

CODAI 2024

**Workshop on Countering Disinformation with  
Artificial Intelligence (CODAI 2024)**

*Proceedings of the Workshop co-located with the 27th European  
Conference on Artificial Intelligence*

**October 20, 2024**

## Preface

This volume contains papers from the 1st Workshop on Countering Disinformation with Artificial Intelligence (CODAI), held at the European Conference on Artificial Intelligence (ECAI) 2024.

Social media platforms which have been designed primarily to allow users to create and share content with others, have become integral parts of modern communication, enabling people to connect with friends, family as well as for broadcasting information to a wider audience. On one side these platforms provide an opportunity to facilitate discussions in an open and free environment. On the flip side, new societal issues have started emerging on these platforms. Among all the issues, the topic of misinformation has been prevalent on these platforms. The term misinformation is an umbrella term which encompasses various entities such as fake news, hoaxes, rumors to name a few. While misinformation refers to non-intentional spread of non-authentic information, the term disinformation points to spreading of a piece of inauthentic information with certain malign intentions.

Initially, researchers have mainly focused on identifying and characterizing misinformation using text based techniques through traditional and advanced NLP techniques. However, with the advancement of techniques and availability of various AI tools, the (mis)information has started appearing in the form of multimodality. For example, a piece of image with incorrect text embedded on it or a morphed video with audio. In addition, the topic of misinformation has impacted individuals and communities from various domains such as medical, political, entertainment, business, etc. This calls for combining forces from different domains. In other words, to counter misinformation computer scientists need to work with domain specialists. To understand the intention a psychologist's inputs can also be vital to understand the reasons for the spreading of misinformation. To summarize, a holistic view is needed to counter the menace of misinformation spread on online social media platforms.

The goal of this workshop is to bring together researchers interested in various domains to not only present their works but also to provide an ecosystem for discussing ideas that facilitate countering the spread of misinformation. We received a total of 17 submissions to the main workshop, of which seven were accepted as oral presentations. Finally, the workshop will feature two distinguished keynote speakers: Paolo Rosso, Universitat Politècnica de València, and David Camacho, Universidad Politécnica de Madrid, Spain.

## **Organizing Committee**

### **Program Chairs**

Rajesh Sharma, University of Tartu, Estonia

Anselmo Peñas, Universidad Nacional de Educación a Distancia, Spain

### **Program Committee**

Rodrigo Agerri, University of the Basque Country (UPV/EHU)

Paolo Rosso, Universidad Politécnica de Valencia (UPV)

Arkaitz Zubiaga, Queen Mary University of London

Harith Alani, Open University, London

Anwitaman Datta, Singapore

Uku Kangur, University of Tartu, Estonia

Shakshi Sharma, Bennett University, India

Johannes Langguth, Simula Research Laboratory, Norway

David Camacho, Universidad Politécnica de Madrid (UPM)

Anselmo Peñas, Universidad Nacional de Educación a Distancia (UNED)

Roberto Centeno, Universidad Nacional de Educación a Distancia (UNED)

Álvaro Rodrigo, Universidad Nacional de Educación a Distancia (UNED)

Rajesh Sharma, University of Tartu, Estonia

Neha Pathak, Indian Institute of Information Technology (IIIT) Delhi

Ahmed Sabir, University of Tartu, Estonia

Giulio Rossetti, CNR, Pisa, Italy

Jan Milan, Applied University of Science, Zurich

Rémy Cazabet, Univ. Lyon 1, Lyon, France

Roshni Chakraborty, University of Tartu, Estonia

## Keynote Talk

# Countering disinformation with AI: discriminating conspiracy theories from critical thinking

Paolo Rosso

Universitat Politècnica de València

**Abstract:** The rise of social media has offered a fast and easy way for the propagation of disinformation and conspiracy theories. Despite the research attention that has received, disinformation detection remains an open problem and users keep sharing texts that contain false statements. In this keynote I will briefly describe how to go beyond textual information to detect disinformation, taking into account also affective and visual information because providing important insights on how disinformation spreaders aim at triggering certain emotions in the readers. I will also describe how psycholinguistic patterns and users' personality traits may play an important role in discriminating disinformation spreaders from fact checkers. Moreover, I will comment on some studies on the propagation of conspiracy theories. In the framework of the PAN Lab at CLEF, we are organising a challenge on oppositional thinking analysis to discriminate between conspiracy narratives and critical thinking. This distinction between critical and conspiracist narratives is vital because considering a message as conspiratorial when it is only oppositional to mainstream views could start a psychosocial process that drives into the arms of the conspiracy communities those who were simply critical about controversial topics such as vaccination or climate change. Most of the work was done in the framework of IBERIFIER, the Iberian media research and fact-checking hub on disinformation funded by the European Digital Media Observatory, and the research projects XAI-DisInfodemics (eXplainable AI for disinformation and conspiracy detection during infodemics), and FAKEnHATE-PdC (FAKE news and HATE speech).

**Bio:** Paolo Rosso is Full Professor of Computer Science at the Universitat Politècnica de València, Spain. His current research interests fall mainly in the area of detection of harmful information in social media, both fake news and hate speech. He is the principal investigator of two related projects: XAI-DisInfodemics on eXplainable AI for disinformation and conspiracy detection during infodemics (PLEC2021-007681), and FAKEnHATE-PdC on FAKE news and HATE speech (PDC2022-133118-I00), both funded by the Spanish Ministry of Science, Innovation and Universities, and by European Union NextGenerationEU/PRTR. He collaborated with the Spanish National Security Department and with the Science and Tech.

**Keynote Talk**

# **Rethinking the problem of disinformation and Artificial Intelligence: boundaries, threats, and trends**

**David Camacho**  
Universidad Politécnica de Madrid

**Abstract:** Disinformation (and more generally misinformation) is spreading everywhere online, causing problems for individuals, societies, and countries. This unchecked dissemination of falsehoods, has nurtured an environment ripe for the proliferation of rumors, propaganda, and hoaxes, exacting a toll on the economic, political, and public health realms, among many other aspects in our daily lives. Confronting this multifaceted adversary demands a united front, drawing upon the collective wisdom and resources of diverse stakeholders including individuals, media entities, governmental bodies, technology firms, and scholars. This keynote endeavours to illuminate the intricate contours of this challenge, delving into some popular Computational techniques such as Machine Learning and Graph Computing as a new set of weapons in the battle against misinformation. Focused primarily on three domains, Natural Language Processing (NLP) and Multimodal Deep Learning (MDL) and Social Network Analysis (SNA), our discourse aims to unveil the potential of these techniques in discerning truth from falsehood. Within the realm of NLP/MDL and SNA, particular attention will be devoted to the facter-check architecture, a novel framework that through the use of ensembles and deep learning techniques based in Transformer technology, enables the identification and tracking of misleading content across the vast expanse of online social networks.

**Bio:** David Camacho is Full Professor at Computer Systems Engineering Department of Universidad Politécnica de Madrid (UPM), he is the head of the Applied Intelligence and Data Analysis research group (AIDA: <https://aida.etsisi.uam.es>), the Director of the PhD program in Computer Science and Technologies of Smart Cities, and the Director of the Master program in Machine Learning and Big Data at UPM. He has published more than 300 journals, books, and conference papers (google scholar). His research interests include Machine Learning (Clustering/Deep Learning), Computational Intelligence (Evolutionary Computation, Swarm Intelligence), Social Network Analysis, Fake News and Disinformation Analysis. He has participated/led more than 60 AI-based R&D projects (National and International: H2020, MCSA ITN-ETN, DG Justice, ISFP, NRF Korea), applied to real-world problems in areas as aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, disinformation countering, or video games among others. He serves as Editor in Chief of Expert Systems from 2023 and sits on the Editorial Board of several journals including Information Fusion, Human-centric Computing and Information Sciences (HCIS), and Cognitive Computation, IEEE Transactions on Emerging Topics in Computational Intelligence (IEEE TETCI), among others. Contact at: [David.Camacho@upm.es](mailto:David.Camacho@upm.es).

# Workshop Program

**9:15 - 9:30** Welcome and opening remarks

**9:30 - 10:30** Invited talk by David Camacho

**10:30 - 11:00** Coffee break

**11:00 - 12:30** Paper presentations

*Analysis of Climate Change Misleading Information in TikTok*

Clara Baltasar, Sergio D'Antonio Maceiras, Alejandro Martin and David Camacho

*Diachronic Political Content Analysis: A Comparative Study of Topics and Sentiments in Echo Chambers and Beyond*

Michele Joshua Maggini, Virginia Morini, Davide Bassi and Giulio Rossetti

*Factoring in context for the automatic detection of misrepresentation*

Bruna Paz Schmid, Annette Hautli-Janisz and Steve Oswald

*Are Misinformation Propagation Models Holistic Enough? Identifying Gaps and Needs*

Raquel Rodriguez-García, Álvaro Rodrigo and Roberto Centeno

**12:30 - 14:00** Lunch break

**14:00 - 15:30** Paper presentations

*Detecting fake news using Twitter social information*

Jesus Maria Fraile Hernandez, Alvaro Rodrigo and Roberto Centeno

*On the Categorization of Corporate Multimodal Disinformation with Large Language Models*

Ana-Maria Bucur, Sónia Gonçalves and Paolo Rosso

*Automated Fact-checking based on Large Language Models:*

*An application for the press*

Bogdan Andrei Baltes, Yudit Cardinale and Benjamin Arroquia Cuadros

**15:30 - 16:00** Coffee break

**16:00 - 17:00** Invited talk by Paolo Rosso

**17:00 - 17:10** Closing remarks

## Table of Contents

Diachronic Political Content Analysis: A Comparative Study of Topics and Sentiments in Echo Chambers and Beyond <i>Michele Joshua Maggini, Virginia Morini, Davide Bass, Giulio Rossetti</i> .....	1
Factoring in Context for the Automatic Detection of Misrepresentation <i>Bruna Paz Schmid, Annette Hautli-Janisz, Steve Oswald</i> .....	11
Detecting fake news using Twitter social information <i>Jesús M. Fraile-Hernández, Álvaro Rodrigo, Roberto Centeno</i> .....	19
On the Categorization of Corporate Multimodal Disinformation with Large Language Models <i>Ana-Maria Bucur, Sónia Gonçalves, Paolo Rosso</i> .....	29
Automated Fact-checking based on Large Language Models: An application for the press <i>Bogdan Andrei Baltes, Yudith Cardinale, Benjamín Arroquia-Cuadros</i> .....	40
Analysis of Climate Change Misleading Information in TikTok <i>Clara Baltasar, Sergio D'Antonio Maceiras, Alejandro Martín David Camacho</i> .....	54
Are Misinformation Propagation Models Holistic Enough? Identifying Gaps and Needs <i>Raquel Rodríguez-García, Álvaro Rodrigo, Roberto Centeno</i> .....	62

# Diachronic Political Content Analysis: A Comparative Study of Topics and Sentiments in Echo Chambers and Beyond

Michele Joshua Maggini<sup>1,\*</sup>, Virginia Morini<sup>2</sup>, Davide Bassi<sup>1</sup> and Giulio Rossetti<sup>3</sup>

<sup>1</sup>*Centro Singular de Investigación en Tecnoloxías Intelixentes da USC, Santiago de Compostela, Galiza, Spain*

<sup>2</sup>*KDD Lab, CNR-ISTI, 56126 Pisa, Italy*

<sup>3</sup>*ISTI-CNR, Pisa, Italy*

## Abstract

Over the past decade, social media platforms have emerged as significant arenas for political discourse and agenda-setting. Among these platforms, Reddit stands out as a prominent site where users actively engage in discussions on controversial topics, often becoming polarized through interactions with like-minded individuals. In this paper, we delve into the realm of political leanings, seeking to understand the predominant topics of interest within echo chambers and whether they diverge from those of unpolarized users. Our primary objective is to ascertain whether echo chambers are characterized by distinct themes discussed therein. Furthermore, we employ cross-sentiment analysis to investigate potential differences in how these themes are perceived across different groups.

## Keywords

natural language processing, political analysis, social network analysis, echo chambers

## 1. Introduction

The rapid growth of social media platforms and online forums has fundamentally reshaped how individuals consume information, share opinions, and engage in political discourse. The proliferation of these online networks has not only transformed the landscape of political communication but has also amplified the formation and influence of echo chambers [1]. Echo chambers are defined as environments where individuals are predominantly exposed to information that reinforces their existing beliefs through repeated exposure to like-minded individuals. This redundancy of content, along with the shared perception of it among users, leads to users' epistemological segregation [2, 3].

This phenomenon has garnered significant scholarly interest due to its potential impact on democratic processes and public opinion [4, 5]. In fact, echo chambers have been observed to contribute to increased polarization, confirmation bias [6], and homophily in online discussions, potentially leading to a distorted perception of reality and hindering constructive debate. The political implications of echo chambers are profound, as they can exacerbate partisan divides and diminish mutual understanding among opposing political groups [7, 8].

These effects are particularly relevant in light of the recent rise of right-wing populist parties. Echo chamber effects, in fact, have been identified as influential contributors to the rise of populist movements. While the roots of populism are multifaceted, scholars have noted the facilitative role of echo chambers in disseminating specialized populist messaging outside mainstream news and party establishments [9, 10, 11]. [12] suggests that individuals, feeling besieged as claimed by populist elites, tend to gravitate towards like-minded groups. Digital media platforms foster the formation and

---

*Proceedings of the 1st Workshop on COUNTERING Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 1–10, October 20, 2024, Santiago de Compostela, Spain*

\*Corresponding author.

✉ michelejoshua.maggini@usc.es (M. J. Maggini); virginia.morini@phd.unipi.it (V. Morini); davide.bassi@usc.es (D. Bassi); giulio.rossetti@isti.cnr.it (G. Rossetti)

🆔 0009-0001-9230-9202 (M. J. Maggini); 0000-0002-7692-8134 (V. Morini); 0000-0003-2025-6559 (D. Bassi); 0000-0003-3373-1240 (G. Rossetti)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



sustenance of homogeneous networks, appealing particularly to populists with their rhetoric of division between "us" and "them." [13] highlights the correlation between populism and the proliferation of "post-truth" politics, wherein factual accuracy is sidelined in favor of personal loyalty and ideological simplicity. Echo chambers, by insulating individuals from dissenting perspectives, can exacerbate this trend, distancing adherents from objective truths.

Yet, as emphasized by [14], empirical evidence supporting a distinct correlation between of right-wing populism leaning and echo chamber dynamics remains scant, with different studies demonstrating fluctuating patterns across different periods and nations [15, 16, 17, 13].

To address this issue, the present study examines how populist political leanings and engagement within or outside echo chambers influence the content and modes of interaction among users on the social media platform Reddit. Indeed, in this social media, users engage with one another by posting and commenting in subreddits aligned with their interests. Using topic modeling, the study investigates distinctions in topic trends among Reddit users based on their political affiliations within and beyond echo chambers. Additionally, it tracks the methods through which discussions are conducted in these different environments. Additionally, the study adopts a diachronic perspective, aiming at providing valuable insights into the evolution of political discourse within echo chambers, identifying shifts in predominant topics and sentiments over time. This approach not only reveals temporal changes but also, in conjunction with our politically fine-grained method, allows for a nuanced examination of how different political affiliations influence the nature of discussions and sentiment expressions within these chambers. In accordance with the findings of [18], which investigated the topological stability of echo chambers, this study hypothesizes that echo chambers will exhibit greater stability in how topics are perceived compared to non-echo chamber structures.

The significance of this study lies in its potential to uncover patterns and trends between "closed online environment" and political communication that may contribute to polarization. By comparing the content and sentiment across politically diverse groups, the aim is to identify whether certain topics or sentiments are more prone to echo chamber effects and how these effects differ across the political spectrum. This study contributes to the broader field of political communication and the ongoing debate about the impact of social media on democratic engagement.

The paper is organized as follows: Section 2 proposes the main contributions and the previous related works constituting the basis of our application; Section 3 introduces the dataset used in this study; Section 4 illustrates the framework, constituted of two parts: Topic Modeling and Cross-Sentiment Analysis; Section 5 reports the data analysis of our case study, reporting the main findings. Finally, Section 6, concludes the paper and provides a look ahead on future research.

## 2. Related Works

Echo chambers are characterized by the reinforcement of ideas, beliefs, or opinions through repeated exposure within an enclosed system, such as online communities or social media networks. The following related works in the area of topic mining in echo chambers highlight the importance of understanding the structure and dynamics of echo chambers, as well as the topics that drive their formation.

**Topological approach** [7] used a network-based approach to identify echo chambers on Facebook, highlighting the role of confirmation bias and homophily in their formation. Similarly, [19] studied the partisan structure inside the retweeting mechanism of political tweets by two networks. They found that the users on the opposite political sides were weakly connected. On the same research line, [5] proposed a method for identifying echo chambers on Twitter by analyzing retweet networks and user ideology. Their findings revealed the existence of polarized echo chambers in political discussions.

**Content approach** [20] has investigated how different social media platforms influence information spread and the creation of echo chambers. By analyzing over 100 million pieces of content on controversial topics from Gab, Facebook, Reddit, and Twitter, two main dynamics were examined: homophily in interaction networks and biased information diffusion. Their findings highlight that

**Table 1**

Original dataset description

Dataset	n. Subreddit	n. Post	n. User
GUN CONTROL	6	180,170	65,111
MINORITIES DISCRIMINATION	6	223,096	52,337
POLITICAL SPHERE	6	431,930	72,399

homophilic clustering is a dominant online behavior, with Facebook exhibiting higher segregation in news consumption compared to Reddit.

Instead, [21] performed a sociolinguistic analysis on tweets from users within echo chambers against those from users outside the echo chamber. Their investigation entails comparative scrutiny of tweet composition, lexical preferences, and thematic emphases, aiming to elucidate potential rationales underlying the observed disparities.

**Mixed approach** [22] focused on diverse subreddits concerning controversial topics and reconstructed the network interaction of users. [22] defined an approach to detect echo chambers on social networks. The framework comprises four steps: (i) the identification of a controversial issue; (ii) the inference of users’ ideology on the controversy; (iii) the construction of users’ debate network; and (iv) the detection of homogeneous meso-scale communities. By modeling the diachronical network’s cohesion and users’ political leaning and interactions, they detected different echo chambers. Authors of [18] proposed an analysis of topological stability and topic detection of the social clusters. By relying upon sentiment analysis and exploiting the textual information coming from sources like posts and comments, the authors investigated how people discussed and perceived a controversial topic. Despite the popularity of that methodology [23, 24], [25] outlined its limitations. Indeed, the viewpoints of diverse users are categorized based on the overall sentiment they convey regarding the topic, rather than their actual alignment on various aspects defining the analyzed subject.

**Textual Forma Mentis Networks** [26, 27, 28] applied a new approach: textual forma mentis networks (TFMN), namely modeling textual concepts as graph neural networks to analyze both semantic and syntactic relationships. That methodology allowed us to simultaneously focus on sentimental, emotional, and rhetorical patterns entailed in online discourses.

Furthermore, [29] suggested applying two emotional lexicons to avoid leading to drastic misinterpretation and conclusions when performing emotion analysis on texts.

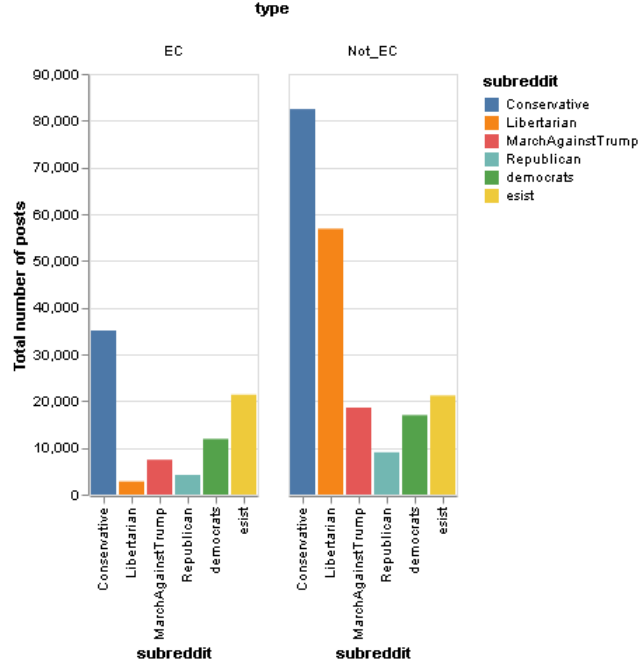
### 3. Data

In this study, three comprehensive datasets compiled, annotated, and preliminarily analyzed in [22, 18] were used. The statistics for the datasets, covering the period from 2017 to July 2019, are presented in Table 1.

In these works, by modeling users’ posts and comments on controversial topics, the authors were able to reveal distinct ideological leanings, categorizing users as *pro-Trump*, *neutral*, or *anti-Trump*.

Subsequently, they introduced a framework for identifying the formation of echo chambers by leveraging both user interaction networks and users’ ideological stances. The communities were delineated using three key metrics: *modularity*, to detect ideologically and topologically homogeneous nodes; *purity*, which measures the product of the frequencies of the most common labels among its nodes; and *conductance*, which calculates the fraction of total edge volume pointing outside the community. In [20], network structures were estimated based on the retention of specific labels within subsets of the network where users shared a common ideology on controversial topics.

[18] focuses on the diachronic evolution of echo chamber topologies. This analysis was enhanced by linking the temporal dimension to the topics discussed, providing insights into the stability of echo chambers over time and the propensity of their members to concentrate on single controversial topics.



**Figure 1:** POLITICAL SPHERE’s posts dataset description.

**Table 2**

Political leaning distribution in echo chambers (EC) and non echo chambers (Not EC) structures.

Leaning	n. Posts in EC	n. Posts in Not EC
antitrump	54,092	75,837
neutral	21,901	47,244
protrump	6,745	31,045

This work focuses on a subset of the dataset: POLITICAL SPHERE, which comprises posts retrieved from the following subreddits as illustrated in Figure 1: r/esist, r/democrats, r/MarchAgainstTrump, r/Conservative, r/Libertarian, and r/Republican. This dataset includes users’ posts categorized by political leaning and echo chamber membership (see Table 2), focusing on discussions related to U.S. politics.

## 4. Methodology

### 4.1. Topic Modeling

In statistics and natural language processing, topic modeling is a commonly used text-mining tool for uncovering hidden semantic structures within a text corpus. In this work, we have applied the BERTopic [30] topic modeling technique to extract topics from texts. BERTopic leverages transformers and c-TF-IDF to create dense clusters, facilitating the generation of easily interpretable topics while retaining key words in the topic extractions. The output of BERTopic consists of generated topics and their probabilities.

Initially, BERTopic converts documents into numerical representations by embedding text in vector space, ensuring that similar texts are positioned closely together, which can be efficiently identified using cosine similarity. To reduce the dimensionality of these representations, we employed UMAP [31], which preserves both local and global information, allowing semantically similar documents to form

clusters while reducing the dataset’s dimensionality. Using HDBSCAN [32], a density-based clustering technique, we detected clusters of various shapes and identified outliers. BERTopic’s outlier reduction method calculates the c-TF-IDF representation for each outlier document and finds the best matching c-TF-IDF topic representation using cosine similarity.

For word-level analysis within topics or clusters, a bag-of-words representation is needed. To highlight differences between clusters, we applied a variant of class-based TF-IDF (c-TF-IDF). Essentially, BERTopic treats all documents within a single category as a single document and then applies TF-IDF. The more significant words within a cluster, the more representative they are of that topic. Consequently, each set of documents is reduced/converted into a single one.

The entire process described above was applied to distinct datasets, differentiating between echo chamber and non-echo chamber contexts.

BERTopic parameters<sup>1</sup> were selected considering the dimension of the echo chambers, aiming at extracting the best representation for our data. With this configuration the aim was looking for few and stable topics to capture the macro-differences preserving both local and global structure in the data. Moreover, with BM25 weighting we stressed the importance of interpretability and diversity in topic representations, reducing the impact of common words while still capturing meaningful bi-grams. Thus, we ensure topics are significant in size and well-represented in the corpus.

The obtained topics were then used to compare online debates taking place within and outside the social clusters about political leaning.

## 4.2. Cross-Sentiment Analysis

To provide a focus on the sentiments and emotions elicited by user-generated contents, we applied two different lexicon-based sentiment analysis algorithms [33]. In details, we leveraged Valence Aware Dictionary and sEntiment Reasoner (VADER) [34] and NRC Emotion lexicon [35]. The former is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. Its sentiment lexicon is composed of a list of lexical features labeled according to their semantic orientation as positive or negative and is attuned to microblog-like contexts. This way, VADER labels the text as positive, neutral, negative, and provides a compound. The NRC Emotion lexicon, on the other hand, assesses the emotional affect conveyed in a text, providing a score for each sentiment or emotion detected in it. Its affective dictionary encompasses approximately 27,000 words, derived from the National Research Council Canada (NRC) affect lexicon and the synonym sets from the WordNet library within the Natural Language Toolkit (NLTK). NRC Emotion Lexicon is constituted by a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). In our case, we used the compound score, calculated by summing the valence scores of each word in the lexicon, adjusting them according to specific rules, and then normalizing the result to range from -1 (most extreme negative) to +1 (most extreme positive). Furthermore, we leveraged the temporal dimension to understand the evolution of the discussions.

## 5. Results

This section reports the experiments’ results on the topic modeling applied on the two networks and the sentiment analysis scores distinguishing the political leanings (anti-trump, pro-trump, neutral) in these clusters.

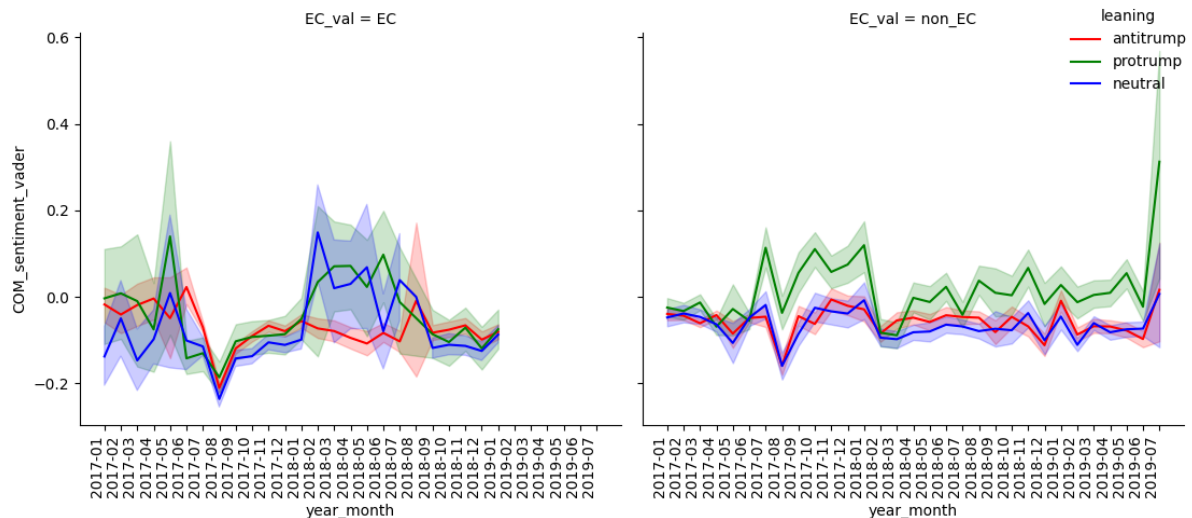
**Topic Modeling** Firstly, we aimed to identify content similarities and dissimilarities between Echo Chambers and Non-Echo Chambers. To extract and analyze the topics, we applied BERTopic. Table 3 present the top 20 most frequent topics in each network. Despite subtle differences in the order and size

<sup>1</sup>UMAP(nneighbors:60, ncomponents:20, mindist:0.05, metric:cosine, randomstate=42); HDBSCAN(minclustersize=90, metric:euclidean, clusterselectionmethod=eom, predictiondata=True); CountVectorizer(stopwords=english, ngramrange=(1, 2)); ClassTfidfTransformer(bm25weighting=True, reducefrequentwords=True); MaximalMarginalRelevance(diversity=0.6), mintopicsize=300).

**Table 3**

Top 20 Topics Frequency in Echo Chamber vs Not Echo Chamber.

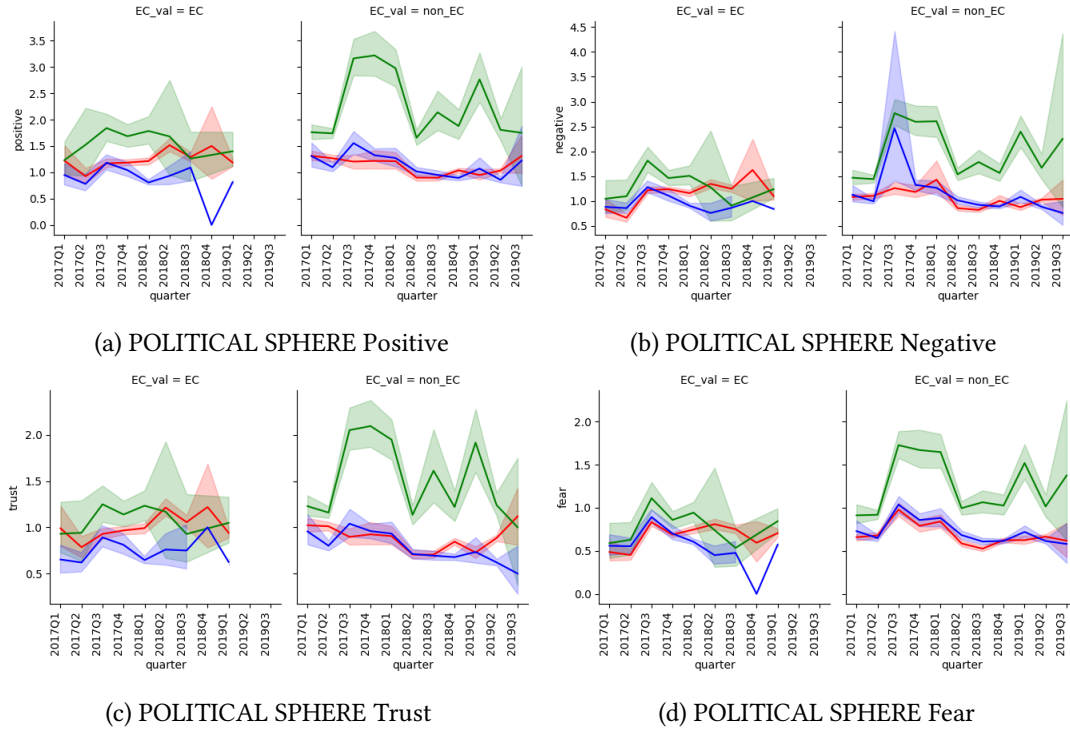
Echo Chamber		Not Echo Chamber	
Topic	Freq.	Topic	Freq.
<i>Democrats</i>	7076	<i>Libertarian party</i>	7707
<i>Roy Moore</i>	6673	<i>Democrats</i>	7460
<i>Conservative</i>	5007	<i>oh_guys_thought_funny</i>	7072
<i>Gun Control, Shootings</i>	4056	<i>Border Wall, Immigration</i>	6906
<i>Border Wall, Immigration</i>	3556	<i>Gun Control, Shootings</i>	6659
<i>Russia, Trump&amp;Putin</i>	3069	<i>Russia, Trump&amp;Putin</i>	5185
<i>Missing</i>	3032	<i>Ben Shapiro</i>	6100
<i>Media, Fake News</i>	2999	<i>Media, Fake News</i>	5370
<i>Taxes</i>	2976	<i>Obamacare, Healthcare</i>	4797
<i>Climate Change</i>	2799	<i>Obama vs Trump</i>	4611
<i>Muslims, Islam</i>	2525	<i>Taxes</i>	4427
<i>FBI, Comey</i>	2460	<i>Transgender, Women</i>	4224
<i>Trump</i>	2334	<i>Climate Change</i>	3935
<i>Obamacare, Healthcare</i>	2228	<i>Capitalism, Socialism</i>	3515
<i>Transgender</i>	1588	<i>Muslims, Islam</i>	3400
<i>Iran, Israel</i>	1545	<i>Facebook, Censorship</i>	3052
<i>NFL, Anthem</i>	1285	<i>Abortion, Parenthood</i>	3027
<i>North Korea, Nuclear War</i>	1280	<i>Brett Kavanaugh</i>	2562
<i>Weinstein Harvey</i>	1189	<i>China, Trade</i>	2359
<i>Robert Mueller</i>	1168	<i>Drugs, cannabis</i>	2353

**Figure 2:** POLITICAL SPHERE VADER compound scores in echo chambers and non-echo chambers grouped by political leaning.

of the two networks, we observed that common topics were discussed with similar frequency in both structures. These topics included Democrats, Conservatives, Libertarians, Gun Control, U.S.-Russia relations, and immigration narratives such as the wall proposed by Trump on the Mexican border. Additionally, in both networks, users debated the perception of popular media outlets like Fox News and CNN as sources of misinformation under the topic "Media, Fake News". Summarizing, that resulted in an homogenous coverage of the contents.

**Sentiment Analysis** To further explore potential discrepancies in the perception of these themes, we analyzed the average sentiment and emotion trends across political leanings in both networks. Firstly, using VADER's compound score, we obtained a general understanding of the trends. As illustrated





**Figure 3:** POLITICAL SPHERE NRC Emotion Lexicon scores for Positive, Negative, Trust and Fear.

in Figure 2, there is a general internal coherence in the patterns, with positive and negative peaks occurring during the same periods across different political leanings. The Non-Echo Chamber leanings showed a generally more neutral evolution than the Echo Chamber and a final peak towards positive perception from pro-trump users. Whereas, in Echo Chamber behavior is more fluctuating. Especially, in 2018, neutral and pro-trump users treated themes positively in contrast with anti-trump. We noted a common negative peak between August and September 2017 in both the networks in conjunction with the Afghanistan conflict being exacerbated and the unveiling of the RAISE Act, a bill introduced under Trump’s government to reduce levels of legal immigration to the United States by halving the number of green cards issued.

Then, to understand the internal coherence between the two networks, we opted to apply a more fine-grained sentiment analysis, namely the NRC Emotion Lexicon, capable of considering ten different sentiments and emotions in a positive range. For the sake of space, we will present only the most relevant plots. By analyzing the users’ sentiments aggregated by learning, we aim to verify their distinct behaviors. Indeed, regardless of the political leaning, Echo Chambers tend to show a less sparse evolution over time. Specifically, antitrump users’ trend often followed neutral ones’ behavior. Non-echo chambers exhibit more volatility in both trust and fear sentiments, with wider confidence intervals and more pronounced peaks and troughs. Fear sentiment peaks are higher and more frequent in Non-Echo Chambers, suggesting more dynamic changes in sentiment outside of echo chambers. Trust levels are generally higher and more stable in Echo Chambers, while fear levels are more. However, both EC and Non-Echo Chambers show similar trends with peaks around the second quarter of 2018, but Non-Echo Chambers has greater variability. Thus, by looking at the data, we can confirm our initial assumption regarding that Echo Chamber’s sentiment is more linear and less fluctuating than in Non-Echo Chambers structures.

Lastly, to validate the coverage of the topics related to their public perception, we used Google Trends<sup>2</sup> data. We tracked users’ searches using our extracted topics as keywords and manually matched these search trends with query trends. This allowed us to confirm that the activity identified in our

<sup>2</sup><https://trends.google.it/trends/>

sentiment analysis corresponded to the specific topics modeled with BERTopic.

## 6. Conclusion

Echo chambers generated in social networks like Reddit, promoting like-minded users interactions, can foster the formation of closed social clusters, where individuals reinforce their shared beliefs by consuming content that aligns with their ideologies. Such processes can then be alimented by political rhetoric gravitating around "in-group/out-group" divisions, such as the one of populist actors, with potential detrimental effects for democratic processes.

To address the lack of empirical research in this field (see Sec.1), in this work we introduced a methodology to assess and analyze the content inside communities reducing the bias towards a single lexicon-based approach.

We considered the first two years of Trump's presidency. Interestingly, on a content level, our study revealed that there is almost no difference in the topics discussed by users in echo chambers and non-echo chambers.

To deepen our understanding of how these topics are discussed, we conducted a diachronic analysis of users' sentiments. This analysis unveiled substantial differences depending on whether users belonged to echo chambers or not, revealing that echo chambers are a more controlled environment, despite the high degree of polarization. This outcome could be explained by the fact that echo chambers are formed by users with the same interests and behaviors. As epistemologically closed clusters, echo chambers' debate processes are more emotionally coherent and do not suffer from high volatility like those in non-echo chambers. Despite not triggering high values of sentiment, users in echo chambers often agree with the rest of the community, reinforcing the auto-exclusive mechanism that enhances the robustness of such networks. This process still promotes the solidification of users' stances.

Additionally, we observed sentiment patterns depending on political leaning. Particularly, pro-trump users in non-echo chambers environments scored high values for each considered sentiment, proving that their vocabulary relies on the usage of more adjectives and more and more heated discussions.

Overall, these results underscore the importance of adopting a fine-grained approach to topic modeling that considers nuanced political orientations, enabling the identification of intricate behaviors at a microscopic level.

However, this study has certain limitations. Firstly, the political leanings of users are determined through a data-driven approach, which may not fully capture the complexity of their political orientations. Secondly, the population under consideration lacks specific social characteristics typically examined in social science studies. It is worth noting that the number of Non-Echo Chamber's users is higher than Echo Chamber's. This could result in biased sentiment analysis results. Additionally, we do not have the tools to collect sensitive variables (such as age, sex, and country of residence), which could significantly enhance the validity and depth of our research findings. Lastly, in 2019 no Echo Chamber in POLITICAL SPHERE was detected. Thus, our plots do not cover this period.

As future research, we plan to delve deeper into user-generated content peculiarities by performing stance detection and conducting rhetorical language analysis to better characterize linguistic differences across users belonging (not belonging) to epistemic enclaves of different political orientations. Such a comprehensive approach will contribute to a deeper understanding of discussion dynamics and the nuances exhibited by dialogues occurring within/outside echo chambers in Reddit.

## Acknowledgments

This work is supported by the EUHORIZON2021 European Union's Horizon Europe research and innovation programme (<https://cordis.europa.eu/project/id/101073351/es>) the Marie Skłodowska-Curie Grant No.: 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA).

Neither the European Union nor the granting authority can be held responsible for them. The authors have no relevant financial or non-financial interests to disclose.

## References

- [1] A. Mahmoudi, D. Jemielniak, L. Ciechanowski, Echo chambers in online social networks: A systematic literature review, *IEEE Access* (2024).
- [2] T. Ulen, Democracy and the internet: Cass r. sunstein, republic.com. princeton, nj. princeton university press. pp. 224. 2001, SSRN Electronic Journal (2001). doi:10.2139/ssrn.286293.
- [3] K. Jamieson, J. Cappella, Echo Chamber: Rush Limbaugh and the Conservative Media Establishment, 2008.
- [4] R. K. Garrett, Echo chambers online?: Politically motivated selective exposure among Internet news users1, *Journal of Computer-Mediated Communication* 14 (2009) 265–285. URL: <https://doi.org/10.1111/j.1083-6101.2009.01440.x>. doi:10.1111/j.1083-6101.2009.01440.x. arXiv:<https://academic.oup.com/jcmc/article-pdf/14/2/265/21491614/jjcmcom0265.pdf>.
- [5] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber?, *Psychological science* 26 (2015) 1531–1542.
- [6] A. Kim, P. L. Moravec, A. R. Dennis, Combating fake news on social media with source ratings: The effects of user and expert reputation ratings, *Journal of Management Information Systems* 36 (2019) 931 – 968. URL: <https://api.semanticscholar.org/CorpusID:149080476>.
- [7] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, The spreading of misinformation online, *Proceedings of the national academy of Sciences* 113 (2016) 554–559.
- [8] S. Flaxman, S. Goel, J. M. Rao, Filter Bubbles, Echo Chambers, and Online News Consumption, *Public Opinion Quarterly* 80 (2016) 298–320. URL: <https://doi.org/10.1093/poq/nfw006>. doi:10.1093/poq/nfw006. arXiv:<https://academic.oup.com/poq/article-pdf/80/S1/298/17120810/nfw006.pdf>.
- [9] S. Engesser, N. Fawzi, A. O. Larsson, Populist online communication: Introduction to the special issue, 2017.
- [10] P. Gerbaudo, Social media and populism: an elective affinity?, *Media, culture & society* 40 (2018) 745–753.
- [11] C. Sandelind, European populism and winning the immigration debate, *Fores*, 2014.
- [12] P. Norris, R. Inglehart, Cultural backlash: Trump, Brexit, and authoritarian populism, Cambridge University Press, 2019.
- [13] S. Waisbord, The elective affinity between post-truth communication and populist politics, *Communication Research and Practice* 4 (2018) 17–34.
- [14] S. Boulianne, K. Koc-Michalska, B. Bimber, Right-wing populism, social media and echo chambers in western democracies, *New media & society* 22 (2020) 683–699.
- [15] F. Esser, A. Stępińska, D. N. Hopmann, Populism and the media: Cross-national findings and perspectives, in: *Populist political communication in Europe*, Routledge, 2016, pp. 365–380.
- [16] A. Haller, K. Holt, Paradoxical populism: How pegida relates to mainstream and alternative media, *Information, Communication & Society* 22 (2019) 1665–1680.
- [17] K. Jacobs, N. Spierings, A populist paradise? examining populists’ twitter adoption and use, *Information, Communication & Society* 22 (2019) 1681–1696.
- [18] E. Cau, V. Morini, G. Rossetti, Trends and topics: Characterizing echo chambers’ topological stability and in-group attitudes, 2024. URL: <https://arxiv.org/abs/2307.15610>. arXiv:2307.15610.
- [19] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter, in: *Proceedings of the international aaai conference on web and social media*, volume 5, 2011, pp. 89–96.
- [20] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, *Proceedings of the National Academy of Sciences* 118 (2021) e2023301118.



- [21] N. Duseja, H. Jhamtani, A sociolinguistic study of online echo chambers on twitter, in: *Proceedings of the third workshop on natural language processing and computational social science*, 2019, pp. 78–83.
- [22] V. Morini, L. Pollacci, G. Rossetti, Toward a standard approach for echo chamber detection: Reddit case study, *Applied Sciences* 11 (2021). URL: <https://www.mdpi.com/2076-3417/11/12/5390>. doi:10.3390/app11125390.
- [23] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, W. Quattrociocchi, Echo chambers: Emotional contagion and group polarization on facebook, *Scientific reports* 6 (2016) 37825.
- [24] D. Wollebæk, R. Karlsen, K. Steen-Johnsen, B. Enjolras, Anger, fear, and echo chambers: The emotional basis for online behavior, *Social Media+ Society* 5 (2019) 2056305119829859.
- [25] M. Amendola, D. Cavaliere, C. De Maio, G. Fenza, V. Loia, Towards echo chamber assessment by employing aspect-based sentiment analysis and gdm consensus metrics, *Online Social Networks and Media* 39 (2024) 100276.
- [26] K. Abramski, L. Ciringione, G. Rossetti, M. Stella, Voices of rape: Cognitive networks link passive voice usage to psychological distress in online narratives, *Computers in Human Behavior* (2024) 108266.
- [27] M. Stella, Cognitive network science for understanding online social cognitions: A brief review, *Topics in Cognitive Science* 14 (2022) 143–162.
- [28] M. Stella, Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media, *PeerJ Computer Science* 6 (2020) e295.
- [29] G. Czarnek, D. Stillwell, Two is better than one: Using a single emotion lexicon can lead to unreliable conclusions, *PLoS ONE* 17 (2022). URL: <https://api.semanticscholar.org/CorpusID:252897016>.
- [30] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, *arXiv preprint arXiv:2203.05794* (2022).
- [31] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426* (2018).
- [32] C. Malzer, M. Baum, Hdbscan( $\epsilon$ ): An alternative cluster extraction method for HDBSCAN, *CoRR abs/1911.02282* (2019). URL: <http://arxiv.org/abs/1911.02282>. arXiv:1911.02282.
- [33] G. Czarnek, D. Stillwell, Two is better than one: Using a single emotion lexicon can lead to unreliable conclusions, *Plos one* 17 (2022) e0275910.
- [34] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: *Proceedings of the international AAAI conference on web and social media*, volume 8, 2014, pp. 216–225.
- [35] S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon, *Computational intelligence* 29 (2013) 436–465.

# Factoring in Context for the Automatic Detection of Misrepresentation

Bruna Paz Schmid<sup>1,\*</sup>, Annette Hautli-Janisz<sup>2</sup> and Steve Oswald<sup>1</sup>

<sup>1</sup>University of Fribourg, Avenue de l'Europe 20, 1700 Fribourg, Switzerland

<sup>2</sup>University of Passau, 94030 Passau, Germany

## Abstract

The aim of the paper is to show how a solid theoretical pragmatic underpinning informs an automatic approach to identifying and classifying misrepresentation in social media. To that end we present a dataset that encodes misrepresentation as well as the source that is misrepresented, building on a set of pragmatically informed annotation guidelines. The performance of standard statistic classifiers for misrepresentation detection is promising. We also perform a fine-grained manual error analysis. The paper closes with a longitudinal analysis of misrepresentation in our dataset and shows that items labelled as misrepresentation increase in years that coincide with political campaigns.

## Keywords

Misrepresentation, pragmatics, natural language processing

## 1. Introduction

A key feature of Trump's political campaign and one-term presidency, which started in early 2017 and ended four years later in 2021, was the strategical use of social media. In October and November 2023 alone, CNN has fact-checked twelve speeches, concluding that his "fall remarks were teeming with false claims - a staggering quantity of misrepresentations, exaggerations and outright lies that made sheer wrongness a central feature of each of his addresses" [1]. Through social media, misrepresentations occur faster and in a more targeted way than in traditional media outlets where political messages are usually assessed in terms of their factual content.

Identifying *misrepresentations*, i.e., a *metarepresentation that is not similar enough to the original representation at the inferential level given a certain context*, in a systematic manner is one building block for helping voters assess the political strategies and worldviews of potential future leaders. But tackling problematic content that is spread in connection to political campaigns is not simply an issue of quickly sifting through large quantities of data. What makes it especially challenging is the quality of the content. As the results of CNN's fact-checking indicate, "sheer wrongness" comes in various forms, one of which is misrepresentation – a notion that may be understood as a form of misinformation, that is, false information.

The aim of the paper is to show how a solid theoretical pragmatic underpinning informs an automatic approach to identifying and classifying misrepresentation in social media. To this end, we present a dataset that encodes misrepresentation as well as the source that is misrepresented, building on a set of pragmatically informed annotation guidelines. The performance of standard statistic classifiers for misrepresentation detection is promising. We also perform a fine-grained manual error analysis. The paper closes with a longitudinal analysis of misrepresentation in our dataset and shows that items labelled as misrepresentation increase in years that coincide with political campaigns.

---

*Proceedings of the 1st Workshop on COUNTERING Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 11–18, October 20, 2024, Santiago de Compostela, Spain*

\*Corresponding author.

✉ bruna.pazschmid@unifr.ch (B. Paz Schmid); annette.hautli-janisz@uni-passau.de (A. Hautli-Janisz);

steve.oswald@unifr.ch (S. Oswald)

🌐 <https://www.fim.uni-passau.de/en/cornlp> (A. Hautli-Janisz); <https://www.steveoswald.ch> (S. Oswald)

🆔 0009-0009-7734-0699 (B. Paz Schmid); 0000-0002-5901-9633 (A. Hautli-Janisz); 0000-0002-5946-1691 (S. Oswald)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The paper proceeds as follows: Section 2 discusses related work in automatically identifying misrepresentation. Section 3 presents the theoretical pragmatic underpinning, followed by a description of the annotation guidelines in Section 4. Section 5 includes information regarding data preprocessing steps, classification, model performance and error analysis. Section 6 discusses longitudinal aspects of the study, while section 7 discusses its results and Section 8 its limitations.

## 2. Related work

Research and governmental-driven efforts alike are trying to contain the effects of misrepresentation by finding ways to automatically identify information online that is misrepresenting the original or is outright false. For instance, the European Commission has sponsored projects aimed at developing AI-tools, such as Monitio, a media monitoring platform that includes fact-checking [2]. The fact-checking system is evidence-based in that it works by retrieving documents from an available collection of news articles which then serve as evidence for the predictions.

However, these approaches to fact-checking and misinformation tend not to differentiate between the various forms of false information. Instead, identification often occurs with the help of vocabularies and datasets labelled based on stance detection [3], truthfulness [4, 5, 6], topic-matching [7], or linguistic features associated with fake news [8, 9].

Misrepresentation itself has rarely been a topic of research in NLP. One exception is Michael Yeomans' study about partisan misrepresentation of political opponents through straw man arguments [10]. In an experimental setting, participants were tasked with articulating their own together with their opponents' positions. They were instructed to write down open-ended responses about the Affordable Care Act. Responses were then labelled depending on whether they were the participants' genuine or imitated positions. At the computational level, texts were scored with sentiment analysis and a lexicon of morally charged words. This was followed by a machine learning model – a logistic LASSO regression - trained to distinguish between texts from opponents and supporters. However, it is difficult to conclude that the study was about the straw man argument specifically since the theoretical underpinning remains unspecified in the paper. As such, the scope of the study appears to have been limited to the analysis of partisan incendiary language accompanying misrepresentation.

In the next section, we will elaborate on our theoretical underpinning. Our approach being theory-driven is what differentiates it from others: Every step is guided by pragmatic theory because, next to identifying misrepresentation, the aim of our study was also to understand the phenomenon from a linguistic and political perspective.

## 3. Pragmatic theory and misrepresentation

The core theoretical underpinning of the present paper is an observation from pragmatics, namely that people tend to show regularity in their language use due to the social aspects of communication [11, pp. 4-6], for instance if people intend to misrepresent or discredit the original. This means that certain patterns in language depend on the context they are embedded in. Therefore, by defining and describing the relevant contexts, we can link patterns of language use to certain pragmatic phenomena such as misrepresentation.

In this study, we build on theories from political discourse analysis [12, 13], pragmatics [14, 15] and philosophy of language [16] in defining misrepresentation as a metarepresentation that is not similar enough to the original representation at the inferential level given a certain context. Example (1) shows a tweet posted by Donald J. Trump on 27 July 2017 in which he claims that the New York Times (belonging to the left-wing spectrum) asserts that 'Fox and Friends' (right-wing spectrum) is the most powerful TV show in America.

**(1) Misrepresentation**

Wow, the Failing @nytimes said about @foxandfriends "....the most powerful T.V. show in America."  
(ID: 890524387773997056)

The original text from an opinion piece in the NYT from 19 July 2017 is shown in (2) and includes some of the text that precedes and follows the quotation:

**(2) Original**

*For years, it was a nontaxing mix of news, lifestyle and conservative couch gab, a warm-up before Fox's day of politics and commentary. Suddenly, for no other reason than its No. 1 fan, it is the most powerful TV show in America. (It's also easily the most-watched cable news morning show, averaging 1.6 million viewers in the year's second quarter, following a post-Trump ratings boost.) [...] [17]*

Based on the theoretical pragmatic underpinning of this paper, a misrepresentation *M* in the political context needs to meet the following criteria *C*:

- C1*: *M* is a metarepresentation in terms of intentionality.
- C2*: There are perceivable structural or componential differences between the original and its metarepresentation.
- C3*: There is noticeable change in the metarepresentation.
- C4*: The difference between the original representation and the metarepresentation results in a difference in comprehension.
- C5*: The difference in comprehension is politically relevant.

Based on *C1*, Example (1) is a metarepresentation, because it contains representative content discernible on a verbatim quotation, on the reported speech verb “said,” and on the use of “[w]ow” to express a psychological state through the positive evaluation of the content of the verbatim quotation. The tweet is also significantly shorter than the overall article, satisfying *C2*.

Regarding *C3*, the quotation is isolated from the article and as a result, there is an emphasis on the content of the quotation which is evaluated positively with “[w]ow.” Criteria *C4* is met for a variety of reasons: For one, the quotation was taken from an opinion article and therefore represents an individual’s opinion (and not necessarily NYT’s). Secondly, Fox & Friends being described as “the most powerful TV show in America” is surprising and concerning given the adverb “[s]uddenly”, which marks unexpectedness. Removing the beginning of the sentence thus changes the overall sentiment in the tweet to a positive one that is absent in the original representation. Thirdly, the author of the original does not appear to believe that the show deserves its new status, since the latter is said to result from “no other reason than its No. 1 fan”, which undermines the inherent quality of the show. Finally, from the first to the second sentence in the original, Fox & Friends develops from “conservative couch gab” to “the most powerful TV show in America”, i.e., the author implies that the world is turned upside down as a result of the former president’s relationship with Fox & Friends.

In terms of *C5*, a politically relevant difference emerges between the original and the misrepresentation: The New York Times’ original article achieves relevance by being identified as a critical piece which deems the show’s new importance to be undeserved and perhaps even dangerous, considering that it is the result of the influence of the president of the United States. From this point of view, it may be a warning against the manipulation of the media for political purposes given the media’s role as a check on governmental power and democracy. This criticism is especially strong since it originates in the media itself. Strikingly, Trump’s metarepresentational rendition achieves political relevance by concluding the opposite, with important contextual implications: Even the New York Times, which is biased and left-wing and is known to be critical of Trump, recognizes how important his favorite show, in which he often participates, is. Additionally, given Trump’s relationship with the

show, presenting it in a positive light may be an attempt to promote himself as well. Thus, it would represent an instance of positive self-presentation. Whereas the original representation is likely an attempt to revise available assumptions and thus change the current state of the world where Fox & Friends is portrayed as undeservedly powerful, the metarepresentation is likely to function as an attempt to strengthen those same assumptions and thus to protect the current state of the world.

## 4. The dataset

The basis of the investigation is the [Trump Twitter Archive](#), a database that contains most tweets posted from Trump’s personal account, @realDonaldTrump, between 2009 and 2021. The site was launched in 2016 and includes 56,571 tweets [18]. The maximum character count of the tweets ranges between 140 and 280. In the following, we discuss the steps taken to prune the dataset in order to be able to model pragmatic theory and misrepresentation in a meaningful way. The dataset and classification code are available at <https://github.com/runastef/auto-ident-trump-misrep.git>.

**Filtering** Two filters are applied consecutively with the aim of increasing the likelihood that the resulting array contains misrepresentation. The first filter extracted tweets containing quotation marks, which usually signal the presence of representative content for instance in the form of reported speech. Therefore, the presence of quotation marks is more likely to be used to comment on an original representation. We exclude tweets predating Trump’s presidential campaign announcement from the selection as well as retweets. The same was attempted for quoted replies by excluding tweets containing the handle @realDonaldTrump. The intention behind excluding retweets and quoted replies was to limit the conversational context of the tweets to the relevant original representations by reducing the amount of representative content. This reduces the contextual complexity of the tweets so as to strengthen the relationship between the utterances and the pragmatic phenomenon of misrepresentation. Eventually, the filter excluded the expressions ‘Nobody’, ‘establishment’, ‘Washington’, ‘elite’, and ‘Congress’. The resulting pre-annotation dataset, which combines both selections, contained a total of 1,737 tweets.

**Annotation study** The annotation of the selected tweets was done by two annotators after instruction, one of them being a co-author of the paper. The annotation guidelines reflect the criteria for misrepresentation discussed in Section 3 in that they are the deciding factors when a tweet is judged as being a misrepresentation of an original. The decision is binary, i.e., ‘misrepresentation’ versus ‘not-misrepresentation’. Inter-annotator agreement with Cohen’s Kappa was 0.765 over the whole dataset, which signals substantive agreement. To increase the quality of the dataset only tweets that both annotators agreed on were included. The resulting dataset has 214 items, 107 are labelled as being an instance of misrepresentation and 107 not being considered misrepresentations.

## 5. Predicting misrepresentation

### 5.1. Preprocessing

In preparation for the application of the text classification algorithms, we normalize, remove noise and anticipate and prevent issues connected to expressions such as URLs. Stop words were removed. Tokenization was done with TweetTokenizer from the NLTK library, which takes into account the specific linguistic structures prevalent in social media.

The list of stop words is updated to reflect Trump’s language use. Since Trump’s language use lacks complexity, removing frequent words may result in the removal of a significant amount of meaning because Trump’s vocabulary contains many elements that would normally count as stop words. As a result, removing such words could influence the classification in a negative way as important patterns linked to his language use might be lost. This might cause an issue with the Naïve Bayes classifier, for example, since it is a probabilistic model that bases its decisions on the frequency with which the



different tokens are present in a certain label during the training phase. Thus, to avoid losing potentially important information, most verbs and conjunctions are kept while pronouns were removed from the list of stop words.

## 5.2. Classification

In the next step, the TfidfVectorizer (Term-Frequency Inverse Document Frequency Vectorizer) from scikit-learn was used for vectorization, i.e., a bag of words representation of all tweets was created, containing the tf-idf values for each word across all tweets. Normally, this vectorizer does its own tokenization, i.e., a library-internal module splits the running text into tokens. For the purpose of this paper, we overwrite this module since tweets have a unique format that can be challenging for tokenization. Since scikit-learn still requires tokenization for internal reasons, we follow the method introduced in David Batista's blog and pass a dummy tokenizer and preprocessor that returns the same input without changing it in scikit-learn [19].

## 5.3. Results

The text classification algorithms employed in this study are: Naïve Bayes, Support Vector Classifier (with a linear kernel) and Random Forest, all imported from the scikit-learn library. All classifiers were left on their standard configurations for learning purposes. On average, all three classifiers performed at around 70% based on mean accuracy (Naïve Bayes: 0.71, SVC: 0.73, Random Forest: 0.68). The accuracy scores ranged between 71% and 72% for Naïve Bayes, between 72% and 74% for SVC, and between 67% and 69% for Random Forest based on a 95% confidence interval. The scores are promising considering that a larger sample size may well improve the performance of the classifiers given that the content of the training data is expected to be more balanced with a larger sample size. The results are promising even when compared with related work: Miranda et al.'s [2] evidence-based automated fact checking platform presents predictions maintained to be correct 58% of the time, and Pérez-Rosas et al.'s [9] fake news detector reportedly presents accuracies between 50% and 76% depending on the domain associated with the dataset that is used. The results also suggest that the Support Vector Machine classifier performed slightly better than the other two.

## 5.4. Error analysis

The performances of Naïve Bayes and SVC are slightly reduced in testing sets containing larger tweets. That is, the performance seems to decrease the higher the number of tokens in the testing set is. This could be due to an imbalance between the training set and the testing set. Random Forest was probably less affected by this because it relies on the decision of multiple classifiers that reach their individual decisions based on a significantly smaller sample size than SVC and Naïve Bayes during the same run.

Incorrect predictions often arise when the language in the tweet is uncharacteristically complex, for instance when a supporter's tweet is copied and posted through Trump's account and contains language that is more complex than Trump's typical writing style. Here is such an example:

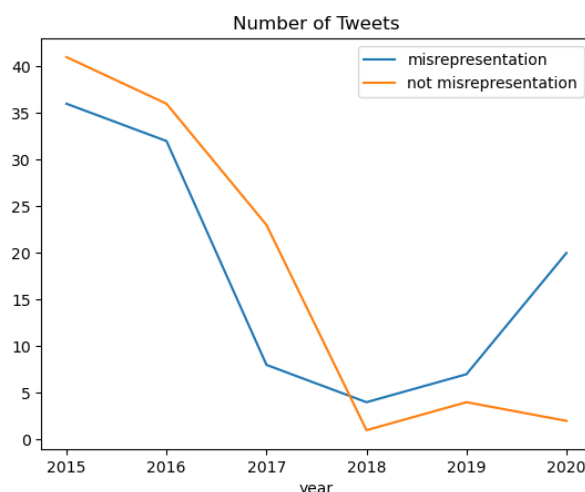
### (3) Copied and posted tweet

*"@racheljoycowley: I'm done with Macy's. Apparently, they follow the trend of trying to force legislation on every American freedom. Done!" (ID: 616646192609558528)*

In an arbitrary run, this tweet was incorrectly predicted as not-misrepresentation by all three classifiers.

## 6. Tracking misrepresentation over time

Having annotated the tweets to be either misrepresentation or not, counting them over the years on the x-axis allows for a longitudinal view of the tweeting behavior of Donald J. Trump.



**Figure 1:** Number of Misrepresentations over time

Figure 1 hints at a possible correlation between the number of tweets containing misrepresentations that Trump posted and the years in which he was active in presidential election campaigns. The overall trend towards less tweets was already evident in the pre-annotation dataset. Thus, it is not surprising. What is interesting is the way in which the trend appears to change abruptly in 2018 and then again in 2019. Both categories experience an increase between 2018 and 2019. However, whereas the instances of not-misrepresentation begin to decrease from 2019 on, the instances of misrepresentation rise sharply.

It is also worth pointing out that given the connection to campaign years, the graph may be interpreted in terms of the reliability of the created dataset since misrepresentation is probably more likely to occur during political campaigns due to the nature of politics.

## 7. Discussion and summary

Future work could include a general analysis of Trump’s language use, which might help to improve the preprocessing. This could take the form of a linguistic features engineering study to determine his writing style. The findings could then be used to create a stop words list that is better able to reflect Trump’s language use although, perhaps, a more generalized approach based on country and political affiliation may be more helpful for the study of misrepresentation itself because it would be easier to generalize. It may also prove to be more practical in terms of implementation.

The process of evaluating the performances of the classifiers could be simplified and improved with an explainer. The LIME (Local Interpretable Model-Agnostic Explanations) Text Explainer was difficult to implement even after replacing the LinearSVC classifier with an SVC classifier with a linear kernel. Initially, the LinearSVC classifier was used, but it was changed into an SVC classifier with a linear kernel. It should produce similar results seeing as LinearSVC is an implementation of SVC. The change was necessary because LinearSVC does not support the function `predict_proba`, which calculates the probabilities for each class prediction. It would have been difficult to evaluate the performance of the LinearSVC classifier without `predict_proba`. LIME was chosen because, in theory, it should work well with all three classifiers as long as the models are able to “predict the probabilities of the categories.” According to Albrecht et al., LIME “works locally by taking a look at each prediction separately. This is achieved by modifying the input vector to find the local components that the predictions are sensitive to” [20, p. 195]. Then, “[f]rom the behavior in the vicinity of the vector, it will draw conclusions about which components are more or less important” and “visualize the contributions and explain the decision mechanism of the algorithm for individual documents” [20, p. 195]. However, LIME’s implementation and interpretation was challenging. As such, in the end, it was not taken into consideration in the error analysis. And yet, evaluating the performance of the classifiers would probably have been significantly

more straightforward with such an explainer. In its absence, the process is considerably slowed down. Improving explainability would have practical implications for future research, for instance, although we favor a theory-driven approach, together with the misrepresentation dataset the findings could be used to improve or expand available vocabularies employed in current fact-checking systems.

To summarize, this study contributes to the research of pragmatically relevant phenomena with computational linguistic methods by discussing how to account for various aspects of the context at different stages of the research process. Specifically, efforts were made to retain contextual information related to the social and political contexts. To this end, a framework was developed for the pragmatic analysis of political misrepresentation with computational methods. Based on the framework, annotation guidelines were written to enable the creation of a misrepresentation dataset, which was then employed to train supervised machine learning algorithms used for text classification. The three algorithms performed at around 70% with SVC performing slightly better than the other two algorithms.

## 8. Limitations

The findings of this study may be limited by the small size of the dataset, the data selection process, data source, and the format of the text data.

The study relied on tweets posted from Trump’s Twitter account. Consequently, the methods applied in this study might yield different results on the discourses of other individuals especially if one considers Trump’s unique language use and political affiliation. Widening the scope of the study to include political discourse from a larger number of politicians is likely to lead to better insight into political misrepresentation. To this end, our study will hopefully provide a basis for further research into a topic that has not received a lot of attention so far.

The relative novelty of pragmatic research into misrepresentation with computational linguistic methods also explains the chosen format. The smaller number of tokens present in tweets were expected to facilitate computation given the theoretical underpinning. Although the small format may limit the study’s generalizability, it facilitates the qualitative analysis of the results which will help us to widen the scope of the analysis in future studies on this topic.

## References

- [1] D. Dale, Trump’s avalanche of dishonesty: Fact-checking 102 of his false claims from this fall, 2023. URL: <https://edition.cnn.com/2023/12/01/politics/trump-dishonesty-avalanche-102-fall-false-claims/index.html>.
- [2] S. Miranda, D. Nogueira, A. Mendes, A. Vlachos, A. Secker, R. Garrett, J. Mitchel, Z. Marinho, Automated fact checking in the news room, in: L. Liu, R. White (Eds.), *The World Wide Web Conference*, ACM, New York, NY, USA, 2019, pp. 3579–3583. doi:10.1145/3308558.3314135.
- [3] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: K. Knight, A. Nenkova, O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, CA, USA, 2016, pp. 1163–1168. doi:10.18653/v1/N16-1138.
- [4] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: M. Palmer, R. Hwa, S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. doi:10.18653/v1/D17-1317.
- [5] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: R. Barzilay, M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Short Papers*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. URL: <https://aclanthology.org/P17-2067>. doi:10.18653/v1/P17-2067.



- [6] Y. Qiao, D. Wiechmann, E. Kerz, A language-based approach to fake news detection through interpretable features and brnn, in: A. Aker, A. Zubiaga (Eds.), *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 14–31. URL: <https://aclanthology.org/2020.rdsm-1.2>.
- [7] A. Miani, T. Hills, A. Bangerter, Loco: The 88-million-word language of conspiracy corpus, *Behavior research methods* 54 (2022) 1794–1817. doi:10.3758/s13428-021-01698-z.
- [8] G. Kuzmin, Larionov, Daniil, D. Pisarevskaya, I. Smirnov, Fake news detection for the russian language, in: A. Aker, A. Zubiaga (Eds.), *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 45–57. URL: <https://aclanthology.org/2020.rdsm-1.5/>.
- [9] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, NM, USA, 2018, pp. 3391–3401. URL: <https://aclanthology.org/C18-1287>.
- [10] M. Yeomans, The straw man effect: Partisan misrepresentation in natural language, *Group Processes & Intergroup Relations* 25 (2022) 1905–1924. doi:10.1177/13684302211014582.
- [11] G. Yule, *Pragmatics*, Oxford introductions to language study, Oxford University Press, Oxford, UK, 2011.
- [12] T. A. van Dijk, Ideology and discourse analysis, *Journal of Political Ideologies* 11 (2006) 115–140. doi:10.1080/13569310600687908.
- [13] J. Wilson, Political discourse, in: D. Tannen, H. E. Hamilton, D. Schiffrin (Eds.), *The Handbook of Discourse Analysis*, Blackwell Handbooks in Linguistics, Wiley Blackwell, Malden and Oxford, 2015, pp. 775–794.
- [14] E.-J. Noh, *Metarepresentation: A Relevance-Theory Approach*, volume 69, John Benjamins Publishing Company, Amsterdam, 2000.
- [15] D. Wilson, D. Sperber, Relevance theory, in: L. R. Horn, G. Ward (Eds.), *The Handbook of Pragmatics*, Wiley, Malden, MA, 2006, pp. 607–632.
- [16] J. R. Searle, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, Cambridge, UK, 2012. doi:10.1017/CBO9781139173452.
- [17] J. Poniewozik, Watching ‘fox & friends,’ trump sees a two-way mirror, 2017. URL: <https://www.nytimes.com/2017/07/19/arts/television/donald-trump-fox-friends.html>.
- [18] B. Brown, Trump twitter archive, 2016. URL: <https://www.thetrumparchive.com/>.
- [19] D. Batista, Applying scikit-learn tfidfvectorizer on tokenized text, 2018. URL: <https://www.davidsbatista.net/blog/2018/02/28/TfidfVectorizer/>.
- [20] J. Albrecht, S. Ramachandran, C. Winkler, *Blueprints for Text Analytics Using Python: Machine Learning-Based Solutions for Common Real World (NLP) Applications*, O’Reilly Media, Inc., Sebastopol, CA, 2021.

# Detecting fake news using Twitter social information

Jesús M. Fraile-Hernández\*, Álvaro Rodrigo and Roberto Centeno

NLP & IR Group at UNED (Spain)

## Abstract

In this paper, the aim is to study whether social information can provide useful information when classifying news. For this purpose, a set of news items in Spanish has been extended with social information. Subsequently, a classifier model has been proposed to carry out this task, mixing the social information previously extracted with the textual information of the news item. Finally, we have studied which social features are the most relevant in this task.

## Keywords

Social information, Classifying news, Classifier model, Social features, Fake news detection

## 1. Introduction

Due to the increase in communication channels in recent decades, users have access to an immense amount of information almost instantaneously. However, it is relatively easy to fall for hoaxes or misinformation on social media.

Traditional models of fake news detection focus on detecting the linguistic characteristics of the news. Subsequently, in [1], pre-trained embeddings were used along with LSTM. Finally, with the emergence of contextual models, [2] leveraged the pre-trained BERT model, to perform transferred learning and identify the veracity of news.

However, due to the difficulty even for a human to discern between true and false news, sometimes the textual information in the news is not enough. In [3] it is proposed at a theoretical level the possibility of creating a hybrid approach that incorporates the linguistic characteristics of the news and an analysis of the networks that are formed around that news. In [4] the author uses different features to identify fake news in popular Twitter threads. In [5] fake news is detected using only the extracted textual information. Regarding hybrid models, the CSI model proposed in [6] performs a characterisation in three modules: capturing, scoring and integrating. In [7], a news detection model is proposed that considers the association of user interactions, the editor's bias and the users' stance towards the news.

The aim of this work is to study whether social information can provide useful information for the detection of fake news. To this end, social information has been collected from Twitter to extend FakeDeS, a relevant corpus of news in Spanish, and a model has been designed to include textual and social information. Furthermore, we intend to study which social features are the most relevant for news classification.

The rest of this paper is structured as follows: Section 2 describes the datasets to be used along with the task to be solved. Section 3 describes the methodology followed including the extraction of social information from Twitter along with the models proposed based on the data they use. Section 4 includes the evaluation metrics used. Section 5 then presents the results, which will be discussed in Section 6. Finally, conclusions and future work are given in Section 7.

---

*Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 19–28, October 20, 2024, Santiago de Compostela, Spain*

\*Corresponding author.

✉ jfraile@lsi.uned.es (J. M. Fraile-Hernández)

ORCID 0009-0001-5474-4844 (J. M. Fraile-Hernández)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Team	Fake	True	$F_{macro}$	Accuracy
<b>GDUFS_DM</b>	<b>0.7666</b>	<b>0.7649</b>	<b>0.7666</b>	<b>0.7657</b>
Haha	0.7548	0.7522	0.7548	0.7535
Chats_	0.7514	0.7690	0.7514	0.7605
SINAI	0.7385	0.7821	0.7385	0.7622
baseline-BERT	0.7321	0.7432	0.7321	0.7378
baseline-BOW-SVM	0.7217	0.7359	0.7217	0.729
Lcad_UFES	0.7102	0.6837	0.7102	0.6976
CITIUS-NLP	0.7098	0.4940	0.7098	0.6311
baseline-CHAR-3-GRAMS-SVM	0.7063	0.6883	0.7063	0.6976
zk15120170770	0.7053	0.6053	0.7053	0.6626
ForceNLP	0.6925	0.4739	0.6925	0.6119
GRX	0.6915	0.5624	0.6915	0.6381
TSIA	0.6860	0.5263	0.686	0.6224
FREE	0.6855	0.6519	0.6855	0.6696
LIMCA	0.6812	0.7027	0.6812	0.6923
ZZWEI	0.6737	0.6794	0.6737	0.6766
Premjithb	0.6576	0.7177	0.6576	0.6906
Sdamian	0.6542	0.75	0.6542	0.7098
Yeti	0.6316	0.609	0.6316	0.6206
Gulu	0.6226	0.476	0.6226	0.5612
Nicksss	0.6119	0.7592	0.6119	0.7028
Bribones tras la esmeralda perdida	0.5835	0.5878	0.5835	0.5857
WSSC	0.5118	0.6657	0.5118	0.6031
Skblaz	0.4838	0.649	0.4838	0.5822

**Figure 1:** Results IberLEF 2021 on the test set.

## 2. Dataset and task

The dataset we will work with is the Spanish Fake News Corpus (FakeDeS) [8], which contains publications in Spanish about different events that were collected from November 2020 to March 2021. Each of these publications is labelled as true or false. Newspaper websites and fact-checking websites were mainly used to collect the information.

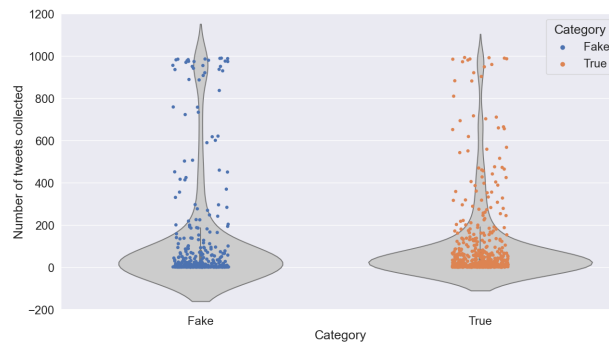
The dataset is divided into 3 files with a total of 1543 news items. Because of the methodology used, it has been decided to merge the training and development files to obtain what we will call the training set. Each of the news items contains information such as the topic, the name of the source, the headline, the text and the link to the news item.

The training set has a total of 971 news items, of which 480 are false and 491 are true. On the other hand, the test set consists of 572 news items, half of which are true and half of which are false. Therefore, we are dealing with balanced data sets.

The topics covered in the training corpus are: politics, entertainment, sport, society, science, health, economy, security and education.

It should be noted that the test set has news related to Covid-19, while the training set does not present any news related to this topic (the most similar are the health news, but in no case do they mention Covid-19). Therefore, the models that are proposed will have to correctly classify this topic without having seen it in the training.

In IberLEF 2021, a shared task was proposed whose objective was to classify a series of news items as true or false. To do so, the FakeDeS corpus described above was used. A report was published in [9], which collected the most important characteristics of the best-performing models. The results of this task by the different participants can be seen in Figure 1. Among the approaches used to solve it, the participants of the GDUFS team, the team that achieved the best accuracy, used a BERT model and sample memory with an attention mechanism. The method consisted of taking the first and last segments of the texts and feeding them into a BERT system, obtaining two embeddings (head and tail). In addition, there is a matrix called ‘sample memory’, which is obtained by taking a random sample of the head and tail embeddings; this matrix is used in an attention mechanism with the rest of the



**Figure 2:** Violin diagram of the number of tweets collected.

texts. In contrast to the GDUFS\_DM approach, the participants of team Haha, the second-placed team, employed feature selection with a weighted tf-idf and a multilayer perceptron. This model not only analysed the content of the news item, but also combined information such as the publisher of the news item or the topic of the news item.

### 3. Methodology

This section describes the methodology used to extract social information from Twitter users. In addition, the models trained according to the type of data they use are presented.

#### 3.1. Social information extraction

The main objective of this work is to study the information provided by social information when detecting fake news, and as mentioned in Chapter 1, there is no corpus in Spanish that contains this information. This is why we decided to extract this information from the social network Twitter, using the API provided by the platform.

For each news item, we searched for those tweets that contained the headline of the news item or the link to it. To solve the problem of the maximum length of the queries, special characters have been eliminated from the news headlines.

According to [4] and [5] there is a series of metadata of the tweets that allow extracting information about whether the user may be prone to the propagation of fake news or the tweet may contain untruthful information. Therefore, it has been decided to extract the following metadata from each of the tweets.

- **Tweet.** Text of the tweet, id of the author, id of the tweet, number of retweets, number of replies to the tweet, number of likes, number of citations of the tweet.
- **User.** username (*str*), user creation date (*date*) ISO 8601, verified user (*bool*), number of followers (*int*), number of followed (*int*), number of tweets (*int*), number of times listed (*int*).

We have managed to extract posts from 41.67% of the total number of news items. Of these, the distribution of the number of tweets collected per news item shows a high concentration in the (0, 200) interval, representing 86% of the news items. Within this interval, it is observed that true news tends to receive more interaction. However, as the number of tweets about a news item increases, it is evident that fake news receives a greater number of interactions. This trend can be seen in the violin diagram presented in Figure 2.

It is worth noting that, although the news is written in Spanish, there are tweets in English or French that talk about the news. This is especially true for news related to Covid-19.

### 3.2. Textual models

In this section, the textual methods used for the binary classification of the news items will be presented. The full text of the news item has been used, so it has had to be preprocessed. For the non-contextual models, urls, emoticons or non-textual expressions, stopwords, the text has been converted to lowercase and the processes of lemmatisation and stemming have been applied. However, for the contextual models, only the urls have been eliminated.

Subsequently, 5 different approaches have been used.

1. Vector space model based on bags of words (BoW).
2. Vector space model using a weighted tf-idf.
3. Bigram counting.
4. Neural Networks and deep learning.
5. Contextual models.

For approaches 1, 2 and 3, Naive Bayes, SVM, Logistic Regression, Decision Trees and Random Forest models have been trained. For approach 4, multilayer perceptrons with input the tf-idf weight vector, multilayer perceptrons and convolutional networks with an embedding layer and multilayer perceptrons, convolutional networks, LSTM, GRU and bidirectional networks with a pre-trained embedding layer. Finally, for approach 5, the BETO model has been selected: Spanish BERT [10] with a final classification layer with two neurons. This model is a BERT model trained with the whole-word masking technique on a large corpus of more than three billion Spanish words.

### 3.3. Models with social information

The methods that use only the social information of the news collected use the following metadata for each published tweet: number of retweets, number of replies, number of likes of the tweet, number of quotes of the tweet, verified user, number of followers, number of followed, number of tweets of the author, number of times the author has been listed. Then, in order to record the impact of the news item on social networks, the number of tweets collected for this news item is added.

To represent all the tweets that talk about a certain news item, an average of the previous characteristics of each tweet has been calculated. Finally, the standard deviation of each characteristic was added. In this way, a data matrix with 20 columns is obtained (where the column relating to the deviation of the number of tweets of the news item is always 0).

Once the feature matrix has been obtained, different learning models have been used with different hyperparameter explorations such as Decision Trees, Random Forest, SVM, Gradient Boosting, Adaptive Boosting, MLP,...

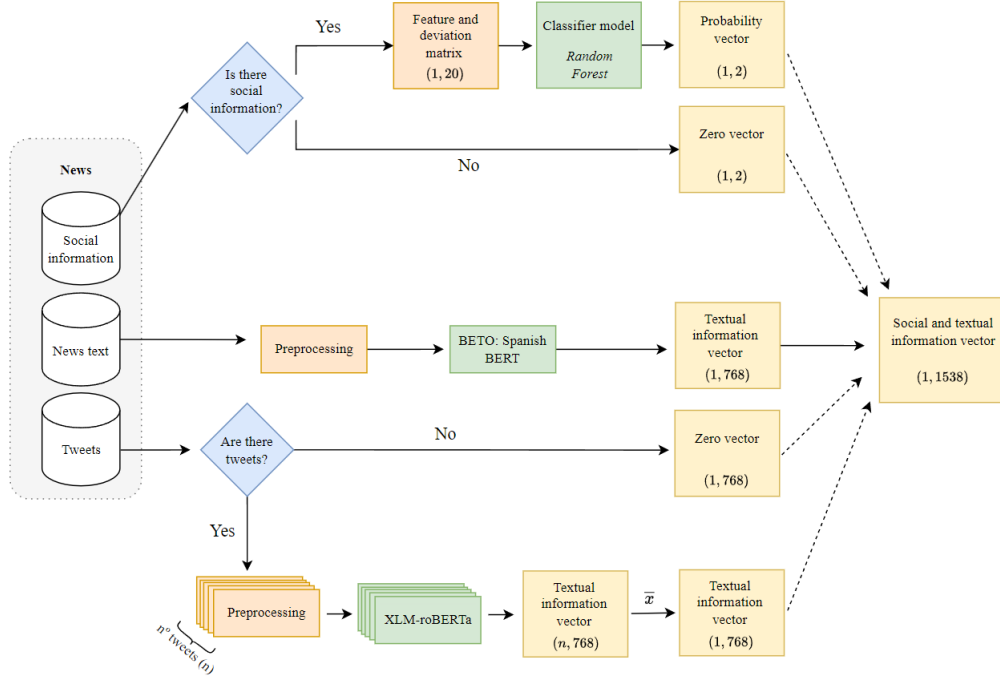
### 3.4. Hybrid model

A hybrid model has been developed that seeks to take advantage of both the textual information provided by the text of the news item and the social information extracted from the Twitter data (both the non-textual information of the section and the text of the tweets collected).

In this model, for each news item, a specialised model is used to classify the news using social information. For this purpose, the best model from the previous subsection (Random Forest) is selected. With this model, for each news item, the probabilities of being true or false are extracted using as input the corresponding row of the matrix of social characteristics with standard deviation described in that section. In the event that no tweets could be extracted from a news item, the output would be a vector of two zeros.

In parallel, the text of the news item is processed using the BETO: Spanish BERT model [11]. The output is a vector of dimension 768.

In parallel to these two processes, for each news item with tweets collected, the text of each tweet is pre-processed (eliminating URLs and tokenising) and subsequently processed using the pre-trained XLM-roBERTa-base model [12]. This transformer model has been trained on a corpus of about 198



**Figure 3:** Workflow of the hybrid model.

million tweets in 8 different languages (Spanish, Arabic, English, French, German, Hindi, Portuguese and Italian) and is specialised in sentiment classification (positive, negative or neutral). In our case, the last layer of the model will be removed, obtaining as output a vector of length 768 that will represent the most relevant features of the text of the tweet.

For each available tweet, the previous process has been carried out, obtaining a vector of length 768. Finally, an average of all the vectors of the tweets of the news item has been made to obtain a vector that represents the tweets of that news item. If the news item had no social information, a vector of zeros is returned.

Then, the three vectors are joined to obtain a vector of dimensionality 1538. This flowchart can be seen in Figure 3

Once all the news has been processed following the previous diagram, several models have been trained such as Decision Trees, Random Forest, SVM, Gradient Boosting, Adaptive Boosting, MLP, ...

## 4. Evaluation

Two different methodologies have been used to evaluate the models, a cross-validation and an evaluation on the test set.

### 4.1. $k$ -fold cross-validation

Cross-validation is one of the most widely used methods to estimate the prediction error of a model with a given set of hyperparameters. A  $k$ -fold (or  $k$ -fold cross-validation) has been used. This method divides the data set, in our case the train set together with the development set, into  $k$  equal parts  $P_1, \dots, P_k$ . For each  $P_n$  the model is trained using the other  $k - 1$  parts and the error in predicting the  $P_n$  data (data never seen by this model) is calculated. By doing this for the  $k$  parts we obtain a set of errors. With these  $k$  errors we calculate their mean and variance to obtain a measure of the average error of that model with those hyperparameters.

It should be noted that this method requires a fairly large computational cost, since for a cross-validation of  $k$ -folds it would be necessary to train  $k$  models. As a general rule, a value of 5 or 10 is



Textual models	F1
TF-IDF (RF)	<b>0.849</b>
BoW (RF)	0.825
Bigramas (RF)	0.822
MLP (Embedding)	0.786
MLP (TF-IDF)	0.751
CNN	0.740
BETO	0.727
GRU	0.678

**Table 1**

Cross-validation results of textual model training.

usually chosen as a good compromise between bias and variance. In our case a 5-fold cross-validation has been used.

## 4.2. Test set evaluation

Finally, for the model that has performed best in the previous cross-validations, the test set will be evaluated. This set will never be seen by the model and will provide a representation of the generalisability of the model.

## 4.3. Evaluation metrics

To evaluate the performance of our classification model, we use the F1 metric. The F1 value will be calculated for both true and false classified news. With this, the value *Macro-F1*, or simply *F1*, will be calculated as the average between the two previous values.

# 5. Results

In this section the results of the various trained models will be presented. For each approach in the section 3 the following results will be shown:

- Within the training of a particular approach, the *Macro-F1* value of the best algorithms used will be shown. The average of the *Macro-F1* values will be reflected using 5-fold cross-validation.
- For each approach, the model with the best *Macro-F1* will be selected during training. Subsequently, it will be retrained with all data and evaluated on the test set. The  $F1_{Fake}$ ,  $F1_{True}$ , *Macro-F1* and the Accuracy of the model will be exposed.

## 5.1. Textual models

The training results of the methods described in section 3.2 are listed in Table 1.

It can be seen that the non-neural models stand out from those using neural networks. This could be due to the fact that the models being used have a large number of parameters to optimise and we have a rather limited data set. It is worth noting that the use of pre-trained embeddings has resulted in lower performance than training the embeddings from scratch. Also noteworthy is the poor performance obtained with recurrent networks, models that have required a large amount of training time and are commonly used for language processing problems. The best performing approach has been to use a weighted tf-idf together with a Random Forest model.

The results of the evaluation of this model on the test set and the results of the teams participating in IberLEF 2021 are shown in Table 4.

Social information models	$F1$
Random Forest	<b>0.845</b>
Gradient Boosting	0.834
Adaptive Boosting	0.826
Extremely Randomized Trees	0.817
Decision Trees	0.797
K-Nearest	0.788
Multilayer Perceptron (MLP)	0.787
SVM	0.785
Passive-Aggressive Classifier	0.785
Perceptron with two hidden layers	0.783
Linear Discriminant Analysis (LDA)	0.781
Multinomial Naive Bayes	0.781
Perceptron with one hidden layer	0.781
Bernoulli Naive Bayes	0.779
Quadratic Discriminant Analysis	0.776
Logistic Regression	0.703

**Table 2**

Cross-validation results of social models.

## 5.2. Social information models

The training results of the methods described in section 3.3 are collected in Table 2.

We can see that the  $F1$  of the models is quite high. Tree-based models occupy the top 5 positions in the list. In addition, those based on clusters of trees stand out from individual decision trees. The best performing approach was a Random Forest model. It should be remembered that this model has only been trained and evaluated with those news items from which it has been possible to extract social information, so the training and test set is smaller than in the rest of the cases.

Due to these results, it has been decided to choose the Random Forest classifier for the social information for the hybrid model, as indicated in section 3.4.

## 5.3. Hybrid model

The training results of the methods described in section 3.4 are listed in Table 3.

In view of the training results, any of the first 2 models would be valid for your choice. The rest of the models have a very similar accuracy to the first three. It has been decided to select logistic regression over decision trees since it is a simpler algorithm, with a smaller number of hyperparameters and with a lower computational cost.

The results of the evaluation of this model on the test set and the results of the teams participating in IberLEF 2021 are shown in Table 4.

## 6. Discussion

This section presents a discussion of the results obtained.

In view of the results shown in Tables 1 and 3, it can be seen that the approach that obtains the best  $F1$  is a model that uses only textual information, more specifically a Random Forest with a weighted tf-idf. This approach obtains a higher  $F1$  compared to other types of models that include social information, so that a priori it could be thought that social information does not provide relevant information.

However, in Table 4, we can see how on the test set the model that uses only textual information obtains worse results compared to the hybrid model. This is due to the fact that when using a tf-idf weight it is possible that there are words in the corpus on which the weight is applied (training news corpus) that do not exist in the test set. This is why models such as transformer networks pre-trained



Hybrid Model	F1
Decision Trees	<b>0.818</b>
Logistic Regression	<b>0.818</b>
SVM	0.809
Linear Discriminant Analysis	0.809
Random Forest	0.809
Gradient Boosting	0.809
Passive-Aggressive Classifier	0.809
Adaptive Boosting (AdaBoost)	0.809
Extremely Randomized Trees	0.809
Quadratic Discriminant Analysis	0.809
Multilayer Perceptron (MLP)	0.809
K-Nearest	0.809
Perceptron with three hidden layers	0.809
Perceptron with two hidden layers	0.808
Perceptron with one hidden layer	0.808
Multinomial Naive Bayes	0.631
Bernoulli Naive Bayes	0.607

**Table 3**

Cross-validation results of hybrid model.

	Fake	True	$F_{macro}$	Accuracy
Textual Models	0.7140	0.7488	0.7314	0.7325
Hybrid Model	<b>0.7900</b>	0.7352	0.7626	<b>0.7657</b>
GDUFS_DM	0.7666	0.7649	<b>0.7666</b>	<b>0.7657</b>
Haha	0.7548	0.7522	0.7548	0.7535
Chats_	0.7514	0.7690	0.7514	0.7605
SINAI	0.7385	<b>0.7821</b>	0.7385	0.7622
baseline-BERT	0.7321	0.7432	0.7321	0.7378
baseline-BOW-SVM	0.7217	0.7359	0.7217	0.7290

**Table 4**

Results on the test set. Including the best participants of IberLEF 2021.

on large corpora will have more generalisation capacity and, therefore, will be able to obtain better results.

Once social information is introduced into the model, a significant increase in results can be seen. This is due to the fact that on the one hand the text is being processed using transformer models with a very high generalisation capacity and that the non-textual social information extracted from Twitter is the same regardless of the subject matter.

Comparing the models with respect to the best classified in IberLEF 2021, Figure 1, it can be seen that the hybrid model is the one that best classifies Fake news. This hybrid model obtains the same Accuracy as the first ranked team.

In addition, a study has been carried out on which social information features are the most relevant for the model. For this purpose, the importance of the permutation set out in [13] has been used. It can be seen that 8 of the 9 most relevant features only depend on the author's information and not on the content or information of the tweet. These 9 features are, in order of importance: listed\_count, following\_count\_std, followers\_count, tweet\_count\_std, followers\_count\_std, quote\_count\_std, verified, verified\_std, tweet\_count. In addition, within these characteristics, the information provided by those obtained from the standard deviation of the set of tweets collected for each news item stands out.

The percentage of importance of the most relevant features used in the logistic regression of the hybrid model has also been calculated. To calculate the importance of each feature,  $f_i$ , the coefficients of the regression,  $w_i$ , have been extracted and the following operation has been carried out  $f_i = e^{w_i}$ . Finally, the percentage of each of them has been calculated. With this, the most relevant characteristic

for the model, with 10 times more importance over the rest, was the variable that corresponds to the probability returned by the Random Forest that a news item is true using the social information of the news item.

## 7. Conclusions and Future Work

Throughout the development of this work, it has been observed how the introduction of social information, combined with textual information, has enabled the classification of news, helping to improve the performance of the models. This suggests that, when solving a problem, it would be useful to add social information to the dataset. However, obtaining this information is quite costly both economically and in terms of time.

Additionally, the importance of social features in classifier models has been studied, concluding that author-related features are more important than tweet-related features. The development of a model that combines all textual and social features achieves similar or better results than models that use only textual information.

However, it is crucial to acknowledge several important limitations:

- **Impractical Approach:** Many of the social signals being harvested are post-facto. While disinformation might actually be spreading, many features (such as the number of reposts) would not have stabilized. Thus, while the current approach of augmenting these signals might work post-facto, it is unlikely to work with live data. Even post-facto, it is unclear whether the approach will scale.
- **Flawed Methodology:** The use of balanced training data, and a small set of data at that, is not meaningful. Particularly, it is unclear how learning from such a small corpus would generalize when new kinds of disinformation arise. In practice, the distribution of disinformation-carrying articles compared to genuine ones is far from balanced. Therefore, any realistic methodology needs to incorporate the ability to handle imbalance and transferability from the learning phase. Moreover, adversary behavior might change to emulate the features of good articles or at least stray away from its current behavior, rendering the specific features used for classification obsolete.
- **Too Static and Small Dataset:** The dataset used is too static and small, and lacks adequate diversity to consider any results conclusive. A variety of distinct datasets ought to be used to determine if the ideas actually work in a more general setting.

As a line of future work, it would be a good approach not only to study the individual social metadata of each user, but also to study a social graph of the followers or followers to see the social relationships that exist between them. Additionally, the dataset should be expanded and diversified, and methods should be developed to handle imbalanced data and adapt to changing adversary behavior.

We acknowledge that this work, while preliminary, can trigger useful discussions and provides a foundation upon which more robust and scalable approaches can be built in the future.

## Acknowledgments

This work was supported by the HAMiSoN project grant CHIST-ERA-21-OSNEM-002, AEI PCI2022-135026-2 (MCIN/AEI/10.13039/501100011033 and EU “NextGenerationEU”/PRTR).

## References

- [1] P. Bharadwaj, Z. Shao, Fake news detection with semantic features and text mining, *International Journal on Natural Language Computing (IJNLC)* Vol 8 (2019).
- [2] R. K. Kaliyar, A. Goswami, P. Narang, Fakebert: Fake news detection in social media with a bert-based deep learning approach, *Multimedia tools and applications* 80 (2021) 11765–11788.

- [3] N. K. Conroy, V. L. Rubin, Y. Chen, Automatic deception detection: Methods for finding fake news, *Proceedings of the association for information science and technology* 52 (2015) 1–4.
- [4] C. Buntain, J. Golbeck, Automatically identifying fake news in popular twitter threads, in: 2017 IEEE International Conference on Smart Cloud (SmartCloud), IEEE, 2017, pp. 208–215.
- [5] M. Albahar, A hybrid model for fake news detection: Leveraging news content and user comments in fake news, *IET Information Security* 15 (2021) 169–177.
- [6] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797–806.
- [7] K. Shu, S. Wang, H. Liu, Exploiting tri-relationship for fake news detection, *arXiv preprint arXiv:1712.07709* 8 (2017).
- [8] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J. J. M. Escobar, Detection of fake news in a new corpus for the spanish language, *Journal of Intelligent & Fuzzy Systems* 36 (2019) 4869–4876.
- [9] H. Gómez-Adorno, J. P. Posadas-Durán, G. B. Enguix, C. P. Capetillo, Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task, *Procesamiento del Lenguaje Natural* 67 (2021) 223–231.
- [10] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, *Pml4dc at iclr 2020* (2020) 1–10.
- [11] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: *PML4DC at ICLR 2020*, 2020.
- [12] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, *Proceedings of the LREC, Marseille, France* (2022) 20–25.
- [13] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, *Bioinformatics* 26 (2010) 1340–1347.

# On the Categorization of Corporate Multimodal Disinformation with Large Language Models

Ana-Maria Bucur<sup>1,2,\*</sup>, Sónia Gonçalves<sup>3</sup> and Paolo Rosso<sup>2,4</sup>

<sup>1</sup>Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania

<sup>2</sup>PRHLT Research Center, Universitat Politècnica de València, Spain

<sup>3</sup>Universidad de Sevilla, Spain

<sup>4</sup>ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain

## Abstract

Disinformation is becoming more prevalent in the corporate sphere, especially as brands choose to promote their products through influencers or micro-celebrities who are perceived as reliable and impartial, but may facilitate false information. The spread of disinformation can have negative economic impacts on companies and brands, which can even affect their reputation. Artificial Intelligence can help detect false information and has become increasingly important in combating disinformation. The current work addresses the problem of characterizing multimodal disinformation targeting corporations and provides a collection of content that spreads disinformation in digital media. The content was manually annotated with information about the target (Organization, Brand, or Other) and the source (Corporate, Advertising, or Other) of the false content. We conduct comprehensive experiments to evaluate the effectiveness of state-of-the-art Unimodal and Multimodal Large Language Models in identifying the source and target of the content.

## Keywords

Corporate Multimodal Disinformation, Multimodal Large Language Models, Spanish

## 1. Introduction and Related Work

According to [1], the concept of disinformation refers to a deliberate and organized attempt to confuse or manipulate people by providing dishonest information. In the corporate sphere, disinformation is gaining more ground. It is orchestrated to persuade audiences and hold great appeal for advertisers who promote their dissemination as a lure “because it fits more easily into people’s prejudices” [2]. The issue can become even more dangerous when we consider that more and more brands choose to promote their products through influencers or micro-celebrities, which can facilitate false information [3]. These opinion leaders are perceived with high levels of reliability and impartiality, allowing them to recommend products and services on various social media platforms and generate word of mouth that brands leverage for their commercialization [4].

The spread of disinformation can be a risk to companies and brands and cause a negative economic impact [5] that can even affect their reputation. Disinformation that can impact a company’s reputation may stem from political, financial, emotional, or internal motivations, such as discontented employees [6]. Therefore, it is important for organizations to manage trusting relationships with the public. Organizations can become victims of individuals and advanced technologies with the intention to damage their reputation for twisted purposes [7] through the use of deepfakes, a new form of fake news that threatens companies, organizations, and brands [8, 9, 10]. As the reputation of organizations can be affected by the spread of disinformation, to protect the corporate image, communication officers need to be aware of strategies to combat it, such as fact-checking. Artificial Intelligence has enabled the implementation of automated approaches capable of detecting false information [11, 12], also from a multimodal perspective [13, 14, 15, 16, 17, 18].

*Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 29–39, October 20, 2024, Santiago de Compostela, Spain*

\*Corresponding author.

✉ ana-maria.bucur@drd.unibuc.ro (A. Bucur); songomgon2@alum.us.es (S. Gonçalves); proso@dsic.upv.es (P. Rosso)

ORCID 0000-0003-2433-8877 (A. Bucur); 0000-0002-5579-7761 (S. Gonçalves); 0000-0002-8922-1242 (P. Rosso)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

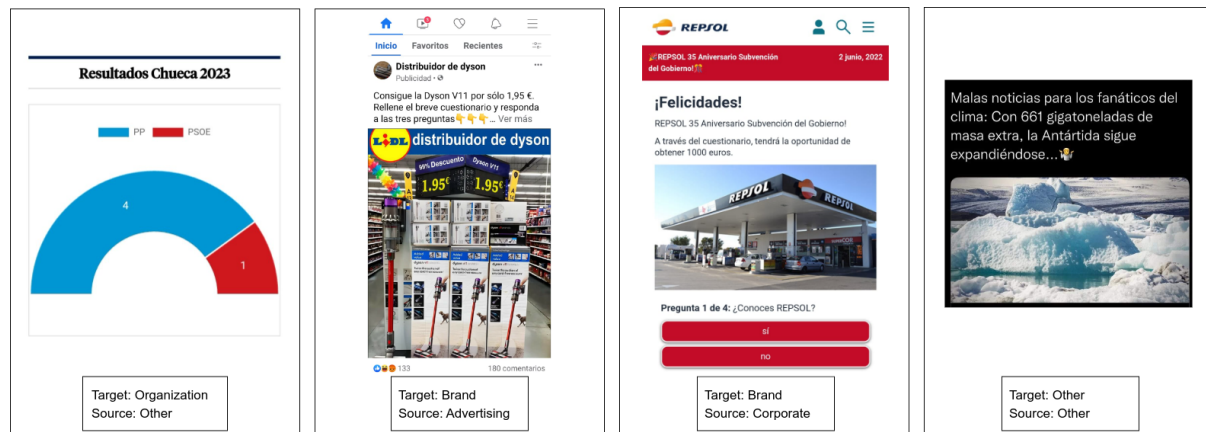


Figure 1: Selected examples of false content. The data is diverse, containing screenshots from social media, websites, etc. Translated text, first image: “Results for Chueca”. Translated text, second image: “Get Dyson V11 for only 1,95 euros. Fill in the short questionnaire and respond to the three questions...”. Translated text, third image: “Congratulations! Repsol 35th anniversary government subsidy! Through the questionnaire, you will have the opportunity to obtain 1000 euros”. Translated text, fourth image: “Bad news for the climate fanatics: with 661 gigatons of extra mass, Antarctica continues to expand...”.

Unlike general disinformation, which can target individuals, events, or broad societal issues, corporate disinformation often has direct financial implications and can damage trust in brands and organizations. Recognizing the unique characteristics and potential impacts of such disinformation, our work aims to deepen the understanding of what are the actors targeted by corporate disinformation and the sources spreading it. By classifying the target of the false content, we can identify whether the affected entity is an organization or a brand. Furthermore, identifying the source will enable affected entities to take action and develop appropriate responses to counter the disinformation being spread about them.

As there are many previous works on multimodal fake content detection [18, 14, 13, 16, 17], we aim to characterize content that has been already fact-checked and confirmed as false. To the best of our knowledge, this is the first time that the problem of multimodal disinformation targeting corporations has been addressed automatically. For this purpose, a collection of multimodal content in Spanish that was already fact-checked is collected and annotated by expert annotators with information about the target and source of the content (Figure 1). Our dataset consists of 534 samples, together with annotations for the target (Organization, Brand, or Other) and the source (Corporate, Advertising, or Other) spreading disinformation. The false content can be targeted at an Organization, such as a company, institution, or an individual representing them. It can also target a Brand or a person associated with it. Alternatively, disinformation can be classified as Other, meaning it is not aimed at an organization or brand but contains misleading information intended to deceive the general population. Furthermore, false content can originate from various sources. It may stem from a Corporate origin, where a corporate entity is responsible for spreading disinformation, rather than just an individual. Alternatively, it could be a result of persuasive Advertising, typically in the form of paid posts on social media. Lastly, false content may originate from Other sources, such as online users disseminating misleading information.

In this paper, we address the problem of characterizing multimodal disinformation targeting corporations. Our work makes the following contributions:

- A collection of multimodal false content (visual and textual information in Spanish) that spread disinformation in digital media on corporations is compiled and annotated with information about the source and target of the false content;
- Comprehensive experiments are conducted to evaluate the effectiveness of state-of-the-art Unimodal and Multimodal Large Language Models (LLMs) in characterizing false content.

## 2. Data Collection

The dataset used in this work is obtained from the IBERIFIER repository<sup>1</sup>, which includes online content that has been fact-checked and verified<sup>2</sup>. IBERIFIER is a project that aims to fight disinformation in digital media in Spain and Portugal, in which data from various fact-checking websites is collected and analyzed. In our research, we specifically focus on false content in Spanish that was verified by EFE Verifica<sup>3</sup> and Maldita.es<sup>4</sup>, as these organizations contributed the most content to the IBERIFIER database. Our dataset consists solely of posts that were confirmed by these fact-checking entities to contain false information. This limits the dataset size, as obtaining fact-checked data is challenging. Our dataset contains 496 samples from Maldita.es and 38 samples from EFE Verifica, with multimodal data represented through both visual and textual information in Spanish. By deliberately focusing on posts that have been verified to contain disinformation, we can more effectively evaluate the performance of pre-trained visual transformer models and LLMs in characterizing deceptive information. This dataset allows us to study and understand how these models identify the different targets and sources spreading disinformation. The dataset is an essential resource for studying the effectiveness of LLMs in classifying false content from visual and textual cues found in images.

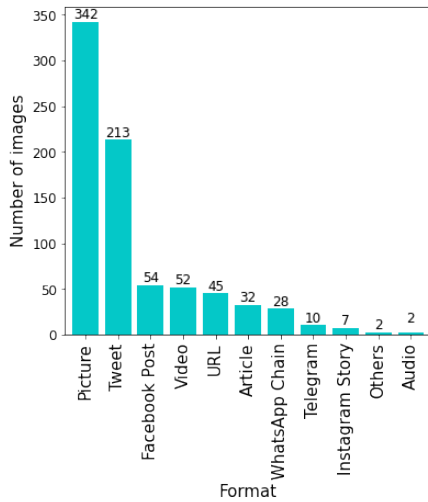


Figure 2: The format of the false content found in the collected data: pictures, screenshots from social media platforms, from different websites, or news articles.

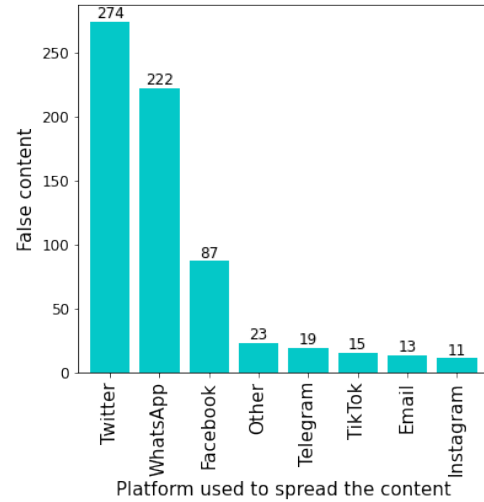


Figure 3: Platforms used to spread the false content. Most of the content was shared on social media platforms and WhatsApp.

For each of the collected images, we also retrieved information about the format of the content and the platform used to spread it using the IBERIFIER API. In Figure 2, we present the various formats of false content. The most common type of false content is represented by pictures, followed by screenshots from social media. Figure 3 shows the platforms used to spread the disinformation content. The data suggests that social media platforms like Twitter, Facebook, TikTok, and Instagram are the primary channels used to spread false content. However, we found that a considerable amount of false information is also shared through messaging apps like WhatsApp.

Two expert annotators have labeled each instance of false content with information about the target and source. The **target** of the disinformation can be an **Organization** (either a company, an institution, or a person representing it), a **Brand** (or a person representing it), or it can be **Other**, meaning that it is not targeted towards an organization or a brand, and it contains false information intending to mislead

<sup>1</sup><https://iberifier.eu/>

<sup>2</sup><https://iberifier.eu/factchecks/>

<sup>3</sup><https://verifica.efe.com/>

<sup>4</sup><https://maldita.es/>



the general population about various topics, such as climate change, immigrants, conspiracy theories, local news. With regard to the different **sources** of false content (i.e. the origin of the content), the content can be of **Corporate** origin (usually, there is an entire corporate entity behind the spread of disinformation, not just an individual), persuasive **Advertising** (usually paid posts on social media), or **Other** - usually false content spread by other users. The Other class also contains false content in which the identity of the spreader does not appear in or cannot be inferred from the image/text (see Figure 1, 1st and 4th example). We obtained a strong agreement between the two annotators (Cohen’s  $\kappa$  0.90). The disagreements between them have been resolved by a senior researcher in the field. The final dataset contains 347 samples targeting an organization, 87 targeting a brand, and 100 targeting other entities. Regarding the sources of the false content, the dataset is comprised of 52 Corporate, 4 Advertising, and 478 Other sources.

We showcase 4 examples from the collected data in Figure 1. The dataset includes different types of disinformation found in digital media, which makes it difficult to identify the source and target spreading the content. The first example shows an image with a figure representing the electoral results from the Chueca neighborhood of Madrid. However, the image is spreading disinformation because the results are actually from a municipality in Toledo with the same name. This is a classic example of how disinformation can be spread by manipulating images and providing false information. The source of the content was classified as *Other* because the origin of the information is unknown, it does not appear in the text or the image. On the other hand, the target is *Organization* because the disinformation publication affects one or more organizations, in this case, political parties (People’s Party (PP)) and Spanish Socialist Workers’ Party (PSOE)).

The second example is a sponsored post from Facebook, asking individuals to complete a brief questionnaire for the chance to purchase a discounted vacuum cleaner. However, this image represents a classic phishing post where individuals are persuaded to share their banking information with malicious entities. This example illustrates how social media platforms can be used to spread phishing scams that can deceive unsuspecting users. The source of the content was categorized as *Advertising* due to the information originating from a clearly identified advertising publication (sponsored content), indicating that the advertising is conducted on a social network through payment. Conversely, the target is identified as *Brand* because the disinformation publication impacts brands, specifically Dyson and Lidl.

The third example is a screenshot from a website that claims to be of Repsol S.A., an energy and petrochemical company from Spain. However, the website is not the real website of the company, and it is used for phishing. Malicious actors are using the website to trick users into sharing their personal data. The content was categorized as *Corporate* because the web page appears to be created by a corporate entity rather than an individual. On the other hand, the target is *Brand*, as it targets Repsol.

In the fourth example, we present a screenshot from social media that is not targeted towards a corporate entity or a brand, and it was labeled as *Other* - trying to mislead the general population. The source of the content was labeled as *Other*, with no information about the source provided in the text or image.

### 3. Methodology

We perform experiments in zero-shot or few-shot settings to evaluate the effectiveness of state-of-the-art visual transformer models and LLMs in characterizing false content within multimodal data.

#### 3.1. Pre-trained Visual Transformer Models

Pre-trained visual transformer models, such as CLIP [19], have shown great performance on downstream tasks without additional training, obtaining competitive results with a supervised baseline. CLIP was pre-trained in a self-supervised manner on a large collection of image-text pairs with a contrastive learning objective. The model was trained to maximize similarity between pairs of the same class and minimize similarity between pairs of different classes. CLIP extracts embeddings by processing the

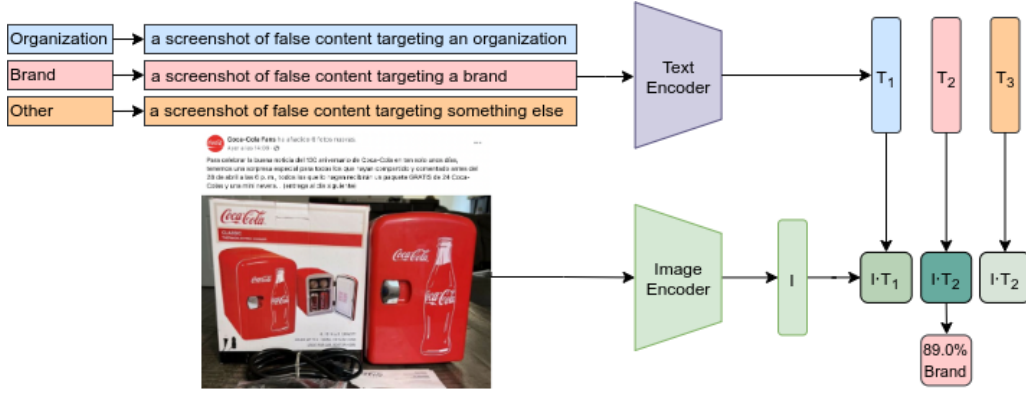


Figure 4: Zero-Shot Classification pipeline for state-of-the-art visual transformer models: CLIP, OpenCLIP, MetaCLIP, SigLIP. Images and class names/descriptions are passed through frozen encoder models, and the final prediction is represented by the text that is most similar to a given image.

image and text through a visual and textual encoder, respectively. The embeddings are then mapped to a shared space where similarities between image-text pairs can be computed. Pre-training allows CLIP to represent images and text with similar content closer in the embedding space while unrelated image-text pairs are represented further apart. In this way, the model can compute the relationship between a given image and its corresponding textual description.

We are exploring the effectiveness of using CLIP and similar models [20, 21] for zero-shot classification. To achieve this, we investigate how well the models can predict the target and the source of online disinformation. The zero-shot classification pipeline is presented in Figure 4. The process involves passing images and texts, in our case, the names/descriptions of the categories, through frozen visual and textual encoder models. The similarity between the image and each category name/description is computed, and the category with the highest similarity score is selected as the final prediction. We conducted our experiments in two settings: by providing the class names as labels and by providing a short definition/description of the content we expect to find for each class. The two types of label names, short and long, are shown in Figure 4. For target classification, we first experimented with short label names such as Organization, Brand, and Other. We also experimented with longer names, such as “a screenshot of false information targeting an organization (a company or an institution)”, etc. Inspired by recent works highlighting the importance of the definitions of the concepts [22], we added more information to the text describing the categories. For the source classification, we followed a similar approach and experimented with both the short label names, such as Corporate, Advertising, and Other, and longer variants.

In our experiments, we have tested the abilities of various pre-trained transformer models like CLIP [19], OpenCLIP [23], MetaCLIP [20], SigLIP [21]. CLIP and OpenCLIP [23] have identical vision transformer architecture, but OpenCLIP was trained on the open-source dataset LAION-2B [24], whereas CLIP was trained on a private dataset of image-text pairs. MetaCLIP [20] uses the same architecture and training regime as above, but the authors ensure that only high-quality image-text pairs are used for pre-training. SigLIP [21] replaces the softmax-based contrastive loss from CLIP with a sigmoid loss. We experiment with different variants of the models, either base, large, or huge, if available.

### 3.2. Large Language Models

With the great success of leveraging LLMs in various vision and language tasks [25, 26, 27, 28], we also choose to test their abilities in characterizing multimodal disinformation shared in digital media. We experiment with two LLMs that have shown good results in language tasks, LLaMa-2 [27], and Mistral [25]. LLaMa is a competitive model, with good results over a suite of benchmarks related to commonsense reasoning, word knowledge, reading comprehension, etc. [27]. Mistral is another LLM



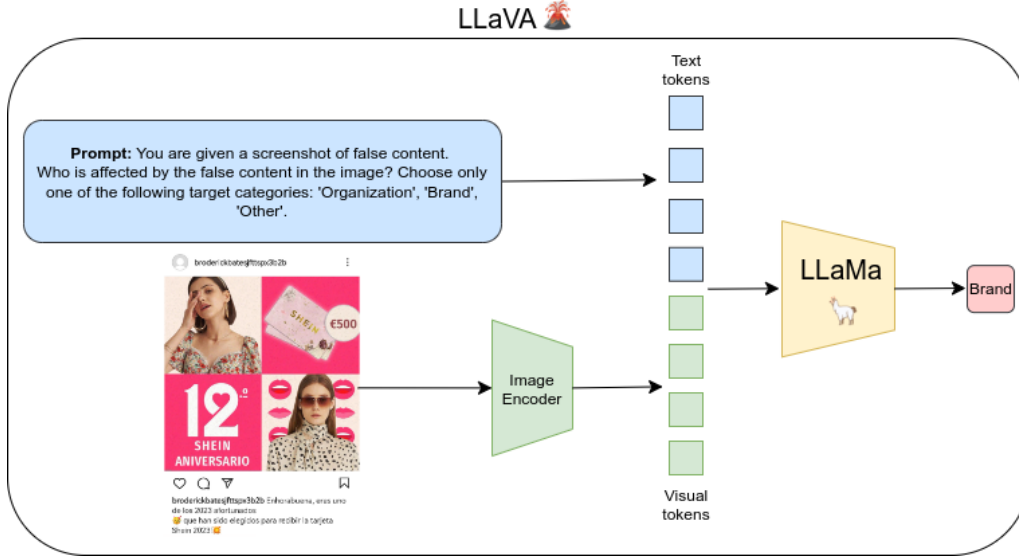


Figure 5: Zero-Shot Classification pipeline with LLaVA. LLaVa uses a language model (in our case, LLaMa) to process both visual information and language instructions, and generate an appropriate response. LLaVa leverages a pre-trained CLIP model to encode visual information from images. These embeddings are then projected into the same word embeddings space and fed into LLaMa. Finally, LLaMa generates a suitable language response.

that surpasses LLaMa-2 on all the tested benchmarks [25]. We chose these two models to evaluate their classification performance on our dataset based solely on the text found in the image and its caption. The text found in images is written in Spanish (as presented in Figure 1) and was extracted using Pytesseract<sup>5</sup>. The caption of the image was generated using BLIP-2 [29]. We conducted zero-shot and few-shot experiments using the aforementioned LLMs.

Although these LLMs are pre-trained on data that is mostly in English, LLaMa, for example, was pre-trained on 1.3B Spanish tokens (0.13% of the total corpus). This amount of pre-training tokens makes it capable of processing Spanish content, although the results may not be as accurate as for English data [30]. No information about the data used for pre-training Mistral models is available [25]. Because the text from the multimodal false content is in Spanish, we chose to include in our experiments a fine-tuned version of LLaMa-7B on Spanish instructions<sup>6</sup>.

### 3.3. Multimodal Large Language Models

In our work, we also conduct experiments using the Multimodal LLM LLaVa [31], which is a general-purpose visual and language model (Figure 5). LLaVa uses a language model (in our case, LLaMa-2 [27]) to process both the visual information from the image and the text of the language instructions. LLaVa uses a pre-trained CLIP vision transformer to process visual input, which is then projected in the same embedding space as the text. The visual and text embeddings are then fed to LLaMa, which generates a suitable language response. In our experiments we use LLaVA-v1.5 [26] and LLaVA-v1.5 Q-Instruct [28]. We chose to use LLaVA-v1.5, as it is an improved version of the original LLaVA, and it achieves state-of-the-art results on various benchmarks related to visual question answering. LLaVA-v1.5 Q-Instruct improves over the aforementioned versions by demonstrating low-level visual perception [28].

<sup>5</sup><https://github.com/madmaze/pytesseract>

<sup>6</sup>[clibrain/Llama-2-7b-ft-instruct-es](https://github.com/laion/LLaMA-2-7B-ft-instruct-es)

## 4. Experimental Setup

As part of our experiments, we tested the zero-shot and few-shot (one-shot) capabilities of various models. Our test set is comprised of 519 samples, as 15 samples were kept to potentially be used for the few-shot settings. We used the open-source implementations for all the models. Due to computational limitations, we only experimented with 7B variants of LLMs and Multimodal LLMs. While generating the output, we use the default temperature of 0.7. Additionally, we post-processed the generated output to remove any punctuation, quotation marks, or explanations generated by the models. The prompts for LLaMa-2-7B and Mistral-7B were written in English. For LLaMa-2-7B-ES, given that it is a model fine-tuned for the Spanish language, we use prompts written in Spanish.

## 5. Results

Model	Labels	Target				Source			
		Weighted-F <sub>1</sub>	Brand	Org.	Other	Weighted-F <sub>1</sub>	Adv.	Corp.	Other
CLIP <sub>base</sub>	Short	29.62	28.38	29.17	32.31	40.57	1.37	15.71	43.16
	Long	47.89	29.20	57.89	27.99	84.62	6.90	<b>48.78</b>	88.45
CLIP <sub>large</sub>	Short	32.77	25.37	36.49	25.77	43.19	1.29	6.40	43.19
	Long	49.95	27.20	59.01	<u>36.97</u>	78.97	2.50	32.52	83.83
MetaCLIP <sub>base</sub>	Short	20.80	31.81	15.10	31.80	48.73	1.22	12.32	52.49
	Long	50.53	30.38	60.03	33.71	70.51	3.28	34.21	74.35
MetaCLIP <sub>large</sub>	Short	19.46	<b>45.69</b>	8.82	35.15	83.69	2.90	26.51	89.62
	Long	14.99	26.67	7.57	31.62	80.46	2.20	<u>41.90</u>	84.56
MetaCLIP <sub>huge</sub>	Short	13.04	20.83	6.13	31.22	82.48	5.71	14.74	89.43
	Long	<u>54.34</u>	28.00	<b>66.37</b>	33.78	85.36	8.70	40.00	90.11
OpenCLIP <sub>base</sub>	Short	10.10	25.11	0.58	31.42	82.73	3.70	0.00	<u>91.14</u>
	Long	36.64	31.10	38.10	36.18	63.83	1.93	23.66	68.02
OpenCLIP <sub>large</sub>	Short	18.66	32.34	10.87	34.88	76.72	2.38	25.00	82.08
	Long	23.29	36.88	17.19	33.53	33.86	1.06	30.77	34.31
OpenCLIP <sub>huge</sub>	Short	<b>55.05*</b>	45.54	62.41	36.75	65.52	3.01	8.33	71.37
	Long	21.42	28.32	15.83	35.57	78.21	1.69	22.78	83.95
SigLIP <sub>base</sub>	Short	21.82	31.40	16.40	33.02	82.90	<b>10.53</b>	7.23	90.60
	Long	29.54	29.43	28.02	35.10	17.14	0.96	29.17	16.03
SigLIP <sub>large</sub>	Short	13.91	37.21	2.87	33.51	<b>86.18*</b>	0.00	9.52	<b>94.03</b>
	Long	51.59	30.45	59.93	<b>39.81</b>	4.16	0.96	34.01	1.26

Table 1: Zero-shot classification using visual transformer models. We present the Weighted F<sub>1</sub>-score, and the F<sub>1</sub>-scores for each of the classes. We present the best results with **bold**, and with underline the second-best results. \* denotes statistically significant differences between best and second-best models using the McNemar-Bowker Test ( $p < 0.05$ ).

We evaluate each model for the two tasks, either target or source classification, by computing F<sub>1</sub> scores for each class. We also measure the performance over each task using Weighted-F<sub>1</sub> score, given that the categories of our dataset are highly imbalanced. We present the results of the zero-shot classification using CLIP, MetaCLIP, OpenCLIP, and SigLIP in Table 1. For the majority of the models and variants, using longer descriptions of the class names improved the results of the classification. The best model for classifying the target of the false multimodal content was OpenCLIP<sub>huge</sub>, obtaining a Weighted-F<sub>1</sub> score of 55.05%. Even if SigLIP<sub>large</sub> obtained an 86.18% Weighted-F<sub>1</sub> score for predicting the source of disinformation, it cannot accurately make predictions for all the categories.

In Table 2, we showcase the performance of the LLMs in zero-shot and few-shot settings. LLaMa-2-7B, Mistral-7B and LLaMa-2-7B-ES use only the text extracted from the image and its generated caption. By providing only one example in the prompt, the performance of LLaMa-2-7B improves by 28.15%. For Mistral-7B, there is a 10.49% improvement in Weighted-F<sub>1</sub> score for target classification, while, for LLaMa-2-7B-ES, the improvement is minimal between zero-shot and few-shot settings. However, the model fine-tuned on Spanish instructions, LLaMa-2-7B-ES, obtained the best Weighted F<sub>1</sub> score of 64.01% in the few-shot setting and second-best Weighted F<sub>1</sub> score of 62.31% in the zero-shot setting.

Model	Target				Source			
	Weighted-F <sub>1</sub>	Brand	Org.	Other	Weighted-F <sub>1</sub>	Adv.	Corp.	Other
<b>LLaMa-2-7B</b> (zero-shot)	14.33	0.00	12.90	31.85	80.71	0.00	0.00	88.94
<b>LLaMa-2-7B</b> (one-shot)	42.48	22.43	50.47	31.00	72.66	<u>2.65</u>	0.00	80.05
<b>Mistral-7B</b> (zero-shot)	49.89	23.53	59.51	38.04	<b>86.98</b>	0.00	4.26	<b>95.43</b>
<b>Mistral-7B</b> (one-shot)	60.38	<b>32.00</b>	74.89	32.62	<u>86.35</u>	0.00	0.00	<u>95.15</u>
<b>LLaMa-2-7B-ES</b> (zero-shot)	<u>62.31</u>	19.23	<u>76.07</u>	<u>50.00</u>	81.81	2.38	<b>41.24</b>	86.11
<b>LLaMa-2-7B-ES</b> (one-shot)	<b>64.01*</b>	<u>24.56</u>	<b>76.41</b>	<b>53.42</b>	78.67	<b>2.96</b>	<u>41.03</u>	82.67

Table 2: Zero-shot and one-shot classification using LLMs. \* LLaMa-2-7B-ES (one-shot) obtains statistically significant improvement over the best English counterpart Mistral-7B (one-shot) in Target prediction (McNemar-Bowker Test,  $p < 0.05$ ).

Model	Target				Source			
	Weighted-F <sub>1</sub>	Brand	Org.	Other	Weighted-F <sub>1</sub>	Adv.	Corp.	Other
<b>LLaVA-v1.5-7B</b>	<b>51.88*</b>	<u>21.37</u>	<b>65.85</b>	<u>27.89</u>	<u>61.68</u>	<u>1.89</u>	<u>8.60</u>	<u>67.12</u>
<b>LLaVA-v1.5-7B (Q-Instruct)</b>	<u>49.68</u>	<b>24.84</b>	<u>60.20</u>	<b>33.22</b>	<b>68.72*</b>	<b>2.65</b>	<b>15.93</b>	<b>74.16</b>

Table 3: Zero-shot classification using LLaVA. \* denotes statistically significant differences between best and second-best models using the McNemar-Bowker Test ( $p < 0.05$ ).

Predicting the target of disinformation is easier, usually relying on specific cues, such as the presence of organizations’ or brands’ logos or names appearing in the picture or written in text. However, predicting the source of disinformation from multimodal content is a harder task, as in many instances, no information about it appears, and the source is unknown. For source classification, the LLMs sometimes only predict the *Other* class, failing to predict other categories. Using the LLaMa-2-7B-ES in one-shot setting with the text from the image and its caption as input was proven to be a suitable approach for target classification, surpassing all other visual models, such as CLIP, MetaCLIP, OpenCLIP and SigLIP. The limitations of general language models trained solely on English data are highlighted by the best performance of LLaMa-2-7B-ES, which was adapted to Spanish data. This further emphasizes the need to develop language-specialized LLMs.

In Table 3, we show the results of LLaVA-v1.5-7B for zero-shot classification. LLaVA-v1.5-7B obtains a better performance of 51.88% Weighted-F<sub>1</sub> score for target classification, while LLaVA-v1.5-7B (Q-Instruct) obtains a better performance for source classification (74.16% Weighted-F<sub>1</sub> score). In zero-shot settings, LLaVA-v1.5-7B outperforms the English-based language-only counterparts, LLaMa-2-7B and Mistral-7B, for target classification, obtaining a Weighted-F<sub>1</sub> score of 51.88%. However, it has a lower performance than LLaMa-2-7B-ES. According to our experiments, while general LLMs pre-trained on mostly English data can provide satisfactory results for identifying false content in our corporate multimodal disinformation dataset, models specifically adapted for a particular language perform better. This is because they can make use of the Spanish text present in the multimodal content, leading to enhanced performance.

## 6. Conclusion

In this paper, our aim was to create a valuable resource for characterizing corporate multimodal disinformation from digital media featuring both visual and textual elements in Spanish, annotated with details about the source and target of the false content. By publishing our dataset, we aim to encourage further research in this area and the development of more effective disinformation characterization technologies. Our comprehensive experiments have assessed the efficacy of state-of-the-art multimodal transformer models and LLMs in characterizing false content within images. Our findings reveal that predicting the target of the false content is easier than predicting the source, as the latter requires information that may not be easily represented in the multimodal data. In terms of zero-shot versus few-shot settings, providing one example for each class improved the performance for target classification by 28.15% for LLaMa-2-7B and 10.49% for Mistral-7B in terms of Weighted-F<sub>1</sub> score. LLaVA, the Multimodal

LLM that we had tested, obtained a Weighted-F<sub>1</sub> score of 51.88% in a zero-shot setting for target classification. The best result for target classification, of 64.01% Weighted-F<sub>1</sub> score, was obtained by LLaMa-2-7B-ES in one-shot setting, suggesting that LLMs specifically adapted for a particular language are needed when processing non-English data.

Our goal is to assist corporate entities in monitoring digital streams for fake news that could potentially harm their reputations. In our future work, we intend to expand our dataset and develop methods for identifying the specific brands and organizations targeted by false content. Moreover, we would like to expand our analysis to recently-released LLMs, such as LLaMa-3<sup>7</sup>, LLaVA-NeXT<sup>8</sup>, GPT-4V [32], Gemini Pro<sup>9</sup>, InstructBLIP [33].

## Limitations

One of the limitations of the current study is the small and imbalanced number of samples in each class from the collected dataset. Our approach relies on data that was already fact-checked, which is challenging to obtain. Due to the insufficient samples in some categories, our models struggle to accurately predict those classes. To address this limitation, our future work will focus on expanding the dataset. Specifically, we will target the collection of more samples for underrepresented classes, such as Brand for target classification and Corporate and Advertising for source classification.

Another limitation is the use of 7B variants of LLMs and Multimodal LLMs in our experiments due to computational limitations. Even if LLaMa-2-7B-ES and LLaVA-v1.5-7B have shown promising results of 64.01% and 51.88% Weighted-F<sub>1</sub> for source classification, using bigger variants of the models could lead to further improvements in the results [34].

## Acknowledgments

The work of Paolo Rosso was in the framework of FAKE news and HATE speech (FAKEHATE-PdC) funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR (PDC2022-133118-I00), Iberian Digital Media Observatory (IBERIFIER Plus) funded by the EC (DIGITAL-2023-DEPLOY-04) under reference 101158511, and Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence (MARTINI) funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGenerationEU/PRTR (PCI2022-135008-2).

## References

- [1] C. Ireton, J. Posetti, Journalism, fake news & disinformation: handbook for journalism education and training, Unesco Publishing, 2018.
- [2] P. Berthon, E. Treen, L. Pitt, How truthiness, fake news and post-fact endanger brands and what to do about it, NIM Marketing Intelligence Review 10 (2018) 18–23.
- [3] S. A. Baker, Alt. health influencers: how wellness culture and web culture have been weaponised to promote conspiracy theories and far-right extremism during the covid-19 pandemic, European Journal of Cultural Studies 25 (2022) 3–24.
- [4] M. De Veirman, V. Cauberghe, L. Hudders, Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude, International journal of advertising 36 (2017) 798–828.
- [5] A. Christov, et al., Economic effects of the fake news on companies and the need of new pr strategies, Journal of Sustainable Development 8 (2018) 41–49.
- [6] A. Reid, What's the damage?. measuring the impact of fake news on corporate reputation can act as a guide for companies to navigate a post-truth landscape, CommunicationDirector.com (2017).

<sup>7</sup><https://ai.meta.com/blog/meta-llama-3/>

<sup>8</sup><https://llava-vl.github.io/blog/2024-01-30-llava-next/>

<sup>9</sup><https://deepmind.google/technologies/gemini/pro/>

- [7] M. Peterson, A high-speed world with fake news: brand managers take warning, *Journal of Product & Brand Management* 29 (2020) 234–245.
- [8] W. A. Galston, Is seeing still believing? the deepfake challenge to truth in politics, *Brookings Institution* (2020).
- [9] S. Gomes-Gonçalves, Los deepfakes como una nueva forma de desinformación corporativa—una revisión de la literatura, *IROCMM: International Review of Communication and Marketing Mix*, 5 (2), 22–38. (2022).
- [10] M. Westerlund, The emergence of deepfake technology: A review, *Technology innovation management review* 9 (2019).
- [11] M. Babakar, W. Moy, The state of automated factchecking, *Full Fact* 28 (2016).
- [12] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, *Computer science review* 47 (2023) 100531.
- [13] Y. Li, B. Jiang, K. Shu, H. Liu, Toward a multilingual and multimodal data repository for covid-19 disinformation, in: *IEEE Big Data, IEEE*, 2020, pp. 4325–4330.
- [14] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, G. Jeon, Towards multimodal disinformation detection by vision-language knowledge interaction, *Information Fusion* 102 (2024) 102037.
- [15] G. Zhang, A. Giachanou, P. Rosso, Scenefnd: Multimodal fake news detection by modelling scene context information, *Journal of Information Science* (2022).
- [16] S. Tufchi, A. Yadav, T. Ahmed, A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities, *International Journal of Multimedia Information Retrieval* 12 (2023) 28.
- [17] A. Wilson, S. Wilkes, Y. Teramoto, S. Hale, Multimodal analysis of disinformation and misinformation, *Royal Society Open Science* 10 (2023) 230964.
- [18] A. Bondielli, P. Dell’Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, *CEUR Workshop Proceedings* (2023).
- [19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *Proceedings of ICML*, 2021, pp. 8748–8763.
- [20] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, C. Feichtenhofer, Demystifying clip data, in: *Proceedings of ICLR*, 2023.
- [21] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: *Proceedings of ICCV*, 2023.
- [22] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, P. Rosso, Definitions matter: Guiding gpt for multi-label classification, in: *Findings of ACL: EMNLP 2023*, 2023, pp. 4054–4063.
- [23] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, Openclip, 2021.
- [24] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, in: *Proceedings of NeurIPS*, volume 35, 2022, pp. 25278–25294.
- [25] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, *arXiv preprint arXiv:2310.06825* (2023).
- [26] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: *Proceedings of ITIF Workshop*, 2023.
- [27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, *arXiv preprint arXiv:2307.09288* (2023).
- [28] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai, et al., Q-instruct: Improving low-level visual abilities for multi-modality foundation models, *arXiv preprint arXiv:2311.06783* (2023).
- [29] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen

- image encoders and large language models, in: Proceedings of ICML, 2023.
- [30] H. Choi, Y. Yoon, S. Yoon, K. Park, How does fake news use a thumbnail? clip-based multimodal detection on the unrepresentative news image, in: Proceedings of the CONSTRAINT Workshop, 2022, pp. 86–94.
  - [31] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Proceedings of NeurIPS, 2024.
  - [32] OpenAI, Gpt-4v(ision) system card, preprint (2023).
  - [33] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. [arXiv:2305.06500](https://arxiv.org/abs/2305.06500).
  - [34] J. Lucas, A. Uchendu, M. Yamashita, J. Lee, S. Rohatgi, D. Lee, Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation, in: Proceedings of EMNLP, 2023, pp. 14279–14305.



# Automated Fact-checking based on Large Language Models: An application for the press

Bogdan Andrei Baltes<sup>1</sup>, Yudith Cardinale<sup>1</sup> and Benjamín Arroquia-Cuadros<sup>1</sup>

<sup>1</sup>*Centro de Estudios en Ciencia de Datos e Inteligencia Artificial (ESenCIA), Valencian International University, C/Pintor Sorolla 21, 46002 València, Spain*

## Abstract

The current proliferation of digital media for the dispersion of news represents advantages given the ease of access but also challenges as the different sources might not necessarily be reliable or fully consistent with each other. Existing solutions for contrasting information include knowledge bases with previously verified information that are often lacking updated information or insightful details. In this context, we propose a framework for enhancing information retrieval from the press to make information more digestible and with the ultimate goal of reducing misinformation. The proposed framework, at the interconnection of automated fact-checking, AI-based reasoning, and ethics, consists of a tool that combines information from several sources and allows users to verify a claim given the information as a knowledge base. The work explores the reasoning capabilities of Large Language Models (LLM) as a new way of automating fact-checking, creating a flexible and dynamic solution. The framework returns a verdict about the claim, as well as a justification and references, building trust for the users. The performance is rigorously evaluated achieving a score of 70% accuracy of classification and justification production for the top-performing models. Equally important, the work studies the ethical challenges of building a framework that changes the way that information from the press is consumed by society. The underlying ethics of the project are discussed both from a perspective for final users and publishing companies, offering guidance for large-scale implementation of the framework. This research poses challenges as well, mainly regarding the capabilities of current and future LLM and the commercial partnership dynamics with publishing companies.

## Keywords

Automated Fact-Checking, Large Language Models, Artificial Intelligence, Ethics

## 1. Introduction

Disinformation and fake news have been combatted through the manual work of journalists at traditional media and fact-checking outlets [1]. Tasks related to fact-checking procedures include contacting the original source via phone or e-mail, consulting alternative sources, and writing and rating the claim and publishing it [2]. While this workflow is complete and consistent, the workforce is often insufficient to monitor every piece of published information, so it is often users' mission to verify whether something they read or heard is true or false [3].

With or without deliberation to spread false information, there are often a variety of sources that are not fully consistent with each other. It is generally not possible for a person to read the same news in many different media to find complementary or contradictory information to get the full picture. Because of this, fact-checking is needed. In addition to the manual efforts of journalists, automated fact-checking (AFC) techniques are being developed mostly by nonprofit fact-checking entities [4]. The limitations of AFC have traditionally been the sensitivity to context that impedes the full automation of fact-checking systems, requiring human supervision. Another direction that AFC has been taking is that of identifying claims and constructing a database of verified claims [5], which is useful for assisting the fact-checker although with a static context.

To make information more digestible, in this work we present a framework that processes information from different sources, solves the user's original question, and indicates where the information comes from, being able to consider the context that is given to the system. An important aspect is that the user of the system is fully aware of the contents of its context and can check it if necessary, adding



trustworthiness. The aim is to help users get a broader perspective on the news from different sources, in order to fight misinformation. This is pursued through the creation of a system that compiles information from the press to be able to verify claims based on the knowledge base, providing a reasoned answer, and having the ability to reference the employed sources, supported by AI-based reasoning and ethics.

In this work, we explore a new approach for automating fact-checking: through the reasoning capabilities of Large Language Models (LLM). We carry out a complete implementation of the framework, starting from a knowledge base crafted from news from Spanish media until the interface where final users can make use of the framework. The functioning of the proposed system is tested, both from a technical and functional perspective, rigorously carrying out an evaluation achieving a score of 70% accuracy of classification and justification production, but also from an ethical standpoint, studying the underlying ethics of the change that people would undergo if the framework were implemented at a large scale and the way media is consumed were modified.

This research revolves around solving problems derived from misinformation and disinformation. The former is defined as "false or inaccurate information", while the latter is adding the notion of the "false or misleading information peddled deliberately to deceive, often in pursuit of an objective" [6]. In particular, this system is intended for journalists as primary end-users. Journalists at fact-checking agencies continuously track claims made by politicians and evaluate the veracity of them, publishing the results for the general public.

The rest of the paper is organized as follows. Section 2 describes recent studies on automated fact-checking and reasoning capabilities of LLM. The proposed Assisted Fact-Checking Framework is presented in Section 3 and the obtained results are discussed in Section 4. The ethics of the proposed work are studied in Section 5 and we present the conclusions in Section 6.

## 2. Related Work

In this section, some studies related to fact-checking and reasoning capabilities of LLM are described.

Fact-checking, in its simplest form, is a practice that verifies whether a claim is true or not. It has, of course, more complex definitions, given that a claim can be technically true but written in a misleading way, or only partially true. The most common workflow when doing the task of fact-checking is searching through multiple sources that can be used to verify the veracity of the claim, assess their reliability, and make a decision on the original claim based on the evidence found in the sources [7].

Traditionally, it consists of manual work carried out by journalists, whether to fact-check published claims in other agencies' work or to assess the correctness of works before being published [2].

The need to automatise fact-checking processes arises from the inability of journalists to verify everything they publish, since this manual work is oftentimes a task that can take up to several days [2]. Sources are not always accessible in literature or on the Internet. While there are official databases or reports such as the National Institute of Statistics (INE) in Spain, there is also manual work to be done when information needs to be verified directly calling an institution, like the Government or the Police.

In the last years, the AI community has dedicated efforts to discuss AFC. The most widely accepted structure for this automation was proposed by Vlachos and Riedel [5] [8], in a sequential process that starts by identifying the claims that need to be checked, looking through sources for the evidence needed to support or refute the claims, and taking a decision considering the given evidence.

There are, however, two issues with the knowledge commonly available to most approaches found in the literature [4]: not all available information is trustworthy, and not all needed information is available.

To overcome these problems, researchers have taken the assumptions that the information included in the employed sources is correct and that the evidence is the information that can be retrieved from there. As the evidence is assumed to be correct, veracity will be defined as the coherence of the claim and the evidence.

This common structure for automated fact-checking can - and should - be adapted to the needs of its

end users (mostly journalists). Regarding the research of these systems, there has sometimes been a lack of collaboration between researchers and journalists [9]. A better collaboration could lead to the solution of some of the issues that AFC systems present, although not all of them can be technically solved.

Furthermore, advances in the field of Generative Artificial Intelligence and specifically LLM are contributing to the transition from simpler natural language processing (NLP) techniques to the usage of the reasoning abilities of more complex models. There have been attempts to integrate LLM in the whole framework for automated fact-checking, using it to detect claims, retrieve evidence and finally, predict a verdict and build a conclusion [10]. However, the results obtained are inferior to those of the state of the art models on datasets like FEVER [11] and WiCE [12] and further research is encouraged.

To better understand the purpose of the research of automated fact-checking, a study has shown that there are eight main intended use cases of automated fact-checking [13]. The study has analysed 100 highly-cited papers, with publication dates ranging from 1998 to 2023, with most studies being from the 2010s. These use cases are listed below, specifying the percentage of the 100 papers where the respective use case is pursued: Automated external fact-checking (22%), Assisted external fact-checking (18%), Assisted media consumption (8%), Scientific curiosity (8%), Assisted knowledge curation (7%), Assisted internal fact-checking (4%), Automated content moderation (4%), Truth-telling for law enforcement (1%).

On the other hand, LLM are AI systems that can process and generate text, to solve a variety of tasks, such as summarisation, translation, question answering [14].

These systems have significantly gained popularity over the recent years. One of the main reasons of this rise was the introduction of the Transformer architecture [15]. This technical breakthrough, along with the ever-growing data collection and generation for training, as well as larger computational abilities, triggered a large wave of more capable language models. The paradigm of their creation shifted from task-specific to task-agnostic training, allowing models to perform a wider range of tasks [16].

One of the desired capabilities of LLM is reasoning. It is a cognitive process designed as the process of thinking about something in order to make a decision. At the intersection of psychology, philosophy and computer science, it is a process that benefits individuals to solve problems and take decisions [17].

Language models have a good performance on specific reasoning tasks, although there is no general agreement on whether or not they have the ability to reason [17]. It has been demonstrated, however, that these models' ability to reason improves considerably with their parameter count. Given this, recently released LLM with over 100 billion parameters are better at reasoning [18].

Performance on reasoning tasks, however, is not only a matter of parameter count. It can be heavily improved through multiple methods, which are commonly classified as [19]:

- **Strategy Enhanced Reasoning.** As LLM usually contain implicit knowledge for reasoning from their pretraining [20], the focus in this method is how to take advantage of this knowledge. The main research area is prompt engineering, which defines how to construct the questions that are fed to the models. It can be single-stage or multi-stage, the latter emulating human reasoning, decomposing a complex problem and reasoning stage by stage. Both cases are also improved by the Chain-of-Thought (CoT) method [21], which generates a series of intermediate reasoning steps by providing demonstrations on the thought process inside the prompt. Other efforts towards Strategy Enhanced Reasoning include Process Optimization [22] and External Engines [23, 24].
- **Knowledge Enhanced Reasoning.** These methods focus on how to use both implicit and explicit knowledge to assist the model in reasoning. Regarding the implicit knowledge, there has been work to take advantage of the implicit knowledge contained in LLM to generate more knowledge and refine results [25]. As for explicit knowledge, efforts have been directed towards reducing hallucinations (the invention of incorrect facts) [26] and improving information retrieval from external files [27].

It is noteworthy to mention that the answer to better reasoning is not necessarily found with more training parameters. Recent research is also focused on smaller models, easier to use in production

environments, using explanations from bigger LLM to become better reasoners [28] [29].

### 3. Assisted Fact-Checking Framework

The framework proposed aims at assessing the improvement of retrieval and consumption of information from the press, in an attempt to improve fact-checking processes and reduce misinformation through machine learning techniques. Hence, the selected narrative of this work is that of assisted media consumption or assisted fact-checking. As seen in the literature review, these use cases contribute to around 30% of the intended uses of automated fact-checking tools [13]. It is important to have a human interaction in the models without it being fully automated, since some steps in fact-checking sometimes need to be done manually (e.g., calling an official source at a Ministry to verify a fact, which cannot be done online).

The proposed framework, illustrated in Figure 1, starts with building a knowledge base of digital media according to the interests of users: the sources and categories of articles of choice. This flexibility allows the framework to be versatile, as it can be used for any type of fact-checking, with information from the press, official documents or any private document base. For the implementation of this work, several pieces on unemployment from Spanish digital media were used as a knowledge base. We recommend that whenever this framework is used for political fact-checking, the knowledge base should ideally consist of a choice of media agencies having different types of audiences [30], to make sure the contents are diverse and can complement or contrast each other.

A machine learning model, specifically an LLM in this case, is used by a data actor (mostly journalists) to verify an input, having the knowledge base as a context. The output of the proposed framework consists of a classification of the given input, as well as a justification with citations to the sources supporting or refuting it. Prompt engineering and retrieval techniques are used to control the behaviour of the language model, in an effort to restrict its context to the given knowledge base without hallucinating information and giving false information to the user [31].

As for the evaluation of the performance of this framework, traditional benchmarks are not useful since the accuracy of the responses are not commonsense reasoning capabilities but depend on specific information from the context. Moreover, human evaluation has shown reproducibility limitations and instability towards the execution of NLP tasks [32]. Hence, the approach shifts from the usage of traditional benchmarks to the evaluation through elements inspired by the LLM-as-a-judge method [33]. In this case, given a fixed knowledge base, an LLM creates a series of potential claims given its context, as well as their classification (supported by the context, refuted by it, or with no information) and their justification. This serves as an evaluation dataset that needs to be manually revised and then utilised to extract performance metrics from the behaviour of the framework, with the same context and several combinations of prompts.

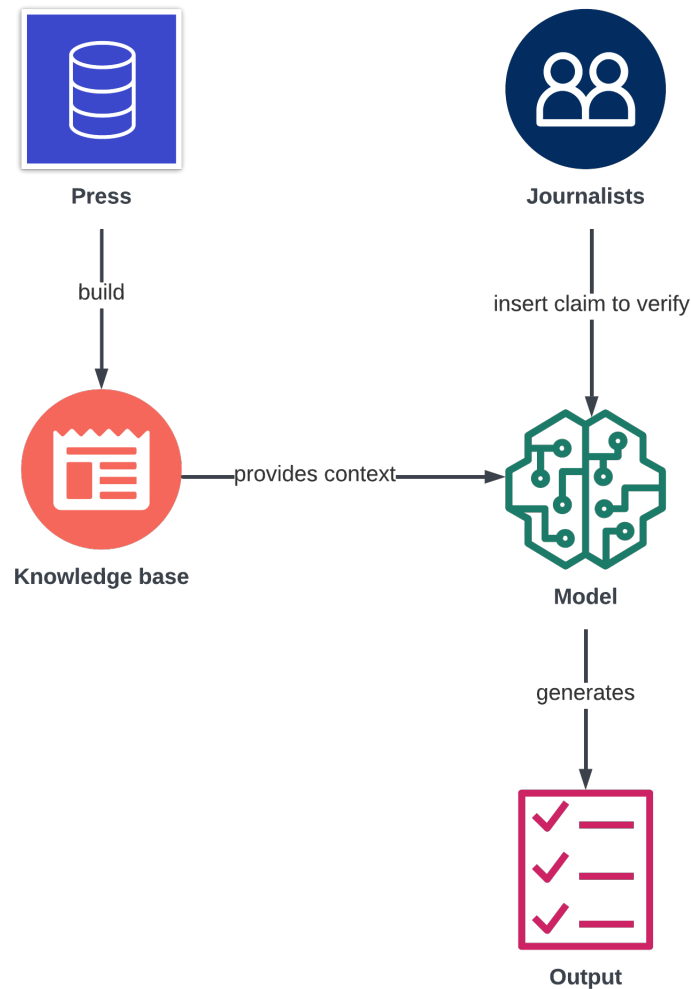
The implementation of the framework was done through LangChain, an open-source framework aimed at developing applications with LLM. Through components of this framework, LLM from OpenAI (gpt-3.5-turbo-1106, gpt-4-0125-preview<sup>1</sup>) and Cohere (Command<sup>2</sup>) were integrated. The embeddings used for this work were also from OpenAI (text-embedding-ada-002) and the LLM were used through API calls to providers offering them at no cost or at a limited one. To perform the evaluation, data consist on several pieces of information on unemployment data from Spanish digital media, in the Spanish language, from the following sites: El Plural, ABC, El Mundo, and Okdiario. These data were stored in an open-source vector database: Faiss. Lastly, the results of the framework were shown through the Gradio interface.

Several techniques were combined to improve the prompt composition [34]:

- Specifying the role: "You are a fact-checker."
- Explicitly asking to only use knowledge from the context provided.

<sup>1</sup><https://platform.openai.com/docs/models/>

<sup>2</sup><https://docs.cohere.com/docs/models>



**Figure 1: Framework for Assisted Fact-Checking from the Press.**

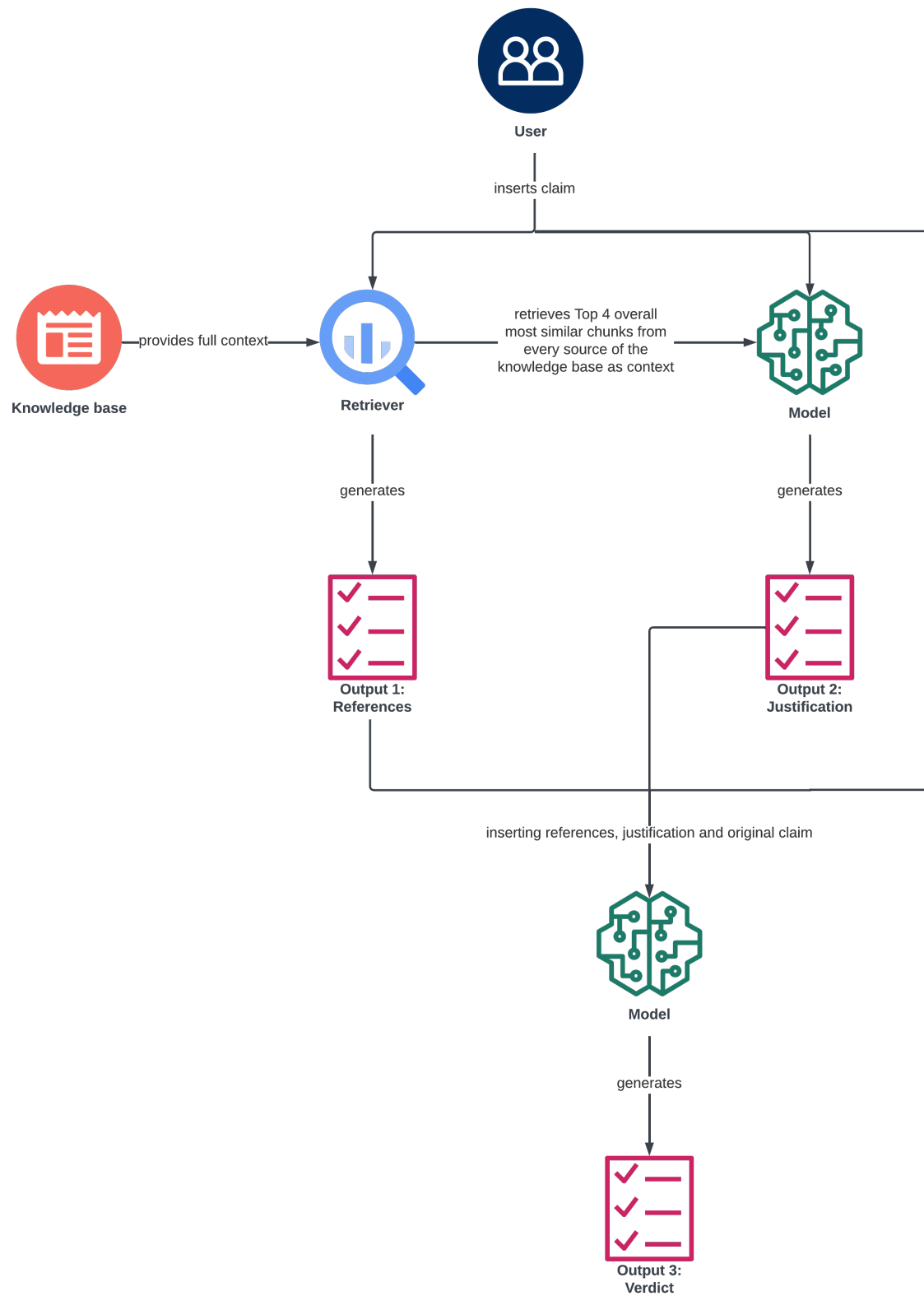
- Explaining the format of the desired output: verdict, justification, and passages of each of the sources either supporting or refuting the given claim.
- Chain-of-Thought [21]: providing an example of how a claim can be verified.

The workflow of the implementation is depicted in Figure 2, through the diverse parts of the described process, leading to the different parts of the output being generated.

The model is invoked twice. In the first call to the model, the prompt only contains instructions about the justification. Specifically, it is told that it is a fact-checker that can only base its answers on the provided context. Then, a description of each of the categories that the system needs to classify the claim into is depicted, however, it is only asked to provide the justification. It is given also an example of how it should work. Finally, the Markdown format to return is specified, and the model is once again reminded to not produce anything that is not in the given context.

Next, the model is invoked for a second time. In this call, it is only asked to create a classification based on the justification from the response of the first call. The prompt once again explains the different values that the categories can have, and the output format (in Markdown) is specified.

As there is no single ground truth for the use case of this work, there is no standard traditional machine learning evaluation method. However, besides the qualitative evaluation, which serves only to identify whether some specific examples were functioning correctly, an additional evaluation is needed to assess the general behaviour of the framework.



**Figure 2: Workflow for the Final Implementation.**

The evaluation was carried out in a mixed approach: automated and manual. The automated stages were the evaluation dataset creation and initial classification grading. On the other side, the classification adjustment and the justification grading were done manually.

GPT-4 was used to generate an evaluation dataset. Iteratively, each of the media pieces was passed to the model with a prompt asking to generate 40 claims from each of the news pieces. The prompt also specified the the claim can have - Supported, Refuted, Half-supported and No information - and a description of each of them, with the demand to classify the generated claim as well. Additionally, it was specified that there should be an equal number of each of the categories to not have unbalanced classes. The output format was demanded to be a Python dictionary in order to use it to create a dataframe afterwards, with three columns: claim, classification, and source (to keep traceability of where the claim was generated from).

After the generation of the evaluation dataset, each model (gpt-3.5-turbo-1106, gpt-4-0125-preview, Cohere Command) was invoked with the same prompt that is used in the final implementation, feeding it as a claim each generation of the evaluation dataset at a time. Therefore, a loop was created, iterating over the 160 claims that were generated for evaluation in total. The method then saved the results of each invocation, appending a column in each row with the verdict (or classification), justification, and the generated references. The results were afterwards opened in a Google Sheets file, where the automatic classification grading was done. A function was implemented to compare the column of the classification created as part of the evaluation dataset, with the one that was extracted from the output through a Google Sheets function. This would result, for each row, in a score of 0 or 1, with the latter being an exact match.

Moving to the manual stage, each row from the 480 generated in total (160 per model) was manually revised to find incorrect classifications from the evaluation dataset - changing the ground truth to an adjusted, correct version - and to find incorrect classifications of each model that could be accepted upon revision, if the justification was supporting it. The accuracy metric is created by the sum of the column of the classification adjustment, dividing by the number of rows evaluated (initially, 160). This would return a result between 0 and 1, later presented as percentage.

The final part of the evaluation is the justification grading. Each justification was graded with an answer correctness metric, assigned a score from 0 to 5. Score 0 corresponds to an output where none of the justification is correct, or it is classified as "No information" even though there exists relevant information in the context to provide a classification, while Score 5 is a justification that is entirely correct.

Justification grading was done manually following the guidance of the criteria above. To avoid inconsistencies, two rounds of grading were conducted, on different days, shuffling the order of the claims in the evaluation dataset. Afterwards, the scores of both rounds were compared, in case any of the claims were graded differently, and a final decision was taken in those cases where the scores varied. This methodology added rigorousness to an otherwise potentially subjective evaluation process.

At the end, after the grading, the score is divided by 5 in order to be a score between 0 and 1. The column of the normalised justification grading is added up and divided by the number of rows that are being evaluated (once again, initially, 160), so the resulting score for the answer correctness metric is also assigned a result between 0 and 1, later presented as percentage. It was deemed important to use both metrics, since both evaluate important functions of the framework that might be used independently.

## 4. Results and Discussion

The final results are shown in Table 1. Results show that the model with the best performance in terms of average of both metrics of classification and justification score is GPT-3.5, scoring over 70% in both metrics. Perhaps counter-intuitively, GPT-4 achieves significantly lower performance in terms of claim classification, although it achieves the top results in terms of justification. Lastly, the justification score of Cohere Command could not be calculated since it had issues with the justification language and format.

The evaluation dataset consisted initially of 160 claims and it underwent a manual revision of quality of generated claims and their classification by the judge model. As described in Section 3, this dataset was



**Table 1**

Quantitative results from evaluation.

Model	Classification Score (%)	Justification Score (%)
GPT-3.5	<b>71.70</b>	70.82
GPT-4	63.52	<b>73.58</b>
Cohere Command	44.65	-

generated by GPT-4, creating balanced categories of classifications (Supported, Refuted, Half-supported and No information) from each of the four sources (El Plural, ABC, El Mundo, and Okdiario).

As seen from the evaluation results, GPT-3.5 is the best-performing model among the three supported. Its strong points are that it had the best scores in classification and a close second place for justification production. Moreover, out of the 114 claims that were correctly classified, 84 of them (73.68%) also got the best score for its justification production.

Furthermore, there has been exactly one case of a claim that was incorrectly classified, but got the maximum score for the justification. It is, in fact, similar to one of the cases mentioned in the incorrect classifications from the evaluation dataset and it has to do with double negations and antonyms. More exactly, the claim was about the unemployment rate improving with regards to the one from 2022. It is probable that the justification was correctly created, but there was confusion with the model with the concept of improvement for terms like unemployment rate. The unemployment rate improved *for society* as the rate declined, but it seems that the model might have understood this phenomenon as an improvement if the rate actually increased. This example showcases the importance of considering both the classification and the justification in the usage of this proposed framework.

Additionally, there have been 17 claims rated with a score of 4 in justification production instead of the maximum of 5. Most of these claims' loss of the final point were related with inexact quantities or approximations. The output of the framework is almost correct in terms of justification, but it has been observed that in some cases, the model starts comparing several mentioned quantities as if they were completely different and not just a mere approximation. Prompting the model to behave in a specific manner when it had to do with approximations was tried in previous implementations in the experimental phase, although it did not have the expected result. If the behaviour of the framework when dealing with approximations improved and the results rated with a 4 were given a score of 5 instead, the results produced by GPT-3.5 could improve an additional 2.14% in terms of justification production, achieving a score of 72.96%. However, it is worth noting that there are cases where approximations are correctly handled. For instance, a claim that said there were 20 million employed people by the end of 2023 was correctly classified as "refuted" by the framework, since the correct number is 21.24 million and it is not a valid approximation.

Figure 3 shows an example of the functioning of the framework powered by GPT-3.5. The user input claims that the number of unemployed people increased in Spain in 2023. The answer given by the system classifies the claim as "Refuted", which is correct as 3/4 sources in the knowledge base support the contrary, whereas there is no relevant information in the fourth source. The justification is an accurate summary of the reasons why the system refutes the claim. Moreover, the references created by the system are also correct.

GPT-4 achieved short of 60% in the classification score, and more than 73%, surpassing GPT-3.5 for the justification production. These results are lower than expected for classification, as both models are from OpenAI but GPT-4 is a newer, bigger model with better results than GPT-3.5 on most benchmarks.

As seen from the contrast of its two scores, it is highlighted that it has lower classification capabilities in this current implementation. There are six claims classified as "No information", while the justification received the highest score possible, creating a faithful and complete reasoning on the claim. It is unknown why this behaviour occurs, since in the final implementation the methodology was changed. Instead of allowing the model to decide directly in the first invocation the classification of a claim, it had to do it in a second call, based only on the justification it previously created. Therefore, these cases should be reduced with this implementation.



**Verificando información desde la prensa**

**Afirmación**  
 Inserte la información que desee verificar  
 El número de parados aumentó en España en 2023

**Modelo**  
 Los modelos más potentes suelen consumir más recursos  
☒ gpt-3.5-turbo-0125 ☐ gpt-4-0125-preview  
☐ cohere

**Clear** **Submit**

**Veredicto:** Refuta  
**Justificación:** La afirmación de que el número de parados aumentó en España en 2023 es refutada por la información disponible en las fuentes abc, elplural y elmundo, las cuales indican que el número de parados disminuyó en España en 2023. Según las fuentes, el paro se redujo en 193.400 personas, situándose la tasa de desempleo en el 11,7%. Por lo tanto, la afirmación es completamente falsa de acuerdo al contexto proporcionado.

**Referencias**  
 Las siguientes fuentes no contienen información al respecto: okdiario  
 Las siguientes fuentes sí contienen información:  
 abc: El Ministerio de Economía, Comercio y Empresa ha destacado en una valoración remitida a los medios que el ritmo de crecimiento del empleo en 2023 fue "casi el triple" que el de 2022 y que el paro se redujo "más del doble que el año anterior" El Ministerio de Economía, Comercio y Empresa ha destacado en una valoración remitida a los medios que el ritmo de crecimiento del empleo en 2023 fue "casi el triple" que el de 2022 y que el paro se redujo "más del doble que el año anterior" 000 empleos y reducir en 193.400 personas el número de parados, situándose la tasa de desempleo en el 11,7% a cierre de 2023 600 personas El total de ocupados se

Figure 3: Output from implementation.

Additionally, there has been a recurring result that arose for 45 claims that impeded getting both a correct classification and a justification. After invoking the prompts, the output was "Understood, I am ready to start. Please, provide a claim and a justification" in 31 occasions, with another 14 cases returning "Verdict: No information. Justification: [Input given as justification]". After the first run of the evaluation, it had looked like an execution error, so the claims that led to this result were executed again. However, the same result was returned and no explanation was found to explain this behaviour, which leads to indicate an instability in the responses of GPT-4 with these prompts.

It is observed from the evaluation that the different models have varied performance levels. The best-performing model, GPT-3.5, achieves a score higher than 70% in both of the metrics that are evaluated. In this configuration, it is safe to say that the framework can be considered reliable when used as a tool for assisted fact-checking or media consumption, in a setting where a human checks the outputs instead of having a fully-automated environment.

One of the requirements needed for the framework was explainability, which is achieved mainly through the creation of references for each output. This is considered to be assured through the similarity search of the pieces of news given as input to the knowledge base. As the context is created through a more manual procedure, rather than given to the LLM to reason about, it is considered to be more reliable.

The creation of references is one of the strong points towards the trustworthiness of the proposed framework. However, it is worth noting that although there are procedures to avoid hallucinations or the invention of unrelated information in the output of LLM, they are not always completely avoided. This is why it is important to disclose to final users that the framework, if used for fact-checking or assisted media consumption, can be prone to occasionally produce such outputs.

Studies seem to suggest that English LLM trained at a very large scale can have almost as good results in other languages although there is still room for improvement [35]. This might also be the case for this research work: the performance could have been lower as the prompts and input data were designed to be used in Spanish. For instance, the confusion at reasoning might have been avoided in English. However, capabilities of language models are increasing at a considerable speed, therefore

the language shall not be an impediment for the adoption of the proposed framework as a solution for assisted fact-checking or media consumption.

Furthermore, given that the different models produce at times contradictory verdicts and reasoning, an improvement point to increase reliability could be a majority voting, or a weighted majority voting based on the evaluated performance of each method. This technique is used in well-known techniques in traditional machine learning such as Random Forests, where each of the trees have a vote and the final decision depends on the answer with most votes.

All in all, the proposed framework for fact-checking with information from the press provides a reliable solution for automating work that was traditionally manual for journalists, as well as opens new possibilities for non-professionals to consume more contrasted information from the news.

## 5. Ethical Considerations

The proposed framework is aimed to enhance the way information is consumed from the press, either with the intention of assisted fact-checking or mere media consumption in a new form. A new form of consuming information can have a considerable impact on society in case it is established. Therefore, it is important to assess ethical considerations of this proposed framework besides its implementation.

The Ethics Guidelines for Trustworthy AI developed by the European Commission in 2019 offer a systematic manner to assess the ethical considerations of the proposed framework through a set of requirements that any trustworthy AI system should meet [36]. The system proposed in this work can be systematically assessed in terms of ethics considering the following requirements from the guidelines:

- **Human agency and oversight:** In terms of human agency and autonomy, it is vital to stress that the proposed framework is a system based on information from the press, which in no case can fully assure that the information on which it relies is totally factual. Therefore, over-reliance shall be avoided. As for the concept of oversight, it is not considered an autonomous system, as it is needed to be overseen by a *Human-in-the-Loop*.
- **Technical Robustness and safety:** A low level of accuracy could create undesired results from the system. However, as it would have human supervision and reference checking, the consequences could not be damaging. The final levels of accuracy achieved through the final implementation would need to be properly communicated to end-users for them to acknowledge the behaviour and limitations of the system in order to align their expectations. Finally, in relation to the training data and assumptions that the LLM were trained on, they have not been observed to lead to adversarial effects during the experiments, mostly due to the explicit prompting to follow instructions and only rely on the data given as a context.
- **Privacy and data governance:** The framework does not use any personal data and only uses publicly available data.
- **Transparency:** There are three main elements that constitute transparency as a requirement. First, traceability is important to track; for this system, the version of the models used, the prompts with which the models are invoked, and the data that make up the knowledge base. Next, explainability is vital for building trust in the AI system, and this is achieved in this framework through the creation of a justification and the references used for that purpose. The last element of transparency is communication. It is clearly communicated that the framework is an AI system and not a human, as well as its benefits, limitations, potential risks, level of accuracy, and error rates. For the sake of transparency, it is also recommended to use open-source models in an industrial implementation of this framework. This research project has only considered closed-source LLM as there was no budget allocated for hardware or API usage. However, transparency would be improved through the usage of open-source models that are presenting performances comparable GPT-3.5 and GPT-4, like models from Llama 2 from Meta<sup>3</sup> and Mixtral from Mistral

<sup>3</sup><https://llama.meta.com/llama2>

AI<sup>4</sup>.

- Diversity, non-discrimination, and fairness: The intention when implementing this work is to always avoid unfair bias. This has been done by carefully crafting the prompts in order to leave out expressions that could leave room for subjectivity. However, it is necessary to include a disclaimer about biases that could already exist. These biases can either be in the data from the press - as the system needs to be faithful to that information - or in the training data of the LLM, although this is less common as the instructions are clearly defined to not use data from the training. Moreover, as it is advisable to consult stakeholders affected by the AI system throughout the whole life cycle, experts in fact-checking from a Spanish fact-checking start-up were consulted about functional feedback regarding the feedback. This was done in order to ensure that the system's design and development was taking into account the actual needs of professionals that could benefit from this system.
- Societal and environmental well-being: The implementation of the proposed framework, if used at a large scale, could have impact on human work and society at large. It would have the potential to change some aspects from journalism, specifically fact-checking, as evidence retrieval would be faster and journalists and fact-checkers could benefit from the time saved to invest in other tasks. On the other side, the usage of this framework for assisted fact-checking could impact favourably society at large by the reduction of misinformation and disinformation. However, it would also pose a challenge: learning a new way to digest information, as facts and claims would already reach people with a justification created, and critical thinking could decrease.
- Accountability: The functioning of the framework is documented and it can be externally audited. Moreover, it is also well-communicated to end-users about the limitations and data sources of the system, since the framework ultimately verifies if a claim is supported or not by the context provided, not if it is factually true or false. Therefore, the responsibility of the accuracy of the information falls under the data sources.

Through this assessment of requirements, it can be concluded that the proposed framework can be considered a responsible application of AI.

The other ethical aspect that needs to be studied prior to any deployment of the proposed framework in a production environment is where the data come from: What should the process to collect news pieces look like? How should the digital outlets be picked in this regard?

The New York Times, one of the longest-running newspapers in the United States, sued OpenAI in December 2023 over content created by ChatGPT [37]. The lawsuit informed of several issues with this content that the newspaper claimed were hurting the brand and functioning of the New York Times. The two most pressing issues were the regurgitation of full articles from the New York Times that ChatGPT would perform if prompted correctly, and its hallucinations where false or inaccurate information was created and then attributed them to the New York Times. Both of these problems can affect the image and the finances of the publisher.

The proposed framework pursues a new way of consuming information, joining several sources to provide a complete picture of the context of a claim users wish to consult. This is always done attributing the original authors, displaying each of the sources' positions on the claim to be checked. Although credit is important, it might not be enough. It is important to know where the information comes from and provide a way to consult the source data directly in case it is needed. However, in case the framework is adopted at large scale, it could reduce the traffic on the information sources' websites.

For all of this, it is considered that the framework, rather than a framework with direct web-scraping, should be proposed as a collaboration with diverse digital newspapers or media outlets.

---

<sup>4</sup><https://mistral.ai/technology/#models>

## 6. Conclusions

In this work, we have studied the interconnection between AI, journalism, and ethics in an attempt to create a framework whose ultimate goal is to reduce misinformation in society. The proposed framework is powered by reasoning capabilities of LLM, as it allows users to contrast a given claim with a previously built knowledge base, based on sources of interest. The claim is classified based on its alignment with the knowledge base, and a justification and references are also returned.

The functioning of the framework is evaluated with a mix of automated and manual techniques, ultimately returning a percentage of classification and justification accuracy. The best-performing model out of the several that have been studied - GPT-3.5 - scored over 70% in both metrics.

It is important to consider all parts of the output - verdict, justification, and references - when using the system, since there are cases when only some of the parts are correct. However, even given this limitation, the tool can serve as a companion for professionals and non-professionals when consuming information. The proposed framework can reliably be used as a tool for assisted fact-checking or assisted media consumption.

Overall, the broad implication of the present research work is that it is possible to use an AI-based framework to enhance the retrieval of information from the press in a responsible manner, showing that AI may be considered a promising companion tool to journalists and non-professionals wanting to contrast information.

We are currently working on testing new state-of-the-art Large Language Models, since they have the potential to improve the current performance. New models are getting rolled out at a very fast pace and benchmark scores are improving.

We are putting focus on open-source releases, since there are models already surpassing GPT-3.5 as they are getting support of the open-source community, so they would be worth testing in the scope of the framework. Minimal code modifications are being performed, since the framework is designed to support any LLM.

We also aim at modifying current methods to increase accuracy and faithfulness. There are several methods that might increase the performance of the current framework without more powerful LLM. One of them could be the combination of the answers of several methods, either by a weighted majority-voting based on the evaluated performance, or through a third call giving the outputs of the models as a new context and letting a model decide based on that information.

## References

- [1] M. F. Çömlekçi, Why do fact-checking organizations go beyond fact-checking? a leap toward media and information literacy education, *International Journal of Communication* 16 (2022) 21.
- [2] V. Moreno-Gil, X. Ramon-Vegas, M. Mauri-Ríos, Bringing journalism back to its roots: examining fact-checking practices, methods, and challenges in the Mediterranean context, *Profesional de la información* 31 (2022).
- [3] M. Himma-Kadakas, I. Ojamets, Debunking false information: investigating journalists' fact-checking skills, *Digital journalism* 10 (2022) 866–887.
- [4] L. Graves, Understanding the Promise and Limits of Automated Fact-Checking, Reuters Institute for the Study of Journalism, University of Oxford (2018). URL: <https://api.semanticscholar.org/CorpusID:13750196>.
- [5] Z. Guo, M. Schlichtkrull, A. Vlachos, A Survey on Automated Fact-Checking, *Transactions of the Association for Computational Linguistics* 10 (2022) 178–206.
- [6] Misinformation versus disinformation, explained | The Foundation for Individual Rights and Expression, <https://www.thefire.org/research-learn/misinformation-versus-disinformation-explained>, Foundation for Individual Rights and Expression.
- [7] B. Borel, *The Chicago Guide to Fact-Checking*, University of Chicago Press, 2016.

- [8] A. Vlachos, S. Riedel, Fact Checking: Task definition and dataset construction, in: *Workshop on Language Technologies and Computational Social Science*, 2014, pp. 18–22.
- [9] L. Dierickx, C.-G. Lindén, A. L. Opdahl, Automated Fact-Checking to Support Professional Practices: Systematic Literature Review and Meta-Analysis, *International Journal of Communication* 17 (2023) 21.
- [10] M. Li, B. Peng, Z. Zhang, Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models, 2023. ArXiv:2305.14623 [cs].
- [11] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 2018, pp. 809–819.
- [12] R. Kamoi, T. Goyal, J. Diego Rodriguez, G. Durrett, WiCE: Real-world entailment for claims in Wikipedia, in: H. Bouamor, J. Pino, K. Bali (Eds.), *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 7561–7583.
- [13] M. Schlichtkrull, N. Ousidhoum, A. Vlachos, The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who, in: *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 8618–8642.
- [14] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., A survey on large language models: Applications, challenges, limitations, and practical usage, *Authorea Preprints* (2023).
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.
- [17] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1049–1065.
- [18] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, *arXiv preprint arXiv:2206.07682* (2022).
- [19] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, H. Chen, Reasoning with Language Model Prompting: A Survey, 2023. ArXiv:2212.09597 [cs].
- [20] B. Paranjape, J. Michael, M. Ghazvininejad, H. Hajishirzi, L. Zettlemoyer, Prompting contrastive explanations for commonsense reasoning tasks, in: *Findings of the Association for Computational Linguistics*, 2021, pp. 4179–4192.
- [21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* 35 (2022) 24824–24837.
- [22] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, *arXiv preprint arXiv:2203.11171* (2022).
- [23] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, A. M. Dai, Mind’s Eye: Grounded Language Model Reasoning through Simulation, in: *The Eleventh International Conference on Learning Representations*, 2023.
- [24] A. Madaan, S. Zhou, U. Alon, Y. Yang, G. Neubig, Language Models of Code are Few-Shot Commonsense Learners, in: *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 1384–1403.
- [25] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, H. Hajishirzi, Generated knowledge prompting for commonsense reasoning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 3154–3169.

- [26] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, Object Hallucination in Image Captioning, in: Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4035–4045.
- [27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: 34th International Conference on Neural Information Processing Systems, 2020.
- [28] S. Li, J. Chen, Y. Shen, Z. Chen, X. Zhang, Z. Li, H. Wang, J. Qian, B. Peng, Y. Mao, et al., Explanations from large language models make small reasoners better, arXiv preprint arXiv:2210.06726 (2022).
- [29] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, A. Awadallah, Orca: Progressive learning from complex explanation traces of gpt-4, arXiv preprint arXiv:2306.02707 (2023).
- [30] F. Guerrero-Solé, The ideology of media: Measuring the political leaning of Spanish news media through Twitter users’ interactions, *Comunicación y sociedad = Communication & Society* 35 (2022) 29–43.
- [31] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, arXiv preprint arXiv:2104.07567 (2021).
- [32] C.-H. Chiang, H.-y. Lee, Can large language models be an alternative to human evaluations?, arXiv preprint arXiv:2305.01937 (2023).
- [33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, *Advances in Neural Information Processing Systems* 36 (2024).
- [34] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).
- [35] J. Armengol-Estapé, O. d. G. Bonet, M. Melero, On the multilingual capabilities of very large-scale english language models, arXiv preprint arXiv:2108.13349 (2021).
- [36] Ethics guidelines for trustworthy AI, Publications Office of the European Union, 2019. URL: <https://data.europa.eu/doi/10.2759/346720>, directorate-General for Communications Networks, Content and Technology (European Commission) and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji.
- [37] E. Helmore, K. Paul, New York Times sues OpenAI and Microsoft for copyright infringement, *The Guardian* (2023). URL: <https://www.theguardian.com/media/2023/dec/27/new-york-times-openai-microsoft-lawsuit>.



# Analysis of Climate Change Misleading Information in TikTok

Clara Baltasar<sup>1</sup>, Sergio D'Antonio Maceiras<sup>1</sup>, Alejandro Martín<sup>1</sup> and David Camacho<sup>1</sup>

<sup>1</sup>Department of Computer System Engineering, Universidad Politécnica de Madrid, Calle de Alan Turing, 28031 Madrid, Spain

## Abstract

In today's digital landscape, social media platforms have become important areas for disseminating information, ranging from legitimate discourse to misinformation, especially on critical topics such as Climate Change. This study employs claim detection and clustering techniques to analyze misleading information within an initial dataset of videos. Initially, the study identified 5,352 videos out of a total of 8,151 that warranted further investigation. Utilizing clustering methods, it was discovered that the prevalence of misinformation was surprisingly lower than anticipated. Most of the clusters showcases content that promotes sustainability and raises environmental awareness, strengthened by corroborated information of fact-checking agency EFE Verifica. Conversely, there are two clusters that focuses on videos propagating misinformation, conspiracy theories, active discussion and debate, highlighting the necessity of consuming media with caution. Looking ahead, combating misinformation necessitates the enhancement of digital literacy and the cultivation of critical thinking skills. This research aims to leverage technology and verified information from credible organizations to identify, analyze, and mitigate the influence of misleading content on social media, thus better understanding its dynamics and reducing its adverse impacts.

## Keywords

Climate change, TikTok videos, Misleading information, Environmental issues

## 1. Introduction

The spread of climate change misinformation on TikTok has become a serious concern for both environmental scientists and social media analysts. Understanding the dynamics and impact of this phenomenon requires in-depth analysis of the most viral content on the platform. Viral content on TikTok often has several key characteristics: visually appealing, emotional and easy to understand. In the context of climate change misinformation, these elements are often used to attract audiences and spread false narratives. For example, many viral videos contain dramatic images of natural disasters or melting ice caps combined with misleading or inaccurate explanations. These videos often contain sensational and apocalyptic predictions, which can evoke strong emotional responses and encourage users to share the content widely.

A perfect example is videos falsely claiming that climate change is a fraud committed by governments or corporations for financial gain. These clips often feature conspiracy theory imagery, which is particularly appealing to viewers who are already skeptical of mainstream science. Additionally, such content creators often pretend to be tipsters, exploiting the feeling of inside information to increase their authority and attract more views.

TikTok's viral misinformation patterns are influenced by the platform's algorithm, which prioritizes content that generates high levels of engagement. Videos with lots of likes, comments and shares appear more frequently on the For You page, which exponentially increases their reach. This creates a feedback loop in which misinformation continuously reaches new users, thereby increasing its spread.

Collaboration features also play a key role in the spread of misinformation. These features allow users to directly interact with existing videos by adding their own comments or creating reaction videos.

---

*Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 54–61, October 20, 2024, Santiago de Compostela, Spain*

✉ cm.baltasar@upm.es (C. Baltasar); sergio.dantonio@upm.es (S. D. Maceiras); alejandro.martin@upm.es (A. Martín); david.camacho@upm.es (D. Camacho)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



Not only does this increase the visibility of the original content, but it also creates a sense of community and brainstorming around misinformation, which further contributes to its virality.

The widespread dissemination of climate change misinformation on TikTok has significant implications for public perception. Studies such as [1] have shown that repeated exposure to misinformation can lead to the formation of false beliefs and increased skepticism towards scientific consensus. On TikTok, where the user base skews younger, the impact is particularly concerning. Young users are still developing their understanding of complex issues like climate change and their perceptions can be heavily influenced by the content they consume on social media. Furthermore, viral misinformation can undermine efforts to promote accurate scientific information.

In this research, we undertake an analysis of misinformation around climate change in TikTok. We filtered 5,352 from a total of 8,151 videos related to conversations around climate. Then, we extracted information from the videos, such as keywords and the transcription. Then we applied clustering algorithms to identify different subtopics and conversations, which allowed us to identify groups promoting negationist theories about climate change.

The rest of this article is organised as follows. Section II presents an analysis of the state-of-the-art literature, Section III describes the methodology, Section IV the results, Section V analyses the presence of hoaxes and finally, Section VI presents a series of conclusions.

## 2. State of the art

In the digital age, social media platforms have become a prime battleground for the dissemination of information and the spread of ideas, both legitimate and otherwise. One such arena is the discussion surrounding climate change, where the impact of social media, AI and algorithmic systems has become a growing concern [2]. The proliferation of misinformation and disinformation on social media platforms has garnered significant attention from researchers, especially concerning critical issues like climate change. This section reviews some relevant contributions in this area, identifying the main trends and findings in the existing literature.

Several studies as [3], agree that fake news is an old concept; it has existed and will exist as long as publishers continue to use misleading information to promote their interests and this has been happening since before the Internet even existed. Nowadays, as [4] points out, misleading information can come in various forms, such as fake news, disinformation or misinformation, which are easily spread through social media.

Many studies have investigated the spread of misinformation on social media. [5] examined the mechanisms through which misinformation is disseminated on platforms like Twitter. They presented a language model who detects fake news spreaders on Twitter. In the same way, other studies as [6] examines the stability and evolution of network structure and discussion topics among a group of prominent climate change deniers. The findings reveal that while the climate change denial network remains stable in terms of size and core group composition, sub-groups continuously emerge and dissolve.

[7] explored the dynamics of misleading information on environmental issues on TikTok, a relatively new and rapidly growing social media platform. This study found that TikTok's short-form video content and highly engaging user interface create a fertile ground for the viral spread of both accurate and inaccurate information. This study, also emphasized the importance of science education in addressing misleading information.

Also, the purpose of the study [8], was to describe content related to climate change on TikTok. Their findings indicate that climate change is being represented on TikTok as a legitimate and anxiety provoking issue. Although only a few videos included in their sample are disinformation, these garnered millions of views. Therefore, they concluded that the presence of credible professionals is essential on platforms like TikTok to increase the chances that messages are as complete as time allows, while also being scientifically sound.

### 3. Architecture development

This section describes the methodology employed to extract, filter and analyze all extracted information from the social network used in this research: TikTok.

#### 3.1. Data collection

The initial phase of our research involved the systematic collection of data using TikTok’s research API designed for developers [9].

We meticulously extracted metadata from videos associated with a selection of environmentally relevant hashtags. These hashtags included #climatechange, #ecofriendly, #sustainability and #ecotok.

Upon securing the initial dataset, we conducted a comprehensive analysis to identify the most frequently used hashtags within these videos. This analysis led to the discovery of additional pertinent hashtags such as #zerowaste, #naturetok, #globalwarming, #climatecrisis, #savetheplanet, #ecology, #plasticfree, #sustainable, #savetheworld, #recycle, #recycling, #upcycling, #saveourplanet, #upcycle, #bethechange, #environment, #climateaction, #climateemergency, #climate, #plasticpollution, #plasticwaste, #savetheocean, #saveouroceans and #eco. This iterative process was repeated meticulously until we had compiled an exhaustive list of the most prevalent hashtags in our dataset, which are those shown above.

Subsequent to the environmental hashtag analysis, we turned our focus to the investigation of misinformation related to climate change. We identified and collected metadata associated with hashtags that propagate misinformation, such as #climatelies, #climatehoaxx, #climatehoaks, #climatelie, #climatehoax, #globalwarmingisfake, #globalwarminghoax, #globalwarmingisahoax, #carbonkleptomania, #globalcooling, #climatechangehoax, #noclimateemergency, #climatescam, #weathermanipulation, #stopglobalwarming and #globalwarmingsucks. And just like before, we repeated this iterative process until we had compiled an exhaustive list of the most prevalent hashtags that propagate misinformation in our dataset, which are those shown above.

Through this rigorous process, we accumulated a total of 8,151 video metadata entries, covering the period from January 2020 to June 2024.

#### 3.2. Data characterization in TikTok

Once all videos were retrieved, the second step was to extract the audio channel, converting from .mp4 to format .mp3, and then we used the Whisper model [10] which is a pre-trained model for automatic speech recognition and speech translation, to get the transcription from the audio of the videos. We obtained a total of 6,998 text transcriptions and we started to work with that.

The next step was to analyze the transcriptions. First, we divided the transcriptions into sentences and two pre-trained models were loaded: the SentenceTransformer model [11], for creating dense vector representations of sentences, and the KeyBERT model [12], for extracting keywords from the transcriptions.

Our aim was to do clustering to all transcriptions and to obtain the most relevant keywords of each cluster. For this, to find the optimal number of clusters for K-Means, the elbow method and silhouette scores were used, and it was found that this number was ten clusters.

Then, we applied K-Means clustering to the embeddings and when we analyzed the most relevant keywords of each cluster we obtained some keywords that didn’t make sense on the issue of climate change, like “foryourpage”, “fyp” or “for watching”, so we established them as stopwords but the results obtained were not much better. After trying several things we realized that many of the videos only had songs or phrases that were not very relevant as sound, so we decided to leave the transcriptions aside.

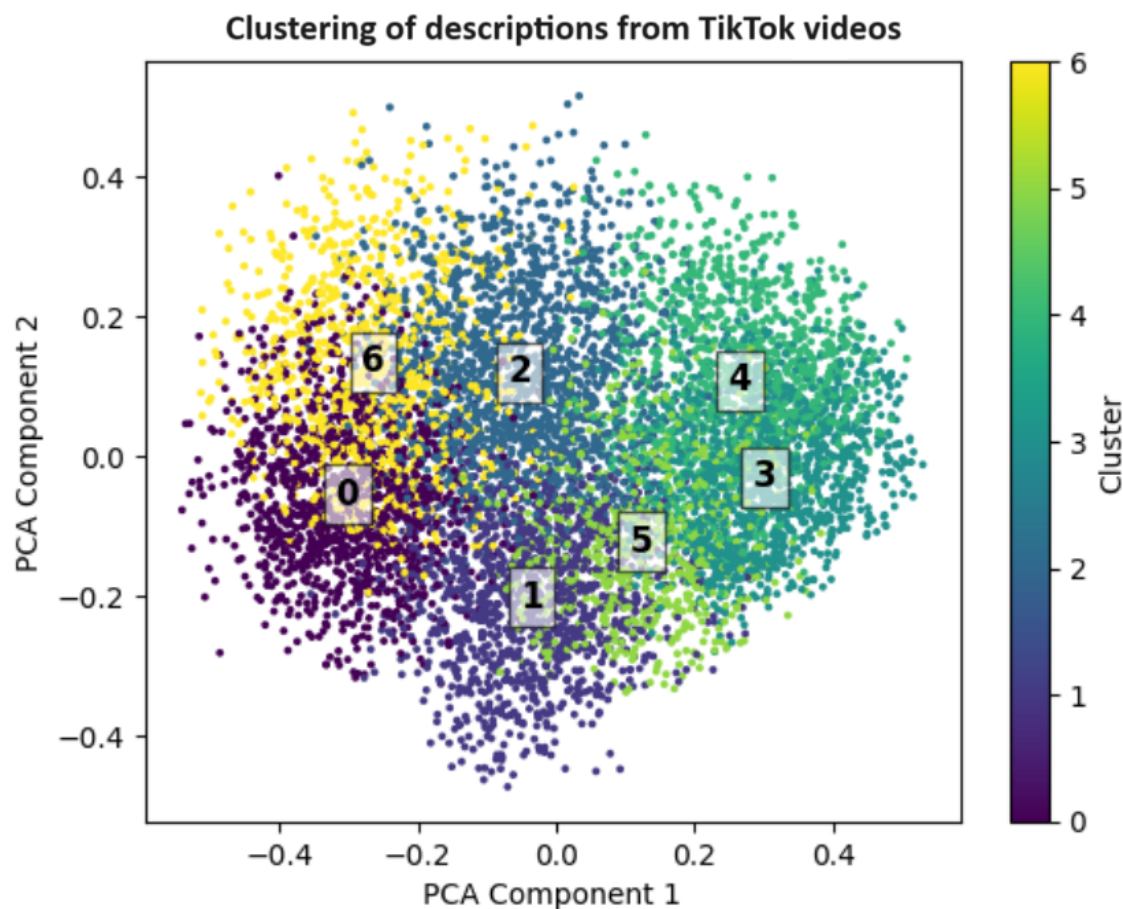
The second option was to use the “video\_description” that we obtained when we extracted metadata from videos using TikTok’s research API designed for developers [9], which owns the description of the video, and apply SentenceTransformer [11] and KeyBERT model [12] to this descriptions.

As we mentioned before, we accumulated a total of 8,151 video metadata entries, covering the period from January 2020 to June 2024, so the data collection we used in the end was larger than expected.

To find the optimal number of clusters for K-Means, the elbow method and silhouette scores were used, and it was found that this number was seven clusters. Then, we applied K-Means clustering to the descriptions of the videos and we obtained seven clusters well-defined, which are shown in the section 4.

## 4. Results

The image below, Fig. 1, represents the clustering of descriptions from 8,151 TikTok videos. We employed K-means clustering and Principal Component Analysis (PCA) to visualize and understand the thematic grouping of these transcriptions. The scatter plot shows the distribution of seven clusters, differentiated by colors. This section provides a detailed analysis of these clusters, discussing the most relevant keywords, potential misinformation and aggregate statistics for likes, views and comments.



**Figure 1:** Clustering of descriptions from TikTok videos.

The clustering algorithm identified seven distinct clusters among the videos descriptions. Each cluster represents a group of videos with similar content based on their descriptions. The cluster analysis is presented below:

- **Group 0.** This cluster, represented in the scatter plot by the color purple, contains 1,260 videos. This cluster is characterized by prominent keywords like "sewing", "upcycling", "thrift", "thriftflip" and "fashion". This suggests a focus on sewing techniques, clothing upcycling and sustainable fashion practices.
- **Group 1.** It is identified in the scatter plot by the color dark blue, comprises 1,300 videos. This cluster is characterized by content related to TikTok, duets, life hacks and fashion. Common

keywords include "duet with", "duet", "lifelifehacks" and "tiktok". This indicates a focus on social interaction and entertainment.

- **Group 2.** It is represented in the scatter plot by the color blue, contains 1,665 videos. This cluster is distinguished by its focus on sustainability, environment and ecological awareness. Keywords include "sustainability", "environment", "eco", "nature" and "ecofriendly". This reflects a strong interest in sustainable practices and environmental protection.
- **Group 3.** Identified in the scatter plot by the color turquoise, comprises 1,243 videos. This cluster is characterized by discussions on weather manipulation, chemtrails and climate-related conspiracy theories. Keywords include "weathermanipulation", "chemtrails", "geoengineering", "weathermodification" and "conspiracytiktok".
- **Group 4.** This cluster, represented in the scatter plot by the color green, contains 1,139 videos. This cluster focuses on climate change, global warming, climate action and its controversies. Common keywords include "climatechange", "globalwarming", "climatecrisis", "climateaction" and "gretathunberg". Also included are terms that deny or question climate change ("climatelies", "globalwarmingisfake" and "climatehoax"). The activist "gretathunberg" is mentioned, suggesting discussions about her influence.
- **Group 5.** This cluster, identified in the scatter plot by the color light green, comprises 636 videos. This cluster is characterized by its focus on climate action, saving the planet and environmental awareness. Keywords include "stopglobalwarming", "savetheworld", "savetheplanet" and "saveouroceans".
- **Group 6.** This last cluster, represented in the scatter plot by the color yellow, contains 908 videos. This cluster is characterized by content related to recycling, waste reduction and sustainable practices. Keywords include "recycle", "recycling", "zero waste", "reuse" and "crafts".

#### 4.1. Keyword Frequency, videos count, likes, views and comments Analysis

To provide the big picture of the results, Table 1 lists the most frequent keywords for each cluster, The total number of videos in each cluster and the total likes, views and comments for videos in each cluster. This data highlights the distribution of videos across the seven thematic clusters and provides insight into the engagement levels of videos in different thematic groups.

An analysis of the keywords and content themes across clusters reveals distinct focal areas and potential areas of misinformation. Clusters like Cluster 0 and Cluster 6 emphasize practical advice and advocacy for environmental issues, showcasing keywords related to sewing, upcycling, recycling and sustainable living. For instance, Cluster 0 includes terms like "sewing", "upcycling" and "thrift" while Cluster 6 is characterized by "recycle", "recycling" and "plasticfree". These clusters tend to present more straightforward and educational content, making them less likely to contain misinformation.

In contrast, Cluster 3 stands out due to the prevalence of terms such as "weathermanipulation", "chemtrails" and "geoengineering", pointing to content related to conspiracy theories. This cluster is more likely to mix factual information with misleading claims, which can negatively influence public perception.

Also Cluster 4, despite containing terms as "climatechange", "globalwarming", "climatecrisis" and "climateaction", it also contains terms as "climatelies", "globalwarmingisfake" and "climatehoax"; which means it focuses on climate change, global warming, climate action and its controversies.

The focus on conspiracy theories contrasts with the more practical and educational nature of the content in Clusters 0 and 6.

When considering the total likes, views and comments of these clusters, significant differences emerge. Cluster 5, for instance, which includes keywords like "stopglobalwarming" and "savetheworld" has the highest number of likes (3,053,581). This suggests that videos in this cluster receive higher acceptance and engagement from users, possibly due to the urgency and global appeal of climate change messaging.

Despite having fewer likes, Cluster 6 has the highest number of views (31,976,184). This indicates that videos in this cluster may be viewed more frequently, possibly due to a growing interest in recycling and

Cluster	Keywords and Frequency	Videos Count	Likes	Views	Comments
0	sewing (37) upcycling (24) thrift (24) thriftflip (20) fashion (19)	1,260	1,160,419	11,666,224	15,321
1	duet with (59) duet (47) bethechange (28) tiktok (25) life hacks inspiration (19)	1,300	839,619	6,303,510	40,536
2	sustainability (63) environment (40) sustainable (40) ecology (37) eco (27)	1,665	1,778,118	18,963,898	39,140
3	weathermanipulation (338) chemtrails (136) weathermodification (69) weathermanipulation lexky (60) coveringthesun weathermanipulation (51)	1,243	681,355	3,707,237	57,980
4	climatechange (238) globalwarming (116) global warming (67) climate change (47) climatecrisis (36)	1,139	1,511,484	10,797,769	48,182
5	stopglobalwarming (162) savetheworld (87) savetheplanet (32) globalwarming (29) saveouroceans (18)	636	3,053,581	17,783,375	52,676
6	recycle (126) recycling (118) recycled (27) plasticfree (18) waste (15)	908	1,621,028	31,976,184	17,701

**Table 1**

Clustering analysis results

zero-waste topics. The higher view count suggests that the audience for recycling content is substantial, even if individual engagement through likes is lower.

Cluster 1, by the other hand, characterized by keywords such as "duet with", "duet" and "bethechange" highlights a significant trend in social media engagement. With 1,300 videos, Cluster 1 focuses on collaborative content, challenges and inspirational messages.

Cluster 3, which includes conspiracy-related keywords, has the highest number of comments (57,980), and Cluster 4 which also mentions "gretathunberg", has the third higher number of comments (48,182), indicating active discussion and debate. This could be due to the controversial nature of the topics, which often elicit strong opinions and engagement from viewers, as we saw in 1. In comparison, Cluster 2, which focuses on sustainability and ecology, has a lower number of comments (39,140) but still shows significant engagement, reflecting interest in environmental issues.



## 5. Disinformation in TikTok

In the digital age, the spread of misinformation, which is often mistakenly believed to be reliable, is a serious problem. As claim detection technology advances, the models for the detection of potentially misleading content have become more accurate. We used an automated claim detection model from Huggingface<sup>1</sup> flagged a significant number of videos, suggesting they were worthy of fact-checking. Specifically, 5,352 videos were initially identified as requiring review due to concerns of possible misrepresentation.

However, after further analysis using clustering techniques, we discovered an interesting finding: the actual number of videos related to disinformation and fake news was significantly lower than the original estimate, as we saw in 4.1. This difference highlights the importance of using more diverse methods to distinguish reviewable content from non-reviewable content.

Using a clustering algorithm allows us to divide tagged videos into distinct groups based on their subject content and inferred features. It's worth noting that we've divided these videos into seven groups, each revealing unique interaction patterns and thematic focus.

Group 0 emerged as a hub for sustainable fashion practices, clothing upcycling and sewing techniques. Our validation process, including consultation of corroborated information of fact-checking agency EFE Verifica [13], confirmed the absence, mostly, of misinformation associated with these keywords. Instead, it underscored a commitment to promoting responsible consumption practices. Group 1 indicates a focus on social interaction, practical tips and entertainment, which, for the most part, does not contain misinformation.

Group 2 is distinguished by its focus on sustainability, environment and ecological awareness. It was underscored a commitment to promoting responsible sustainability and ecological discourse. Group 3 is characterized by discussions on weather manipulation, chemtrails and climate-related conspiracy theories. Despite not all videos in this cluster perpetuating misinformation, those that did sparked intense debates and polarized discussions among viewers. Group 4 focuses on climate change, global warming, climate actions and its controversies. Although many of the videos are based on climate actions, there are also some other certain controversies and debates.

Group 5 is characterized by its focus on climate action, saving the planet and environmental awareness. Our validation process, confirmed the absence, mostly, of misinformation associated with these keywords. Group 6 is characterized by content related to recycling, waste reduction and sustainable practices. The commitment to promoting responsible sustainability and recycling practices was highlighted.

## 6. Conclusion

As we have seen, in the digital era, social media platforms have become critical battlegrounds where misinformation about critical issues like climate change proliferates. Our study utilized claim detection and clustering techniques to analyze a substantial dataset of videos. Initially flagging 5,352 videos for potential misinformation, we found through clustering that the actual prevalence of misinformation, particularly in Cluster 3 and some in Cluster 4, despite existing, was lower than expected.

The rest of clusters predominantly featured content promoting sustainability and environmental awareness, corroborated by information of fact-checking agency EFE Verifica [13]. In contrast, Cluster 3 and 4 contained videos with keywords associated with misinformation and conspiracy theories. Collaboration with multilingual fact-checking platforms underscored the need for vigilant media consumption practices.

In summary, our study contributes to understanding and addressing the challenges posed by misinformation in digital media, emphasizing the importance of informed media consumption and collaborative efforts in safeguarding information integrity.

---

<sup>1</sup>[https://huggingface.co/Nithiwat/xlm-roberta-base\\_claim-detection](https://huggingface.co/Nithiwat/xlm-roberta-base_claim-detection)

## Acknowledgements

This work has been funded by Grant PLEC2021-007681 (XAI-DisInfodemics), by the “European Union” or by the “European Union NextGenerationEU/PRTR”, by grant PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR”, by Calouste Gulbenkian Foundation, under the project MuseAI - Detecting and matching suspicious claims with AI and by European Comission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511).

## References

- [1] U. K. H. Ecker, S. Lewandowsky, J. Cook, et al., The psychological drivers of misinformation belief and its resistance to correction, *Nature Reviews Psychology* 1 (2022) 13–29. URL: <https://doi.org/10.1038/s44159-021-00006-y>. doi:10.1038/s44159-021-00006-y.
- [2] V. Galaz, H. Metzler, S. Daume, A. Olsson, B. Lindström, A. Marklund, Ai could create a perfect storm of climate misinformation, 2023. arXiv:2306.12807.
- [3] M. Aldwairi, A. Alwahedi, Detecting fake news in social media networks, *Procedia Computer Science* 141 (2018) 215–222. URL: <https://www.sciencedirect.com/science/article/pii/S1877050918318210>. doi:<https://doi.org/10.1016/j.procs.2018.10.171>, the 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops.
- [4] P. N. Petratos, Misinformation, disinformation, and fake news: Cyber risks to business, *Business Horizons* 64 (2021) 763–774. URL: <https://www.sciencedirect.com/science/article/pii/S000768132100135X>. doi:<https://doi.org/10.1016/j.bushor.2021.07.012>, cIBER SPECIAL ISSUE: CYBERSECURITY IN CRISIS.
- [5] J. Huertas-Tato, A. Martín, D. Camacho, Bertuit: Understanding spanish language in twitter with transformers, *Expert Systems* 40 (2023) e13404. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13404>. doi:<https://doi.org/10.1111/exsy.13404>. arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13404>.
- [6] A. Yang, Exploring the network and topic stability in climate change deniers’ disinformation network: A longitudinal study (2024).
- [7] B. Braga Pereira, S. Ha, Environmental issues on tiktok: Topics and claims of misleading information, *Journal of Baltic Science Education* 23 (2024) 131–150. doi:10.33225/jbse/24.23.131.
- [8] F. J. Basch CH, Yalamanchili B, #climate change on tiktok: A content analysis of videos, *J Community Health* 47(1) (2022) 163–167. doi:10.1007/s10900-021-01031-x.
- [9] T. Company, Tiktok api, 2024. URL: <https://developers.tiktok.com/>.
- [10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. URL: <https://arxiv.org/abs/2212.04356>. doi:10.48550/ARXIV.2212.04356.
- [11] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. URL: <https://arxiv.org/abs/1908.10084>.
- [12] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: <https://doi.org/10.5281/zenodo.4461265>. doi:10.5281/zenodo.4461265.
- [13] Agencia EFE, EFE Verifica, ??? URL: <https://efe.com/>.



# Are Misinformation Propagation Models Holistic Enough? Identifying Gaps and Needs

Raquel Rodríguez-García<sup>1</sup>, Álvaro Rodrigo<sup>1</sup> and Roberto Centeno<sup>1</sup>

<sup>1</sup>NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain

## Abstract

Misinformation has experienced increased online diffusion, mainly due to the low control of published content and low interest in fact-checking it from social media users. Many efforts have focused on misinformation-related tasks, although typically centered on one perspective, such as shared texts or users' connections. There is a lack of holistic integrations of these local and global perspectives. Misinformation propagation models allow us to simulate how misinformation spreads through social media, and they are a way to combine both of those dimensions. In this work, we present a comprehensive study of the state of the art in this task to highlight these approaches' limitations and to establish the requirements for these models to approach misinformation propagation from a more holistic perspective.

## Keywords

Rumor Propagation, Fake News, Multi-agent Systems, Epidemiological Models

## 1. Introduction

Misinformation has proven to have a perverse effect by manipulating the public through different techniques, such as appealing to their emotions or fears to foster its believability [1]. It has negatively affected democratic processes, such as the 2016 and 2020 US Elections [2, 3], and spread potentially harmful content, such as the misinformation regarding the COVID-19 pandemic [4]. Many efforts are underway to determine what distinguishes fake content from other information [5], to detect its presence [6], or what users are more susceptible [7]. At the micro level, fake news detection is addressed by analyzing the information within a message. Recent efforts exploit Large Language Models (LLMs) [8] for their enhanced performance. Other methods have explored the detection from a more rounded standpoint, exploiting characteristics from Twitter (now X) threads [9], such as the depth of the tree, or subjective metrics such as biases and credibility [10], outperforming state-of-the-art models.

At the macro or social network level, there have been efforts to detect profiles sharing misinformation [11], showing that information on user interactions improves results obtained using only user information. The detection of bots is also explored through user features and network topology [12], showing how bot formations foster high propagation rates. It has also been approached from the lens of the differing stances within communities [13]. These features, from user characteristics to network topology, prove informative for these tasks [14].

From these efforts, we notice a general lack of holistic integration. Some approaches to detect spreaders have integrated information from different levels [15], such as shared information, user profiles, and ego networks. Nonetheless, most efforts focus on disjointed perspectives, either local [8] or global [12]. Holistic integration might limit the risk misinformation poses [16], especially considering the complexity of organized campaigns. Propagation models, which allow us to simulate how misinformation disseminates online, are a way to combine both dimensions.

There are significant efforts toward modeling the users and their psychological capabilities or behaviors [17, 7], although none includes the shared information. Some approaches have considered

*Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence (CODAI), co-located with the 27th European Conference on Artificial Intelligence (ECAI), pages 62–73, October 20, 2024, Santiago de Compostela, Spain*

✉ rrodriguez@lsi.uned.es (R. Rodríguez-García); alvarory@lsi.uned.es (Á. Rodrigo); rcenteno@lsi.uned.es (R. Centeno)

🌐 <https://sites.google.com/view/nlp-uned/people/%C3%A1lvaro-rodrigo-yuste> (Á. Rodrigo);

<http://nlp.uned.es/~rcenteno/indice.php> (R. Centeno)

🆔 0009-0000-6964-5956 (R. Rodríguez-García); 0000-0002-6331-4117 (Á. Rodrigo); 0000-0001-9095-4665 (R. Centeno)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the topics of the messages, their emotion, or users' common interests [18, 19, 20], disregarding the value of the content by itself. Regarding the macro level, most efforts employ synthetic networks [21]. Other approaches have used real topologies without the matching shared information [22], although it is crucial to the diffusion [23].

As it becomes apparent, misinformation has been commonly studied from the perspective of separate signals. Although propagation models present an opportunity to connect them, there is still a lack of research. The content of the shared information plays a significant role in the diffusion [24], differing from real information [25]. Current efforts disregard this component, which seems counterintuitive, given that the users interact with a message, textual or otherwise. With this work, we aim to highlight current limitations within these models, affecting their holistic integration and exploring the requirements for proper experimental frameworks.

This paper is structured as follows. We review state-of-the-art propagation models in Section 2. In Section 3, we expose their limitations, and in Section 4, we identify the requirements for these holistic models. Finally, in Section 5, we highlight our conclusions.

## 2. State of the Art

This section reviews the state-of-the-art propagation models. We start with early approaches in Section 2.1, then continue with epidemiological-based models in Section 2.2. In Section 2.3, we introduce non-epidemiological models, while Section 2.4 covers agent-based social models.

### 2.1. Early Approaches

Based on ordinary differential equations (ODEs), epidemiological models have been extensively employed to study the diffusion of a virus within a population [26]. These models introduce one or several infected individuals into a population. The disease spreads amongst those susceptible until it has affected the whole group, or its diffusion slowly stops. These models divide the population into exclusive categories: *Susceptible*, *Infected*, and *Removed*. These are the states the users are in regarding the disease, and they give this model its name: the SIR model.

One of the first approaches to information diffusion adapts this epidemiological model [27] as an “intellectual epidemic”, initially devised for its application in Information Retrieval. This approach creates a simile between the spread of a disease and the dissemination of information. Using the concepts in the epidemiological model as an analogy, the disease is now an idea or a piece of information, and the individuals are readers waiting to come into contact with it.

Stemming from the initial epidemiological model [26], other variations were proposed, such as the Daley-Kendall (DK) or Ignorant-Spreader-Stifler (ISS) model [28], including rumor-specific concepts, such as a decay rate to symbolize the forgetting of the information or its “news value”. A later adaptation, the Maki-Thompson model [29], simplifies the former by altering the rate at which spreaders turn into stiflers.

These previous models, and others in this section, might rely on stochastic or deterministic processes. In a stochastic process, the transitions between the compartments are probabilistic (finite-state Markov Chain). In a deterministic model, transitions are expressed through differential equations. A deterministic model is simpler than a stochastic one. However, it presents some drawbacks, such as the transition rates being proportional to the population size [30] and not allowing for individual behavior or network heterogeneity. Stochastic models incorporate randomness and are also more realistic [31], at the cost of higher complexity [27].

### 2.2. Extension of Epidemiological Models

Based on the previous models, a formal definition of information diffusion we will use for these next sections corresponds with the interactions between a population of  $N$  individuals, with an underlying graph (directed or undirected)  $G = (V, E)$  for a set of vertices  $V = \{v_0, \dots, v_{N-1}\}$  and a set of edges

$E = \{e_0, \dots, e_{E-1}\}$  that connects them. A node represents a user, and the edges between the users denote the connections, either explicit (follower-followee relationships) or implicit (interaction-based). Diffusion would be measured in users' internal stance (state) regarding the information per time unit.

Many other models inspired by epidemiological diffusion have been proposed since its early approximation, adapted to rumor diffusion, such as the *Susceptible-Infected* (SI) model [32], where users carry the information forever. The *Susceptible-Infected-Susceptible* (SIS) model [33], where the population would become *Susceptible* again, reflecting that users might forget the information. Lastly, the *Susceptible-Infected-Recovered-Susceptible* (SIRS) model [34] considers the possibility of gaining immunity after going through the infection (*Recovered*), and the possibility of losing it after some time (*Susceptible*).

These models face the problem of a clear divergence between information and epidemic transmission and the complexity of the former. Information diffusion depends on many factors, such as network topology or social interactions. Epidemiological models work on the assumption of a homogeneously interacting population, which contrasts with complex social media networks, facing unexpected deviations from the results obtained in epidemic fields [35, 36]. Another shortcoming involves the compartments for the population. Individuals might not get *Infected* but rather turn *Fact Checkers* against misinformation or undergo a period of indecision. Due to these limitations, other models aim to include complex factors not directly extracted from epidemiological behaviors but inspired by their interactions.

In the *Susceptible-Exposed-Infected-Recovered* (SEIR) model [37], individuals might go through an *Exposed* state after being in contact with an *Infected* node. Some variations consider the fuzziness of a rumor and a hesitating mechanism before sharing [38], a *Skeptic* state where users never share the information received (SEIZ) [39], or a transition to a *Recovered* state (SEIZR) [40]. The *Susceptible-Known-Infected-Recovered* (SKIR) model [41] creates a state for the individuals that spread the anti-rumor, drawing inspiration from evolutionary game theory for users' behaviors. Also modeling their opinions, the *Susceptible-Positively Infected-Negatively Infected-Recovered* (SPNR) model [42] includes two different stances towards the rumor: *Positively* or *Negatively Infected*. Regarding their emotion, the Emotion-based SIS model (ESIS) [19] classifies the message into seven differently weighted classes, such as fear or happiness, thus rendering some emotions more effective for spreading.

Other more complex models consider more states, such as the SCNDR model [43], where *Susceptible* users in this model might turn *Credulous*, *Neutrals*, or *Denies*, as well as turn *Recovered*. Besides believing the information or not, individuals might share it, not act or warn other users. The ICSAR model [44] considers the states: *Ignorant*, *Carrier*, *Spreader*, *Advocate* and *Removed*. These states can be further classified based on whether their information is a rumor or the truth. While users might transition between the different states and stances, *Advocate* and *Removed* are sink states, thus reflecting how users might not be persuaded to change their opinion.

As it becomes apparent, many models have drawn inspiration from epidemiology studies. Although they have been extended to account for information diffusion particularities, they still struggle to reflect intricate behavior. Dividing the population into compartments simplifies the problem, but it faces the difficulty of reflecting complex social behavior with a discrete label. As an example, in the  $IS_1S_2C_1C_2R_1R_2$  model [45], the difference between the *Super Authoritative* and *Authoritative* or *Super Rumor Spreader* and *Rumor Spreader* states might have more to do with node qualities and network position rather than a state in a finite state machine.

### 2.3. Non-Epidemiological Models

Although epidemiological models have been extensively used in information diffusion, other mathematical models have been proposed. Some include *Independent Cascades*, the *Linear Threshold Maximization* model, or *Hawkes Processes*.

*Independent Cascades* start with a set of active nodes [46]. With each step, they might activate other surrounding inactive nodes with a set probability dependent on the connecting edge. There is only one chance for a node to activate its neighbors. This model has been used in the diffusion of information [47], showing that the dynamics can reflect those of social media [48, 49]. Other variations do not limit

influence to a one-time-only event but a window [50].

The *Linear Threshold* model [51] establishes a threshold of surrounding neighbors for the users to change their behavior. Once a node is active, it cannot be deactivated. Other variations introduce weights between the nodes to account for social dynamics [52]. Further adaptations also introduce user information and the similarity between previously shared content [53].

*Multivariate Hawkes Process* is a type of stochastic point process model characterized by its ability to self-excite. *Hawkes Point Process* model was originally proposed to investigate earthquake events [54]. These models have been used for information diffusion on social media [55] and to devise how to mitigate its effects [56].

There are other lesser-known models, such as *Push-Pull* [57], which employs a pair-wise interaction where the user shares their information to attempt to “push” or “pull” others. *Markov Chains* have also been explored for this task, both discrete-time [58], and continuous-time [59].

These previous models are more commonly *Influence Maximization* problems. Normally, a higher-level controller supervises the optimization and the simulations, so individuality is limited. Additionally, some of these problems are NP-Hard. Although there are many efforts towards its reduction and optimization [47, 50], time complexity is high.

Beyond the epidemiological analogy, other models have been proposed inspired by different naturally occurring phenomena. Such is the case of the *Energy Model* [60] and the *Forest Fire Model* [61]. The *Energy Model* is based on the physical theory of heat energy. This model alters the traditional paradigm of a binary value for the diffusion, whether the user is infected or not, and leverages a continuous range of agreement with the rumor, constituting their “energy”.

The *Forest Fire Model* [61] is influenced by the process of fire spreading in a forest. Drawing inspiration from the diverse factors that affect the formation and spread of fire, it creates a simile with social interactions. The forest density relates to users’ ego networks, and the area’s topography relates to the account activity. Further extensions allow users to receive the information without sharing it and a similarity score between them to assess their probability of sharing [18]. Although textual characteristics are included, they are used to model the users to establish similarity scores through matching keywords, not as part of the shared content.

## 2.4. Agent-Based Social Models

A more recent trend is to exploit the potential of Agent-Based Social Systems (ABSS). Most previous models assume homogeneity in user behavior, influence, or topology, which is limiting [30]. ABSS also adopts compartmental epidemiological models while solving those issues.

The SIR epidemiological model has been adapted to ABSS technologies [62]. This model considers *Infected* users might get *Cured* by realizing the rumor is fake and stop sharing. Other studies distinguish malicious and regular users and study their influence and susceptibility based on a belief system [63]. Similarly, it has been extended to account for bots and influencers with different behaviors [64], as well as time dynamics or trust measures between agents [65]. These approaches face the problem of only focusing on user-specific characteristics.

Other efforts have modeled individual processes in users’ perceptions, such as an uncertainty-based SIR model, where uncertainty is modeled through ambiguity and ignorance [66], or a cognitive-inspired model where belief is measured based on dissonance and exposure [67]. Other common social effects and theories, such as homophily or social influence, have also been studied, such as a segregation between gullibles and skeptics within the population [68], aiding the spread of a rumor, or social context based on similarity and influence propagation [69].

Social sciences have been another interesting topic of research. The Big Five model [70] has been estimated to explore user similarity [20], homophily regarding political views [21], or a trust model based on users’ identity, behavior, and relationships [71]. Game theory and decision theory have also been studied in the context of fake news [7], introducing common deception strategies to benefit from the uncertainty. Social Impact Theory has also been used for modeling rumors [72] by introducing other components such as persuasiveness or environmental bias. Lastly, echo chambers are also explored



from different levels [17]: individual, environmental, and technological. Based on their experiments, the individual level is enough to polarize the networks, but adding the other two components generates more distinct groups.

The last and more recent approach exploits the capacity of large language models (LLMs) to simulate the opinions shared [73, 74]. Each agent potentially has different individual traits, personalities, and memories, and they can engage in discussions where they can reflect on their opinions and update them as needed. This new framework allows for fully customizable and rich environments to simulate how disinformation spreads.

An advantage of these approaches is the ability to test complex social-based behavior, such as simultaneous information [75] or real discussions between the agents. Although mathematical models have been used with centralized and decentralized measures [22], ABSS is more versatile and has been studied further in this context to identify influential nodes and delay the diffusion process [76], to study the simultaneous spread of a rumor and its counterinformation [75], and other measures based on user attention [64]. The main problem in many of these studies is the lack of real data validation. When including some of these social theories, the need arises to determine information from the user that might not be easily extracted or determined. This forces the models to employ estimations or distributions, which introduce biases.

### 3. Current Limitations

Propagation simulation models have some limitations, which affect their holistic integration. We can summarize them based on the five main areas we explore below.

**Users.** Whenever users are characterized, their metrics are established through probability distributions or means that cannot be validated, partly due to their complexity and the difficulty of extracting them from real data. Some proposals also employ psychological models without contemplating that associations between social media usage and these traits are not always found [77]; they might not align with the modeled behavior, or they might vary over time.

**Content.** Users have been the main focus in this area. Few studies consider the message through incomplete dimensions [78] or to establish user similarity based on posted content [18]. As such, content within the diffusion has not been explored. There is also a very pointed focus on the dichotomy of fake and real news. A priori, information is unverified and might remain so. Focusing on the characteristics rather than the truth value seems more realistic and valuable.

**Network.** In most cases, the topologies used are synthetic or do not match the real diffusion. This makes it impossible to connect users with their characteristics and topology, although it is an essential component. Another limitation is creating a network with as many users as participants in the conversation, which already creates an implicit bias. Although it would be computationally impossible to include all the users in any social media network, only including those participating makes another issue arise: predicting when users will not participate.

**Internal state.** Most studies measure interaction based on states, which reflect an internal measure of the users participating in the diffusion. Messages are used to make an abstraction of the users' state. This also allows intermediate states to reflect user behaviors that cannot be found in the real data. This situation can be avoided by using the messages directly, reflecting diffusion more accurately since users can share more than one message, but their state would remain a bounded constant. Messages were only employed in one study [63], aggregating diffusion into zero messages, over 500, and in-between. This would suggest that 500 retweets have the same relevance as 50,000, which should not be correct.

**Evaluation.** In most cases, validation is done through empirical evaluation or the analysis of mathematical properties of the diffusion within the networks. Although mathematical properties provide a theoretical background, real complex networks are characterized by their non-trivial features, which do not appear in synthetic graphs. Regarding empirical evaluation, incomplete data is most commonly employed, which forces the issue of its validity. Some approaches have been evaluated aggregating at the time level [63], which dismisses how relevancy works in social media: 500 retweets

Dataset	Content	Temporal	Network	User	Stance	Topic
FakeNewsNet[79]	✓	✓	✗	✓	✗	Politics
Palin and Obama[80]	✗	✓	✗	✓	✓	Politics
ReCOVery[81]	✓	✓	✗	✓	✗	COVID-19
CoAID[82]	✓	✓	✗	✓	✗	COVID-19
MediaEval[11]	✓	✗	✓	✗	✓	Conspiracies
PHEME-9[83]	✓	✓	✓	✓	✓	General
SNAP[84]	✗	✗	✓	✗	✗	-

**Table 1**

Available datasets for information propagation models and their main characteristics

in 10 minutes do not equate to 500 retweets in 10 days.

## 4. Requirements for a Properly Experimental Framework

A proper experimental framework is required to overcome propagation models' limitations. Within this empirical evaluation, it is important to recreate the scenarios of a news piece's diffusion on social media. Datasets that contain the necessary information to evaluate the models are crucial. Below, we explore the most relevant information required for this process.

Information of the shared content (**Content**) and the users (**User**). The information shared is an essential part of the diffusion. This includes the initial post, external website links, or visual content. In terms of the users, since they are the main focus of these models, it is important to have enough information on user metrics and engagement to properly characterize them.

Temporal information of when the texts are shared, by whom, and which users engage with it at what given times (**Temporal**). Besides the texts, we need to know the timestamps of when those posts are shared to determine the evolution of the news: whether more information is added or corrected, as well as how many times it appears at different times. Within the simulation, it determines when users are engaging, which is crucial for the evaluation.

The social network (**Network**). This is an important element in the diffusion of content online. Although synthetic networks might reflect some properties of real social media networks, they pose a significant limitation since diffusion inherently depends on those connections. Users with millions of followers will have higher chances of broadcasting information than new users.

Posts labeled with their stance (**Stance**). This is a relevant measure to study and evaluate diffusion in terms of epidemiological-based models. Distinguishing between *Infected* and *Vaccinated* is essential, equivalent to users' stance towards a post (*Support* or *Oppose*).

After establishing these requirements, we review available datasets to determine their suitability. In Table 1, we include the most relevant ones we found and their relevant characteristics. We have excluded datasets created ad-hoc since they are not publicly available and typically require retrieving new data and those centered around topics unrelated to misinformation.

As illustrated in Table 1, most datasets focus on the **Text** and **Temporal** aspects (the tweets and timestamps), and the **User** information from the poster. Some datasets, such as MediaEval, anonymize the tweets by removing the time when tweets were posted and removing the information from the users. These features are essential to establish the diffusion of information. Regarding this type of content, the SNAP collection does not provide diffusion information; it only shares the topologies from social media networks. Although it is valuable information, the diffusion that matches the network is deemed necessary. Some other collections, such as PHEME-9, include information regarding the users' state or stance towards the information. This information is also essential to evaluate epidemiological-based models.

Only two of the listed datasets include **Network** information associated with the diffusion: PHEME-9 and MediaEval. MediaEval poses an additional problem due to its topology, created based on an interaction network. It is also significantly filtered and skewed: 3,800 tweets are associated with a network of 1.7



million nodes and 270 million edges. Based on these available resources, we can see that most current available datasets do not provide enough information for a proper evaluation. This is an important limitation and highlights the need for more publicly available content for the community to further research efforts into mitigating fake news.

## 5. Conclusions and Future Work

Current misinformation-related tasks and approaches show a clear divide between the micro level, or the content of the information, and the macro level, or the social network. There is a lack of holistic integration between the different tools to address misinformation. Propagation models are one tool that would allow a holistic approach by studying the diffusion of online misinformation from local and global perspectives.

With this work, we have studied current approaches to propagation diffusion models, from early approaches with the SIR epidemiological model to non-epidemiological models, such as the *Forest Fire Model*, and agent-based systems. From these approaches, we have appreciated some common limitations that constrain the holistic view. Within these constraints, the most relevant one is disregarding the information shared within the network, treated as a black box. To overcome these limitations, we have determined the main requirements for a proper experimental framework that would allow us to overcome them.

In terms of future work, we believe it is paramount to focus on overcoming these limitations by developing new models that consider the impact of the messages on the users. Additionally, posing new evaluation frameworks that overcome the limitations of the users' stances, such as focusing on the messages, is another interesting research avenue. Lastly, developing new publicly available datasets with the required information for these models is crucial for evaluating these models.

## Acknowledgments

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación), through the DeepInfo project PID2021-127777OB-C22 (MCIU/AEI/FEDER, UE) and the HOLISTIC ANALYSIS OF ORGANISED MISINFORMATION ACTIVITY IN SOCIAL NETWORKS project (PCI2022-135026-2).

## References

- [1] C. Martel, G. Pennycook, D. G. Rand, Reliance on Emotion Promotes Belief in Fake News, *Cognitive Research: Principles and Implications* 5 (2020) 1–20. doi:10.1186/s41235-020-00252-3.
- [2] E. Ferrara, H. Chang, E. Chen, G. Muric, J. Patel, Characterizing Social Media Manipulation in the 2020 U.S. Presidential Election, *First Monday* 25 (2020). doi:10.5210/fm.v25i11.11431.
- [3] A. Guess, B. Nyhan, J. Reifler, Selective Exposure to Misinformation: Evidence from the Consumption of Fake News during the 2016 U.S. Presidential Campaign, *European Research Council* 9 (2018) 4.
- [4] J. Y. Cuan-Baltazar, M. J. Muñoz-Perez, C. Robledo-Vega, M. F. Pérez-Zepeda, E. Soto-Vega, Misinformation of COVID-19 on the Internet: Infodemiology Study, *JMIR Public Health and Surveillance* 6 (2020) 1–9. doi:10.2196/18444.
- [5] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. doi:10.18653/v1/D17-1317.
- [6] S. Raza, C. Ding, Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach, *International Journal of Data Science and Analytics* 13 (2022) 335–362. doi:10.1007/s41060-021-00302-z.

- [7] C. Kopp, K. B. Korb, B. I. Mills, Information-Theoretic Models of Deception: Modelling Cooperation and Diffusion in Populations Exposed to "Fake News", *PLOS ONE* 13 (2018) e0207383. doi:10.1371/journal.pone.0207383.
- [8] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection, *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (2024) 22105–22113. doi:10.1609/aaai.v38i20.30214.
- [9] C. Buntain, J. Golbeck, Automatically Identifying Fake News in Popular Twitter Threads, in: 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 208–215. doi:10.1109/SmartCloud.2017.40.
- [10] P. Bazmi, M. Asadpour, A. Shakery, Multi-View Co-Attention Network for Fake News Detection by Modeling Topic-Specific User and News Source Credibility, *Information Processing & Management* 60 (2023) 103146. doi:10.1016/j.ipm.2022.103146.
- [11] K. Pogorelov, D. T. Schroeder, S. Brenner, A. Maulana, J. Langguth, Combining Tweets and Connections Graph for FakeNews Detection at MediaEval 2022, in: *MediaEval 2022*, volume 3583, 2022, pp. 1–4.
- [12] G. Caldarelli, R. De Nicola, F. Del Vigna, M. Petrocchi, F. Saracco, The Role of Bot Squads in the Political Propaganda on Twitter, *Communications Physics* 3 (2020) 1–15. doi:10.1038/s42005-020-0340-4.
- [13] K. Neha, V. Agrawal, S. Chhatani, R. Sharma, A. B. Buduru, P. Kumaraguru, Understanding Coordinated Communities through the Lens of Protest-Centric Narratives: A Case Study on #CAA Protest, in: *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 2024, pp. 1123–1133. doi:10.1609/icwsm.v18i1.31377.
- [14] K. Shu, H. R. Bernard, H. Liu, *Studying Fake News via Network Analysis: Detection and Mitigation*, Springer International Publishing, Cham, 2019, pp. 43–65. doi:10.1007/978-3-319-94105-9\_3.
- [15] S. Sharma, R. Sharma, Identifying Possible Rumor Spreaders on Twitter: A Weak Supervised Learning Approach, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8. doi:10.1109/ijcnn52387.2021.9534185.
- [16] A. Peñas, J. Deriu, R. Sharma, G. Valentin, J. Reyes-Montesinos, Holistic Analysis of Organised Misinformation Activity in Social Networks, in: *Disinformation in Open Online Media*, Springer Nature Switzerland, Cham, 2023, pp. 132–143. doi:10.1007/978-3-031-47896-3\_10.
- [17] D. Geschke, J. Lorenz, P. Holtz, The Triple-filter Bubble: Using Agent-based Modelling to Test a Meta-theoretical Framework for the Emergence of Filter Bubbles and Echo Chambers, *British Journal of Social Psychology* 58 (2019) 129–149. doi:10.1111/bjso.12286.
- [18] S. Kumar, M. Saini, M. Goel, B. S. Panda, Modeling Information Diffusion in Online Social Networks Using a Modified Forest-Fire Model, *Journal of Intelligent Information Systems* 56 (2021) 355–377. doi:10.1007/s10844-020-00623-8.
- [19] Q. Wang, Z. Lin, Y. Jin, S. Cheng, T. Yang, ESIS: Emotion-based Spreader-Ignorant-Stifler Model for Information Diffusion, *Knowledge-Based Systems* 81 (2015) 46–55. doi:10.1016/j.knosys.2015.02.006.
- [20] L. Milli, Opinion Dynamic Modeling of News Perception, *Applied Network Science* 6 (2021) 76. doi:10.1007/s41109-021-00412-4.
- [21] A. Coates, T. Muller, S. Sirur, Simulating the Impact of Personality on Fake News, in: *TRUST@AAMAS*, 2021, pp. 1–12.
- [22] A. N. Zehmakan, C. Out, S. Hesamipour Khelejan, Why Rumors Spread Fast in Social Networks, and How to Stop It, in: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 2023*, pp. 234–242. doi:10.24963/ijcai.2023/27.
- [23] M. Karnstedt, M. Rowe, J. Chan, H. Alani, C. Hayes, The Effect of User Features on Churn in Social Networks, in: *Proceedings of the 3rd International Web Science Conference, WebSci '11, Association for Computing Machinery, New York, NY, USA, 2011*, pp. 1–8. doi:10.1145/2527031.2527051.

- [24] B. Horne, S. Adali, This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News, *Proceedings of the International AAAI Conference on Web and Social Media* 11 (2017) 759–766. doi:10.1609/icwsm.v11i1.14976.
- [25] S. Vosoughi, D. Roy, S. Aral, The Spread of True and False News Online, *Science* 359 (2018) 1146–1151. doi:10.1126/science.aap9559.
- [26] W. O. Kermack, A. G. McKendrick, A Contribution to the Mathematical Theory of Epidemics, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115 (1927) 700–721. doi:10.1098/rspa.1927.0118.
- [27] W. Goffman, V. A. Newill, Generalization of Epidemic Theory: An Application to the Transmission of Ideas, *Nature* 204 (1964) 225–228. doi:10.1038/204225a0.
- [28] D. J. Daley, D. G. Kendall, Epidemics and Rumours, *Nature* 204 (1964) 1118–1118. doi:10.1038/2041118a0.
- [29] D. P. Maki, M. Thompson, *Mathematical Models and Applications*, Prentice-Hall, 1973.
- [30] D. J. Daley, D. G. Kendall, Stochastic Rumours, *IMA Journal of Applied Mathematics* 1 (1965) 42–55. doi:10.1093/imamat/1.1.42.
- [31] T. Britton, Stochastic Epidemic Models: A Survey, *Mathematical Biosciences* 225 (2010) 24–35. doi:10.1016/j.mbs.2010.01.006.
- [32] D. Shah, T. Zaman, Rumors in a Network: Who's the Culprit?, *IEEE Transactions on Information Theory* 57 (2011) 5163–5181. doi:10.1109/tit.2011.2158885.
- [33] M. Kimura, K. Saito, H. Motoda, Efficient Estimation of Influence Functions for SIS Model on Social Networks, in: *Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI'09*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2009, pp. 2046–2051. doi:10.5555/1661445.1661772.
- [34] R. Escalante, M. Odehnl, A Deterministic Mathematical Model for the Spread of Two Rumors, *Afrika Matematika* 31 (2019) 315–331. doi:10.1007/s13370-019-00726-8.
- [35] M. Nekovee, Y. Moreno, G. Bianconi, M. Marsili, Theory of Rumour Spreading in Complex Social Networks, *Physica A: Statistical Mechanics and its Applications* 374 (2007) 457–470. doi:10.1016/j.physa.2006.07.017.
- [36] R. Pastor-Satorras, A. Vespignani, Epidemic Spreading in Scale-Free Networks, *Physical Review Letters* 86 (2001) 3200–3203. doi:10.1103/PhysRevLett.86.3200.
- [37] C. Wang, K. Xu, G. Zhang, A SEIR-based Model for Virus Propagation on SNS, in: *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies, IEEE, Xi'an, 2013*, pp. 479–482. doi:10.1109/EIDWT.2013.86.
- [38] L.-L. Xia, G.-P. Jiang, B. Song, Y.-R. Song, Rumor Spreading Model Considering Hesitating Mechanism in Complex Social Networks, *Physica A: Statistical Mechanics and its Applications* 437 (2015) 295–303. doi:10.1016/j.physa.2015.05.113.
- [39] F. Jin, E. Dougherty, P. Saraf, Y. Cao, N. Ramakrishnan, Epidemiological Modeling of News and Rumors on Twitter, in: *Proceedings of the 7th Workshop on Social Network Mining and Analysis, ACM, Chicago Illinois, 2013*, pp. 1–9. doi:10.1145/2501025.2501027.
- [40] L. M. A. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, C. Castillo-Chávez, The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models, *Physica A: Statistical Mechanics and its Applications* 364 (2006) 513–536. doi:10.1016/j.physa.2005.08.083.
- [41] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, M. Xu, Rumor Propagation Dynamic Model Based on Evolutionary Game and Anti-Rumor, *Nonlinear Dynamics* 95 (2019) 523–539. doi:10.1007/s11071-018-4579-1.
- [42] Y. Bao, C. Yi, Y. Xue, Y. Dong, A New Rumor Propagation Model and Control Strategy on Social Networks, in: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 2013*, pp. 1472–1473. doi:10.1109/ASONAM.2013.6785909.
- [43] W. Hong, Z. Gao, Y. Hao, X. Li, A Novel SCNDR Rumor Propagation Model on Online Social Networks, in: *2015 IEEE International Conference on Consumer Electronics - Taiwan, IEEE, 2015*, pp. 154–155. doi:10.1109/ICCE-TW.2015.7216829.
- [44] N. Zhang, H. Huang, B. Su, J. Zhao, B. Zhang, Dynamic 8-State ICSAR Rumor Propagation Model

- Considering Official Rumor Refutation, *Physica A: Statistical Mechanics and its Applications* 415 (2014) 333–346. doi:10.1016/j.physa.2014.07.023.
- [45] Y. Zhang, Y. Su, L. Weigang, H. Liu, Rumor and Authoritative Information Propagation Model Considering Super Spreading in Complex Social Networks, *Physica A: Statistical Mechanics and its Applications* 506 (2018) 395–411. doi:10.1016/j.physa.2018.04.082.
- [46] D. Kempe, J. Kleinberg, É. 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*, ACM, Washington, D.C., 2003, pp. 137–146. doi:10.1145/956750.956769.
- [47] N. P. Nguyen, G. Yan, M. T. Thai, S. Eidenbenz, Containment of Misinformation Spread in Online Social Networks, in: *Proceedings of the 4th Annual ACM Web Science Conference*, ACM, Evanston Illinois, 2012, pp. 213–222. doi:10.1145/2380718.2380746.
- [48] A. Kalogeratos, K. Scaman, L. Corinzia, N. Vayatis, Chapter 24 - Information Diffusion and Rumor Spreading, in: *Cooperative and Graph Signal Processing*, Academic Press, 2018, pp. 651–678. doi:10.1016/B978-0-12-813677-5.00024-9.
- [49] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-Tracking and the Dynamics of the News Cycle, in: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 497–506. doi:10.1145/1557019.1557077.
- [50] W. Lee, J. Kim, H. Yu, CT-IC: Continuously Activated and Time-Restricted Independent Cascade Model for Viral Marketing, in: *2012 IEEE 12th International Conference on Data Mining*, IEEE, 2012, pp. 960–965. doi:10.1109/icdm.2012.40.
- [51] D. J. Watts, A Simple Model of Global Cascades on Random Networks, *Proceedings of the National Academy of Sciences* 99 (2002) 5766–5771. doi:10.1073/pnas.082090499.
- [52] Y. Zhuang, A. Arenas, O. Yağan, Clustering Determines the Dynamics of Complex Contagions in Multiplex Networks, *Physical Review E* 95 (2017) 012312. doi:10.1103/PhysRevE.95.012312.
- [53] C. Lagnier, L. Denoyer, E. Gaussier, P. Gallinari, Predicting Information Diffusion in Social Networks Using Content and User's Profiles, in: *Advances in Information Retrieval*, Springer, Berlin, Heidelberg, 2013, pp. 74–85. doi:10.1007/978-3-642-36973-5\_7.
- [54] A. G. Hawkes, Spectra of Some Self-Exciting and Mutually Exciting Point Processes, *Biometrika* 58 (1971) 83–90. doi:10.2307/2334319.
- [55] Y. Jiang, M. D. Porter, Simulating Fake News Dissemination on Twitter with Multivariate Hawkes Processes, in: *2022 IEEE International Conference on Big Data (Big Data)*, IEEE, Osaka, Japan, 2022, pp. 3597–3606. doi:10.1109/BigData55660.2022.10020285.
- [56] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, H. Zha, Fake News Mitigation via Point Process Based Intervention, in: *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 2017, pp. 1097–1106.
- [57] M. Caglar, O. Ozkasap, A Chain-Binomial Model for Pull and Push-Based Information Diffusion, in: *2006 IEEE International Conference on Communications*, IEEE, Istanbul, 2006, pp. 909–914. doi:10.1109/ICC.2006.254823.
- [58] D. A. Vega-Oliveros, L. d. F. Costa, F. A. Rodrigues, Rumor Propagation with Heterogeneous Transmission in Social Networks, *Journal of Statistical Mechanics: Theory and Experiment* 2017 (2017) 023401. doi:10.1088/1742-5468/aa58ef.
- [59] T. Zhu, B. Wang, B. Wu, C. Zhu, Maximizing the Spread of Influence Ranking in Social Networks, *Information Sciences* 278 (2014) 535–544. doi:10.1016/j.ins.2014.03.070.
- [60] S. Han, F. Zhuang, Q. He, Z. Shi, X. Ao, Energy Model for Rumor Propagation on Social Networks, *Physica A: Statistical Mechanics and its Applications* 394 (2014) 99–109. doi:10.1016/j.physa.2013.10.003.
- [61] V. Indu, S. M. Thampi, A Nature - Inspired Approach Based on Forest Fire Model for Modeling Rumor Propagation in Social Networks, *Journal of Network and Computer Applications* 125 (2019) 28–41. doi:10.1016/j.jnca.2018.10.003.
- [62] E. Serrano, C. A. Iglesias, Validating Viral Marketing Strategies in Twitter via Agent-Based Social Simulation, *Expert Systems with Applications* 50 (2016) 140–150. doi:10.1016/j.eswa.2015.



- 12.021.
- [63] A. Averza, K. Slhoub, S. Bhattacharyya, Evaluating the Influence of Twitter Bots via Agent-Based Social Simulation, *IEEE Access* 10 (2022) 129394–129407. doi:10.1109/ACCESS.2022.3228258.
  - [64] A. Gausen, W. Luk, C. Guo, Can We Stop Fake News? Using Agent-Based Modelling to Evaluate Countermeasures for Misinformation on Social Media, in: *ICWSM Workshops*, 2021, pp. 1–5.
  - [65] Q. F. Lotito, D. Zanella, P. Casari, Realistic Aspects of Simulation Models for Fake News Epidemics over Social Networks, *Future Internet* 13 (2021) 76. doi:10.3390/fi13030076.
  - [66] J.-H. Cho, S. Rager, J. O'Donovan, S. Adali, B. D. Horne, Uncertainty-Based False Information Propagation in Social Networks, *ACM Transactions on Social Computing* 2 (2019) 1–34. doi:10.1145/3311091.
  - [67] N. Rabb, L. Cowen, J. P. De Ruiter, M. Scheutz, Cognitive Cascades: How to Model (and Potentially Counter) the Spread of Fake News, *PLOS ONE* 17 (2022) e0261811. doi:10.1371/journal.pone.0261811.
  - [68] M. Tambuscio, D. F. M. Oliveira, G. L. Ciampaglia, G. Ruffo, Network Segregation in a Model of Misinformation and Fact-Checking, *Journal of Computational Social Science* 1 (2018) 261–275. doi:10.1007/s42001-018-0018-9.
  - [69] W. Li, Q. Bai, M. Zhang, A Multi-agent System for Modelling Preference-Based Complex Influence Diffusion in Social Networks, *The Computer Journal* 62 (2019) 430–447. doi:10.1093/comjnl/bxy078.
  - [70] P. Costa, R. McCrae, Personality in Adulthood: A Five-Factor Theory Perspective, *Management Information Systems Quarterly - MISQ* (2002). doi:10.4324/9780203428412.
  - [71] R. F. Muhammad, S. Kasahara, Agent-Based Simulation of Fake News Dissemination: The Role of Trust Assessment and Big Five Personality Traits on News Spreading, *Social Network Analysis and Mining* 14 (2024) 75. doi:10.1007/s13278-024-01235-8.
  - [72] S.-H. Tseng, T. Son Nguyen, Agent-Based Modeling of Rumor Propagation Using Expected Integrated Mean Squared Error Optimal Design, *Applied System Innovation* 3 (2020) 48. doi:10.3390/asi3040048.
  - [73] Y. Liu, X. Chen, X. Zhang, X. Gao, J. Zhang, R. Yan, From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News, *arXiv preprint* (2024). ArXiv:2403.09498.
  - [74] J. Pastor-Galindo, P. Nespoli, J. A. Ruipérez-Valiente, Large-Language-Model-Powered Agent-Based Framework for Misinformation and Disinformation Research: Opportunities and Open Challenges, *IEEE Security & Privacy* 22 (2024) 24–36. doi:10.1109/MSEC.2024.3380511.
  - [75] J. Brainard, P. R. Hunter, Misinformation Making a Disease Outbreak Worse: Outcomes Compared for Influenza, Monkeypox, and Norovirus, *SIMULATION* 96 (2019) 365–374. doi:10.1177/0037549719885021.
  - [76] C. Marshall, J. Cruickshank, C. O'Riordan, Identifying Influential Nodes to Inhibit Bootstrap Percolation on Hyperbolic Networks, in: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, Barcelona, 2018, pp. 1266–1273. doi:10.1109/ASONAM.2018.8508248.
  - [77] D. Azucar, D. Marengo, M. Settanni, Predicting the Big 5 Personality Traits from Digital Footprints on Social Media: A Meta-Analysis, *Personality and Individual Differences* 124 (2018) 150–159. doi:10.1016/j.paid.2017.12.018.
  - [78] Y. Wang, D. Jin, C. Yang, J. Dang, Integrating Group Homophily and Individual Personality of Topics Can Better Model Network Communities, in: *2020 IEEE International Conference on Data Mining (ICDM)*, IEEE, Sorrento, Italy, 2020, pp. 611–620. doi:10.1109/ICDM50108.2020.00070.
  - [79] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media, *Big Data* 8 (2020) 171–188. doi:10.1089/big.2020.0062.
  - [80] V. Qazvinian, E. Rosengren, D. R. Radev, Q. Mei, Rumor Has It: Identifying Misinformation in Microblogs, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 1589–1599. URL: <https://aclanthology.org/D11-1147>.

- [81] X. Zhou, A. Mulay, E. Ferrara, R. Zafarani, ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 3205–3212. doi:10.1145/3340531.3412880.
- [82] L. Cui, D. Lee, CoAID: COVID-19 Healthcare Misinformation Dataset, *arXiv preprint* (2020). doi:10.48550/arXiv.2006.00885, arXiv:2006.00885.
- [83] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, P. Tolmie, Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads, *PLOS ONE* 11 (2016) e0150989. doi:10.1371/journal.pone.0150989.
- [84] J. Leskovec, A. Krevl, SNAP Datasets: Stanford Large Network Dataset Collection, <http://snap.stanford.edu/data>, 2014.