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 593–602, New York, NY, USA, 2014. Association for Computing Machinery.

[6] Ahmed Abouzeid, Ole-Christoffer Granmo, Christian Webersik, and Morten Goodwin. Learning automata-based misinformation mitigation via hawkes processes. *Information Systems Frontiers*, pages 1–20, 2021.

[7] Ahmed Abouzeid, Ole-Christoffer Granmo, Christian Webersik, and Morten Goodwin. Socially fair mitigation of misinformation on social networks via constraint stochastic optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11801–11809, 2022.

[8] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[9] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[10] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, WebSci '21, page 187–195, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[12] Liesbeth Allein and Marie-Francine Moens. Checkworthiness in automatic claim detection models: Definitions and analysis of datasets. *CoRR*, abs/2008.08854, 2020.

[13] Blake E Ashforth and Fred Mael. Social identity theory and the organization. *Academy of management review*, 14(1):20–39, 1989.

[14] Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France, May 2020. European Language Resources Association.

[15] Isabelle Augenstein. Towards explainable fact checking. *arXiv preprint arXiv:2108.10274*, 2021.

[16] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *CoRR*, abs/1802.04291, 2018.

[17] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 258–265. IEEE, 2018.

[18] Hui Bai, Jan Voelkel, Johannes Eichstaedt, and Robb Willer. Artificial intelligence can persuade humans on political issues. 2023.

[19] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[20] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[21] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[22] Shai Ben-David and R. Borbely. Exploiting task relatedness for multiple task learning. *Lecture Notes in Computer Science*, 2777, 08 2002.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[24] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. Directive explanations for monitoring the risk of diabetes onset: Introducing directive data-centric explanations and combinations to support what-if explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 204–219, New York, NY, USA, 2023. Association for Computing Machinery.

[25] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[26] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.

[27] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz

Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[28] Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. Youtube recommendations and effects on sharing across online social platforms. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[29] Grégoire Burel, Tracie Farrell, and Harith Alani. Demographics and topics impact on the co-spread of covid-19 misinformation and fact-checks on twitter. *Information Processing Management*, 58(6):102732, 2021.

[30] Yinqiong Cai, Yixing Fan, Jiafeng Guo, Fei Sun, Ruqing Zhang, and Xueqi Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *CoRR*, abs/2103.04831, 2021.

[31] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[32] Tanmoy Chakraborty, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. Information retrieval algorithms and neural ranking models to detect previously fact-checked information. *Neurocomputing*, 557:126680, 2023.

[33] Chonghao Chen, Fei Cai, Xuejun Hu, Wanyu Chen, and Honghui Chen. Hhgn: A hierarchical reasoning-based heterogeneous graph neural network for fact verification. *Information Processing & Management*, 58(5):102659, 2021.

[34] Chonghao Chen, Fei Cai, Xuejun Hu, Jianming Zheng, Yanxiang Ling, and Honghui Chen. An entity-graph based reasoning method for fact verification. *Information Processing & Management*, 58(3):102472, 2021.

[35] Emily Chen, Ashok Deb, and Emilio Ferrara. # election2020: the first public twitter dataset on the 2020 us presidential election. *Journal of Computational Social Science*, pages 1–18, 2021.

[36] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[37] D. Cheriton. From doc2query to doctttttquery. 2019.

[38] Matthew Childs, Cody Buntain, Milo Z. Trujillo, and Benjamin D. Horne. Characterizing youtube and bitchute content and mobilizers during u.s. election fraud discussions on twitter. In *14th ACM Web Science Conference*

2022, WebSci '22, page 250–259, New York, NY, USA, 2022. Association for Computing Machinery.

[39] Sera Choi, Ashley A Anderson, Shelby Cagle, Marilee Long, and Nicole Kelp. Scientists' deficit perception of the public impedes their behavioral intentions to correct misinformation. *Plos one*, 18(8):e0287870, 2023.

[40] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Valensise, Emanuele Brugnoli, Ana Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10, 10 2020.

[41] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10, 2020.

[42] Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. The effect of people recommenders on echo chambers and polarization. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):90–101, May 2022.

[43] Dennis Collaris, Leo M. Vink, and Jarke J. van Wijk. Instance-level explanations for fraud detection: A case study. *CoRR*, abs/1806.07129, 2018.

[44] Antoine-Nicholas Condorcet. *Outlines of an Historical View of the Progress of the Human Mind*. Lulu. com, 2009.

[45] Michael Crawshaw. Multi-task learning with deep neural networks: A survey, 2020.

[46] Lei Cui, Furu Wei, and Ming Zhou. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–413, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[47] Zhuyun Dai and Jamie Callan. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1533–1536, New York, NY, USA, 2020. Association for Computing Machinery.

[48] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 126–134, New York, NY, USA, 2018. Association for Computing Machinery.

[49] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 126–134, New York, NY, USA, 2018. Association for Computing Machinery.

[50] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[51] Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629, 1955.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[53] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. *CoRR*, abs/2105.09284, 2021.

[54] Yasan Ding, Bin Guo, Yan Liu, Yunji Liang, Haocheng Shen, and Zhiwen Yu. Metadetector: Meta event knowledge transfer for fake news detection. *ACM Trans. Intell. Syst. Technol.*, 13(6), sep 2022.

[55] Evelyn Douek. The rise of content cartels. *Knight First Amendment Institute at Columbia*, 2020.

[56] Elizabeth Dubois, Sara Minaeian, Ariane Paquet-Labelle, and Simon Beaudry. Who to trust on social media: How opinion leaders and seekers avoid disinformation and echo chambers. *Social media+ society*, 6(2):2056305120913993, 2020.

[57] 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.

[58] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1041–1048, Madison, WI, USA, 2011. Omnipress.

[59] Martin J Eppler and Jeanne Mengis. The concept of information overload-a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004) the information society: An international journal, 20 (5), 2004, pp. 1–20. *Kommunikationsmanagement im Wandel: Beiträge aus 10 Jahren= mcminstitute*, pages 271–305, 2008.

[60] Fatima Ezzeddine, Luca Luceri, Omran Ayoub, Ihab Sbeity, Gianluca Nogara, Emilio Ferrara, and Silvia Giordano. Characterizing and detecting state-sponsored troll activity on social media, 2023.

[61] Francesco Fabbri, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. Rewiring what-to-watch-next recommendations to reduce radicalization pathways. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2719–2728, New York, NY, USA, 2022. Association for Computing Machinery.

[62] J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci. Automatic knowledge extraction from documents. *IBM Journal of Research and Development*, 56(3.4):5:1–5:10, 2012.

[63] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1097–1106. PMLR, 06–11 Aug 2017.

[64] Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. Knowledge does not protect against illusory truth. *Journal of experimental psychology: general*, 144(5):993, 2015.

[65] Zhengcong Fei. Memory-augmented image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1317–1324, May 2021.

[66] Emilio Ferrara. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models, 2023.

[67] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. Characterizing social media manipulation in the 2020 us presidential election. *First Monday*, 2020.

[68] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, July 2022. Association for Computational Linguistics.

[69] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), jul 2018.

[70] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1797–1806. ACM, 2017.

[71] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, November 1987.

[72] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1), jan 2018.

[73] Valerio La Gatta, Luca Luceri, Francesco Fabbri, and Emilio Ferrara. The interconnected nature of online harm and moderation: Investigating the cross-platform spread of harmful content between youtube and twitter. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA, 2023. Association for Computing Machinery.

[74] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlí. Covid-19 sentiment analysis based on tweets. *IEEE Intelligent Systems*, 38(3):51–55, 2023.

[75] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. An emotional analysis of false information in social media and news articles. *ACM Trans. Internet Technol.*, 20(2), apr 2020.

[76] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 877–880, New York, NY, USA, 2019. Association for Computing Machinery.

[77] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[78] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press, 2018.

[79] Tamar Ginossar, Iain J. Cruickshank, Elena Zheleva, Jason Sulskis, and Tanya Berger-Wolf. Cross-platform spread: vaccine-related content,

sources, and conspiracy theories in youtube videos shared in early twitter covid-19 conversations. *Human Vaccines & Immunotherapeutics*, 18(1):1–13, 2022. PMID: 35061560.

[80] Laura Glitsos and James Hall. The pepe the frog meme: an examination of social, political, and cultural implications through the tradition of the darwinian absurd. *Journal for Cultural Research*, 23(4):381–395, 2019.

[81] Yevgeniy Golovchenko, Cody Buntain, Gregory Eady, Megan A. Brown, and Joshua A. Tucker. Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 u.s. presidential election. *The International Journal of Press/Politics*, 25(3):357–389, 2020.

[82] Genevieve Gorrell, Ahmet Aker, Kalina Bontcheva, Elena Kochkina, Maria Liakata, Arkaitz Zubiaga, and Leon Derczynski. SemEval-2019 task 7: RumourEval, Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[83] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision, 2022.

[84] Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. Learning to transduce with unbounded memory. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 1828–1836, Cambridge, MA, USA, 2015. MIT Press.

[85] Hao Guo, Yijun Liu, Jianmin Wang, and Yonghui Wu. Knowledge graph-based multi-label classification with label embeddings and label dependency. *Knowledge-Based Systems*, 198:105965, 2020.

[86] Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. Matchzoo: A learning, practicing, and developing system for neural text matching. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1297–1300, New York, NY, USA, 2019. ACM.

[87] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *CoRR*, abs/1903.06902, 2019.

[88] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[89] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[90] Michael Hameleers, Thomas E. Powell, Toni G.L.A. Van Der Meer, and Lieke Bos. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301, 2020.

[91] Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. Learning-to-rank with BERT in tf-ranking. *CoRR*, abs/2004.08476, 2020.

[92] Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. Happenstance: Utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):327–338, Jun. 2023.

[93] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. Happenstance: Utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. *arXiv preprint arXiv:2205.14484*, 2022.

[94] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[95] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3651–3655, New York, NY, USA, 2022. Association for Computing Machinery.

[96] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics.

[97] Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus*, 2(3), 03 2023. pgad035.

[98] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2333–2338, New York, NY, USA, 2013. Association for Computing Machinery.

[99] Benjamin WK Hung, Anura P Jayasumana, and Vidarshana W Bandara. Insight: A system to detect violent extremist radicalization trajectories in dynamic graphs. *Data & Knowledge Engineering*, 118:52–70, 2018.

[100] Roland Imhoff, Felix Zimmer, Olivier Klein, João HC António, Maria Babinska, Adrian Bangerter, Michal Bilewicz, Nebojša Blanuša, Kosta Bovan, Rumena Bužarovska, et al. Conspiracy mentality and political orientation across 26 countries. *Nature human behaviour*, 6(3):392–403, 2022.

[101] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan. Automatic detection of entity-manipulated text using factual knowledge. In *Association for Computational Linguistics*, pages 86–93, 2022.

[102] Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 596–605. IEEE, 2022.

[103] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[104] Julie Jiang, Xiang Ren, and Emilio Ferrara. Retweet-bert: Political leaning detection using language features and information diffusion on social networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):459–469, Jun. 2023.

[105] Shan Jiang and Christo Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

[106] Quanliang Jing, Di Yao, Xinxin Fan, Baoli Wang, Haining Tan, Xiangpeng Bu, and Jingping Bi. Transfake: Multi-task transformer for multimodal enhanced fake news detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[107] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734, 2017.

[108] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[109] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[110] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery.

[111] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA, 2019. Association for Computing Machinery.

[112] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950, 2019.

[113] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. The hateful memes challenge: Competition report. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR, 06–12 Dec 2021.

[114] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[115] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread

of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 324–332, 2018.

[116] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021.

[117] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[118] 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.

[119] Bisera Kostadinovska-Stojchevska and Elena Shalevska. Internet memes and their socio-linguistic features. *European journal of literature, language and linguistics studies*, 2(4), 2018.

[120] 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.

[121] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120, 2023.

[122] Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, 21(3):103–156, 2020.

[123] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[124] Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. Misinformation detection using multitask learning with mutual learning for novelty

detection and emotion recognition. *Information Processing Management*, 58(5):102631, 2021.

[125] Maram Kurdi, Nuha Albadi, and Shivakant Mishra. "video unavailable": Analysis and prediction of deleted and moderated youtube videos. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 166–173, 2020.

[126] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[127] Valerio La Gatta, Chiyu Wei, Luca Luceri, Francesco Pierri, and Emilio Ferrara. Retrieving false claims on twitter during the russia-ukraine conflict. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1317–1323, New York, NY, USA, 2023. Association for Computing Machinery.

[128] Orestis Lampridis, Dimitra Karanatsiou, and Athena Vakali. Manifesto: A human-centric explainable approach for fake news spreaders detection. *Computing*, 104(4):717–739, apr 2022.

[129] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[130] Mark Ledwich, Anna Zaitsev, and Anton Laukemper. Radical bubbles on youtube? revisiting algorithmic extremism with personalised recommendations. *First Monday*, 27(12), Dec. 2022.

[131] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, page 535–541, Cambridge, MA, USA, 2000. MIT Press.

[132] Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online, June 2021. Association for Computational Linguistics.

[133] Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369, 2017.

[134] Dongyue Li, Huy Nguyen, and Hongyang Ryan Zhang. Identification of negative transfers in multitask learning using surrogate models. *Transactions on Machine Learning Research*, 2023. Featured Certification.

[135] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media, 2023.

[136] Liunian Harold Li, Mark Yatskar, D Yin, CJ Hsieh, and KW Chang. Visualbert: A simple and performant baseline for vision and language. arxiv 2019. *arXiv preprint arXiv:1908.03557*.

[137] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

[138] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020.

[139] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2019.

[140] Qing Liao, Heyan Chai, Hao Han, Xiang Zhang, Xuan Wang, Wen Xia, and Ye Ding. An integrated multi-task model for fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 01 2021.

[141] Chi-Chin Lin, Yi-Ching Huang, and Jane Yung-jen Hsu. Crowdsourced explanations for humorous internet memes based on linguistic theories. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2(1):143–150, Sep. 2014.

[142] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *CoRR*, abs/2106.04554, 2021.

[143] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[144] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.

[145] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[146] Yue Liu, Chakkrit Tantithamthavorn, Li Li, and Yepang Liu. Explainable ai for android malware detection: Towards understanding why the models perform so well? In *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*, pages 169–180, 2022.

[147] Kuan-Chieh Lo, Shih-Chieh Dai, Aiping Xiong, Jing Jiang, and Lun-Wei Ku. All the wiser: Fake news intervention using user reading preferences. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 1069–1072, New York, NY, USA, 2021. Association for Computing Machinery.

[148] Marcelo Sartori Locatelli, Josemar Caetano, Wagner Meira Jr., and Virgilio Almeida. Characterizing vaccination movements on youtube in the united states and brazil. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 80–90, New York, NY, USA, 2022. Association for Computing Machinery.

[149] 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.

[150] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019.

[151] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[152] 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.

[153] Luca Luceri, Stefano Cresci, and Silvia Giordano. Social media against society. *The Internet and the 2020 Campaign*, 2021.

[154] Luca Luceri, Silvia Giordano, and Emilio Ferrara. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 417–427, 2020.

[155] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[156] Wei Lyu and George L Wehby. Community use of face masks and covid-19: Evidence from a natural experiment of state mandates in the us: Study examines impact on covid-19 growth rates associated with state government mandates requiring face mask use in public. *Health affairs*, 39(8):1419–1425, 2020.

[157] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. Visual question answering with memory-augmented networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6975–6984, 2018.

[158] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 585–593, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[159] Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. Did i see it before? detecting previously-checked claims over twitter. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 367–381, Berlin, Heidelberg, 2022. Springer-Verlag.

[160] Lance E Mason, Dan Krutka, and Jeremy Stoddard. Media literacy, democracy, and the challenge of fake news. *Journal of Media Literacy Education*, 10(2):1–10, 2018.

[161] Arak Mathai and Panagis Moschopoulos. The distribution of the standard f-ratio in one-way anova with multinomially distributed cell sizes. *International Journal of Mathematical and Statistical Sciences*, 4, 01 1995.

[162] Michele Mazza, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Investigating the difference between trolls, social bots, and humans on twitter. *Computer Communications*, 196:23–36, 2022.

[163] Lee McIntyre. *Post-truth*. MIt Press, 2018.

[164] Youssef A. Mejjati, Darren Cosker, and Kwang In Kim. Multi-task learning by maximizing statistical dependence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3465–3473, 2018.

[165] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online, November 2020. Association for Computational Linguistics.

[166] Donald Metzler and W. Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 472–479, New York, NY, USA, 2005. Association for Computing Machinery.

[167] Alessandro Miani, Thomas Hills, and Adrian Bangerter. Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, 8(43):eabq3668, 2022.

[168] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[169] Adam Mills, Christine Pitt, and Sarah Lord Ferguson. The relationship between fake news and advertising: Brand management in the era of programmatic advertising and prolific falsehood. *Journal of Advertising Research*, 59:3–8, 03 2019.

[170] Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. Predicting readers' sarcasm understandability by modeling gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[171] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1291–1299, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[172] Tamar Mitts, Nilima Pisharody, and Jacob Shapiro. Removal of anti-vaccine content impacts social media discourse. In *14th ACM Web Science Conference 2022*, WebSci '22, page 319–326, New York, NY, USA, 2022. Association for Computing Machinery.

[173] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Memory graph networks for explainable memory-grounded question answering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 728–736, Hong Kong, China, November 2019. Association for Computational Linguistics.

[174] Patricia Moravec, Randall Minas, and Alan R Dennis. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business research paper*, (18-87), 2018.

[175] Vincenzo Moscato, Giuseppe Napolano, Marco Postiglione, and Giancarlo Sperlì. Multi-task learning for few-shot biomedical relation extraction. *Artificial Intelligence Review*, pages 1–21, 2023.

[176] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[177] Goran Muric, Yusong Wu, and Emilio Ferrara. Covid-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*, 7(11):e30642, 2021.

[178] Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James Glass. FAKTA: An automatic end-to-end fact checking system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 78–83, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[179] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *CoRR*, abs/1911.03854, 2019.

[180] 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. Survey Track.

[181] Preslav Nakov and Giovanni Da San Martino. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19, Online, November 2020. Association for Computational Linguistics.

[182] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mücahid Kutlu, and Yavuz Selim Kartal. Overview of the CLEF-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. *CoRR*, abs/2109.12987, 2021.

[183] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.

[184] Christoph G Nguyen, Sabrina J Mayer, and Susanne Veit. The impact of emotions on polarization. anger polarizes attitudes towards vaccine mandates and increases affective polarization. *Research & Politics*, 9(3):20531680221116571, 2022.

[185] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022.

[186] Bo Ni, Zhichun Guo, Jianing Li, and Meng Jiang. Improving generalizability of fake news detection methods using propensity score matching. *CoRR*, abs/2002.00838, 2020.

[187] Ingram Niblock, Jacob Wallis, and Albert Zhang. Understanding global disinformation and information operations. 2022.

[188] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[189] Ping Nie, Yuyu Zhang, Xiubo Geng, Arun Ramamurthy, Le Song, and Daxin Jiang. DC-BERT: decoupling question and document for efficient contextual encoding. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and*

development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 1829–1832. ACM, 2020.

[190] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6859–6866, 07 2019.

[191] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Coordinated behavior on social media in 2019 uk general election. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):443–454, May 2021.

[192] Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. In *14th ACM Web Science Conference 2022*, WebSci '22, page 348–358, New York, NY, USA, 2022. Association for Computing Machinery.

[193] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424, 2019.

[194] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *CoRR*, abs/1904.08375, 2019.

[195] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[196] Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018.

[197] Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. Challenges and opportunities in information manipulation detection: An examination of wartime russian media. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, 2022.

[198] Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. Challenges and opportunities in information manipulation detection: An examination of wartime russian media. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[199] Andrew Patel and Jason Sattler. Creatively malicious prompt engineering. *WithSecure Intelligence*, 2023.

[200] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019. The Cognitive Science of Political Thought.

[201] Denis Peskoff and Brandon M Stewart. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, 2023.

[202] Fabio Petroni, Tim Rocktäschel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *CoRR*, abs/1909.01066, 2019.

[203] Francesco Pierri, Matthew R DeVerna, Kai-Cheng Yang, David Axelrod, John Bryden, and Filippo Menczer. One year of covid-19 vaccine misinformation on twitter: Longitudinal study. *Journal of Medical Internet Research*, 25:e42227, 2023.

[204] Francesco Pierri, Luca Luceri, and Emilio Ferrara. How does twitter account moderation work? dynamics of account creation and suspension during major geopolitical events. *arXiv preprint arXiv:2209.07614*, 2022.

[205] Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 65–74, New York, NY, USA, 2023. Association for Computing Machinery.

[206] Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 65–74, New York, NY, USA, 2023. Association for Computing Machinery.

[207] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[208] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and Q. M. Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4051–4070, 2023.

[209] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[210] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

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

[212] 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*, 21:1–67, 2020.

[213] Chahat Raj and Priyanka Meel. Arcnn framework for multimodal infodemic detection. *Neural Networks*, 146:36–68, 2022.

[214] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

[215] Steve Rathje, Jay J. Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118, 2021.

[216] Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[217] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics, 2019.

[218] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 131–141, New York, NY, USA, 2020. Association for Computing Machinery.

[219] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 131–141, New York, NY, USA, 2020. Association for Computing Machinery.

[220] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

[221] Gal Ron, Effi Levi, Odelia Oshri, and Shaul Shenhav. Factoring hate speech: A new annotation framework to study hate speech in social media. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 215–220, Toronto, Canada, July 2023. Association for Computational Linguistics.

[222] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 232–239, New York, NY, USA, 2017. Association for Computing Machinery.

[223] Anderson Rossanez, Julio Cesar Dos Reis, Ricardo da Silva Torres, and Hélène de Ribaupierre. Kgen: a knowledge graph generator from biomedical scientific literature. *BMC medical informatics and decision making*, 20(4):1–24, 2020.

[224] Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1325–1334, New York, NY, USA, 2020. Association for Computing Machinery.

[225] Federico Ruggeri, Marco Lippi, and Paolo Torroni. Membert: Injecting unstructured knowledge into BERT. *CoRR*, abs/2110.00125, 2021.

[226] Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. Spillover of antisocial behavior from fringe platforms: The unin-

tended consequences of community banning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):742–753, Jun. 2023.

[227] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

[228] Zeve Sanderson, Megan A Brown, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review*, 2021.

[229] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[230] Simone Santini. Evaluation vademecum for visual information system. In Minerva M. Yeung, Boon-Lock Yeo, and Charles A. Bouman, editors, *Storage and Retrieval for Media Databases 2000*, volume 3972, pages 132 – 143. International Society for Optics and Photonics, SPIE, 1999.

[231] Fernando P. Santos, Yphtach Lelkes, and Simon A. Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50):e2102141118, 2021.

[232] Matheus Schmitz, Goran Muric, and Keith Burghardt. Quantifying how hateful communities radicalize online users. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 139–146, 2022.

[233] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. If you have a reliable source, say something: Effects of correction comments on covid-19 misinformation. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):896–907, May 2022.

[234] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL-HLT 2022*, NAACL-HLT '22, Seattle, Washington, USA, 2022.

[235] 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.

[236] Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 186–195, 2021.

[237] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018.

[238] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multimodal media manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[239] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. Semeval-2020 task 8: Memotion analysis - the visuo-lingual metaphor! *CoRR*, abs/2008.03781, 2020.

[240] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3), apr 2019.

[241] Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States, July 2022. Association for Computational Linguistics.

[242] Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. Detecting and understanding harmful memes: A survey. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track.

[243] Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim?, 2023.

[244] Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5468–5481, Online, August 2021. Association for Computational Linguistics.

[245] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Finding streams in knowledge graphs to support fact checking. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 859–864, 2017.

[246] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA, 2019. Association for Computing Machinery.

[247] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[248] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, sep 2017.

[249] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[250] Michael Siering, Jascha-Alexander Koch, and Amit V Deokar. Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems*, 33(2):421–455, 2016.

[251] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565, 2021.

[252] Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online, June 2021. Association for Computational Linguistics.

[253] Molly J Simis, Haley Madden, Michael A Cacciatore, and Sara K Yeo. The lure of rationality: Why does the deficit model persist in science communication? *Public understanding of science*, 25(4):400–414, 2016.

[254] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020.

[255] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. *CoRR*, abs/2112.04482, 2021.

[256] Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1322–1331, May 2022.

[257] Amir Soleimani, Christof Monz, and Marcel Worring. Bert for evidence retrieval and claim verification. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 359–366, Cham, 2020. Springer International Publishing.

[258] Chenguang Song, Kai Shu, and Bin Wu. Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, 58(6):102712, 2021.

[259] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975, 2016.

[260] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. Visual question answering using deep learning: A survey and performance analysis. *CoRR*, abs/1909.01860, 2019.

[261] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? *CoRR*, abs/1905.07553, 2019.

[262] Digital Forensic Research Lab Stanford Internet Observatory, Center for an Informed Public. The long fuse: Misinformation and the 2020 election. Stanford Digital Repository: Election Integrity, 2021.

[263] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[264] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. *CoRR*, abs/2107.06912, 2021.

[265] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2440–2448, Cambridge, MA, USA, 2015. MIT Press.

[266] Ling Sun, Yuan Rao, Yuqian Lan, Bingcan Xia, and Yangyang Li. Hg-sl: Jointly learning of global and local user spreading behavior for fake news early detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5248–5256, Jun. 2023.

[267] Vishnuprasad Padinjaredath Suresh, Gianluca Nogara, Felipe Cardoso, Stefano Cresci, Silvia Giordano, and Luca Luceri. Tracking fringe and coordinated activity on twitter leading up to the us capitol attack, 2023.

[268] 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. *CoRR*, abs/1704.07506, 2017.

[269] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16, 1979.

[270] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics.

[271] 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 Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web – ISWC 2019*, pages 309–324, Cham, 2019. Springer International Publishing.

[272] Abhinav Kumar Thakur, Filip Ilievski, Hông Ân Sandlin, Zhivar Sourati, Luca Luceri, Riccardo Tommasini, and Alain Mermoud. Multimodal and explainable internet meme classification, 2023.

[273] 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.

[274] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019.

[275] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. I want to break free! recommending friends from outside the echo chamber. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 23–33, New York, NY, USA, 2021. Association for Computing Machinery.

[276] Hanghang Tong, Christos Faloutsos, and Jianyong Pan. Fast random walk with restart and its applications. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, pages 613–624. IEEE, 2006.

[277] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13(9):e0203958, 2018.

[278] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[279] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018*, pages 517–524, 2018.

[280] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693, 2020.

[281] Amir Vakili Tahami, Kamyar Ghajar, and Azadeh Shakery. Distilling knowledge for fast retrieval-based chat-bots. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2081–2084, New York, NY, USA, 2020. Association for Computing Machinery.

[282] Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online, April 2021. Association for Computational Linguistics.

[283] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *CoRR*, abs/2012.12975, 2020.

[284] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 275–284, New York, NY, USA, 2018. Association for Computing Machinery.

[285] 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.

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

[287] Emily K Vraga and Leticia Bode. Using expert sources to correct health misinformation in social media. *Science communication*, 39(5):621–645, 2017.

[288] Emily K Vraga and Leticia Bode. I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, 21(10):1337–1353, 2018.

[289] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014.

[290] Emily L. Wang, Luca Luceri, Francesco Pierri, and Emilio Ferrara. Identifying and characterizing behavioral classes of radicalization within the qanon conspiracy on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):890–901, Jun. 2023.

[291] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[292] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey, 2023.

[293] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery.

[294] Yijun Wang, Yuezhang Huang, and Yun Yang. Multi-modal transformer for fake news detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4090–4100, 2020.

[295] Yuping Wang, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini. A multi-platform analysis of political news discussion and sharing on web communities. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1481–1492, 2021.

[296] Andrew Ward, L Ross, E Reed, E Turiel, and T Brown. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, pages 103–135, 1997.

[297] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), nov 2010.

[298] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. 2015. Publisher Copyright: © 2015 International Conference on Learning Representations, ICLR. All rights reserved.; 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

[299] Tom Wilson and Kate Starbird. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1), 2020.

[300] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3), June 2008.

[301] Ke Wu, Song Yang, and Kenny Q. Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, pages 651–662, 2015.

[302] Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. 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 4644–4653, Hong Kong, China, November 2019. Association for Computational Linguistics.

[303] Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

[304] Yiping Xia, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. Disinformation, performed: self-presentation of a russian ira account on twitter. *Information, Communication & Society*, 22(11):1646–1664, 2019.

[305] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 55–64, New York, NY, USA, 2017. Association for Computing Machinery.

[306] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. Heterogeneous network representation learning: A unified framework with survey and benchmark. *TKDE*, 2020.

[307] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science*, 5(2):1511–1528, 2022.

[308] Kai-Cheng Yang and Filippo Menczer. Anatomy of an ai-powered malicious social botnet, 2023.

[309] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[310] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, sep 2023. Just Accepted.

[311] Zichao Yang, Diyi Yang, Chris Dyer, X. He, Alex Smola, and E. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, 2016.

[312] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:3208–3216, 05 2021.

[313] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6995–7004, June 2021.

[314] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1007–1014, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[315] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA, 2018. Association for Computing Machinery.

[316] 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.

[317] ChengXiang Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141, 2008.

[318] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 334–342, New York, NY, USA, 2001. Association for Computing Machinery.

[319] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network representation learning: A survey. *IEEE Transactions on Big Data*, 6(1):3–28, 2020.

[320] Qiang Zhang, Hongbin Huang, Shangsong Liang, Zaiqiao Meng, and Emine Yilmaz. Learning to detect few-shot-few-clue misinformation. *CoRR*, abs/2108.03805, 2021.

[321] Ting Zhang, Bang Liu, Di Niu, Kunfeng Lai, and Yu Xu. Multiresolution graph attention networks for relevance matching. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 933–942, New York, NY, USA, 2018. Association for Computing Machinery.

[322] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2023.

[323] Xiaoying Zhang, Shu Xie, Hongtao Liu, and Maosong Sun. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2018.

[324] Yuchen Zhang, Qiang Yang, and Ding Zhou. Multi-label classification via knowledge graph embeddings and soft-constrained label propagation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pages 3350–3356. AAAI Press, 2018.

[325] Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. Fake news early detection: A theory-driven model. *Digital Threats*, 1(2), jun 2020.

[326] Xinyi Zhou, Kai Shu, Vir V. Phoha, Huan Liu, and Reza Zafarani. "this is fake! shared it by mistake":assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3685–3694, New York, NY, USA, 2022. Association for Computing Machinery.

[327] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multimodal fake news detection. In Hady W. Lauw, Ee-Peng Lim, Raymond Chi-Wing Wong, Alexandros Ntoulas, See-Kiong Ng, and Sinno Jialin Pan, editors, *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 354–367. Springer, 2020. Publisher Copyright: © Springer Nature Switzerland AG 2020. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.; 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2020 ; Conference date: 11-05-2020 Through 14-05-2020.

[328] Xinyi Zhou and Reza Zafarani. Network-based fake news detection: A pattern-driven approach. *SIGKDD Explor. Newsl.*, 21(2):48–60, nov 2019.

[329] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5), sep 2020.

[330] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2072–2084. Curran Associates, Inc., 2021.

[331] Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, WWW '19, page 2472–2482, New York, NY, USA, 2019. Association for Computing Machinery.

[332] Ron Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *CoRR*, abs/2012.08290, 2020.

[333] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China.

[334] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science?, 2023.

[335] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11, 03 2016.

Author's publications

Journal Publications

J1 V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì. *Covid-19 sentiment analysis based on Tweets.* IEEE Intelligent Systems, 2023. DOI: 10.1109/MIS.2023.3239180

J2 T. Chakraborty, V. La Gatta, V. Moscato, G. Sperlì. *Information retrieval algorithms and neural ranking models to detect previously fact-checked information.* Neurocomputing, 2023. DOI: 10.1016/j.neucom.2023.126680

J3 A. Ferraro, A. Galli, V. La Gatta, M. Postiglione. *Benchmarking Open Source and Paid Services for Speech to Text: An Analysis of Quality and Input Variety.* Frontiers in Big Data, 2023. DOI: 10.3389/fdata.2023.1210559

J4 V. La Gatta, V. Moscato, M. Pennone, M. Postiglione, G. Sperlì. *Music Recommendation via Hypergraph Embedding.* IEEE Transactions on Neural Networks and Learning Systems, 2022. DOI: 10.1109/TNNLS.2022.3146968

J5 A. Barducci, S. Iannaccone, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, S. Zavota. *An end-to-end framework for information extraction from Italian resumes.* Expert Systems with Applications, 2022. DOI: 10.1016/j.eswa.2022.118487

J6 V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì. *CASTLE: Cluster-aided space transformation for local explanations.* Expert Systems with Applications, 2021. DOI: 10.1016/j.eswa.2021.115045

J7 V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì. *PASTLE: Pivot-aided space transformation for local explanations.* Pattern Recognition Letters, 2021. DOI: 10.1016/j.patrec.2021.05.018

J8 V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì. *An Epidemiological Neural Network Exploiting Dynamic Graph Structured Data Applied to the COVID-19 Outbreak*. IEEE Transactions on Big Data, 2020. DOI: 10.1109/TBDATA.2020.3032755

Conference Publications

C1 V. La Gatta, L. Luceri, F. Fabbri, E. Ferrara. *The Interconnected Nature of Online Harm and Moderation: Investigating the Cross-Platform Spread of Harmful Content between YouTube and Twitter*. 34th ACM International Conference on Hypertext and Social Media (HT2023). DOI: 10.1145/3603163.3609058

C2 V. La Gatta, C. Wei, L. Luceri, F. Pierri, E. Ferrara. *Retrieving false claims on Twitter during the Russia-Ukraine conflict*. Companion Proceedings of the ACM Web Conference 2023 (WWW2023). DOI: 10.1145/3543873.3587571

C3 M. Postiglione, G. Esposito, R. Izzo, V. La Gatta, V. Moscato, R. Piccolo. *Harnessing multi-modality and expert knowledge for adverse events prediction in clinical notes*. International Conference on Image Analysis and Processing (ICIAP).

C4 G. Riccio, A. Romano, A. Korsun, M. Cirillo, M. Postiglione, V. La Gatta, A. Ferraro, A. Galli, V. Moscato. *Healthcare Data Summarization via Medical Entity Recognition and Generative AI*. The 2nd Italian Conference on Big Data and Data Science (ITADATA2023). CEUR Workshop Proceedings

C5 A. Ferraro, A. Galli, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, F. Amato. *HEMR: Hypergraph Embeddings for Music Recommendation*. Symposium on Advanced Database System (SEBD2023). CEUR Workshop Proceedings

C6 A. Ferraro, A. Galli, V. La Gatta, V. Moscato, M. Postiglione, G. Sperlì, F. Moscato. *Unsupervised Anomaly Detection in Predictive Maintenance using Sound Data*, Symposium on Advanced Database System (SEBD2023). CEUR Workshop Proceedings

C7 A. Ferraro, A. Galli, V. La Gatta, M. Postiglione. *A Deep Learning pipeline for Network Anomaly Detection based on Autoencoders*. Proceedings of the 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE2022). DOI: 10.1109/MetroXRAINE54828.2022.9967598