A Multi-Faceted NLP Analysis of Misinformation Spreaders in Twitter

Dimosthenis Antypas, Alun Preece, Jose Camacho-Collados
Cardiff NLP, School of Computer Science and Informatics
& Security, Crime and Intelligence Innovation Institute
Cardiff University, United Kingdom
{AntypasD,PreeceAD,CamachoColladosJ}@cardiff.ac.uk

Abstract

Social media is an integral part of the daily life of an increasingly large number of people worldwide. Used for entertainment, communication and news updates, it constitutes a source of information that has been extensively used to study human behaviour. Unfortunately, the open nature of social media platforms along with the difficult task of supervising their content has led to a proliferation of misinformation posts. In this paper, we aim to identify the textual differences between the profiles of user that share misinformation from questionable sources and those that do not. Our goal is to better understand user behaviour in order to be better equipped to combat this issue. To this end, we identify Twitter (X) accounts of potential misinformation spreaders and apply transformer models specialised in social media to extract characteristics such as sentiment, emotion, topic and presence of hate speech. Our results indicate that, while there may be some differences between the behaviour of users that share misinformation and those that do not, there are no large differences when it comes to the type of content shared.

1 Introduction

The emerging popularity of social platforms such as Facebook, Twitter, and WhatsApp has revolutionised the way information is disseminated and consumed (Fac, 2023; Murthy, 2018; Deshmukh, 2015). People are able to express their sentiments, share their opinions on multiple topics, and to discuss and influence each other with ease and at a speed that has transformed not only how we communicate but also how we perceive the world around us. Unfortunately, the capacity to reach a vast audience within seconds, along with the challenges that arise with verifying an ever-expanding volume of content, has created fertile ground within social media for malicious actors, or unaware users, to spread misinformation. The recent examples of fake news related to the COVID-19 pandemic (Evanega et al., 2020), and the ongoing war in Ukraine (Pierri et al., 2023) demonstrate that misinformation in social media is a complex problem with far-reaching implications for society, democracy, and information integrity.

Combating misinformation in social media is a topic that is studied extensively in academia (Vosoughi et al., 2018; Pennycook et al., 2020) and in the natural language processing (NLP) community (Su et al., 2020) specifically, among others. Common approaches of dealing with misinformation include defining the problem as a classification task (Serrano et al., 2020; Hamid et al., 2020) and classifying a post as fake or not; with fact-checking (Thorne and Vlachos, 2018) often defined as an information retrieval task (Lazarski et al., 2021). However, research regarding the agents that share misinformation is rather limited in comparison (Shu et al., 2020; Rangel et al., 2020; Dou et al., 2021) particularly when it comes to analysing language-specific features.

In this paper, we focus on misinformation in Twitter and perform an analytical comparison between different types of user based on their content shared online and the reliability of their sources. To this end, we first compiled three diverse datasets in which spreaders of misinformation are categorised using different techniques. Then, we perform an exhaustive analysis of the content of these users by leveraging transformer-based language models specialised on social media tasks such as sentiment analysis, emotion recognition, topic categorisation and hate speech detection. The main contributions of this paper are the following: (1) we gather and consolidate existing and new Twitter datasets related to misinformation spreaders; and (2) we extract insights for the behaviour of such users in comparison with users sharing content from reliable sources.

2 Related Work

The study of identifying misinformation has been a prominent area of research in recent years. Initially, efforts focused on addressing the problem through classification, either in a binary or multiclass context. Some studies delved into examining the spread of true and false information online work on the topic of information dissemination (Vosoughi et al., 2018). Meanwhile, others opted for a data mining approach in the realm of fake news detection on social media, utilising various features and machine learning algorithms to classify news articles as true or false (Shu et al., 2017).

Moreover, beyond binary classification, researchers explored multiclass classification methods. For instance, Castillo et al. (2011) investigated the credibility of information on Twitter and proposed a framework categorizing tweets into four groups: true, false, unverified, and non-informative. Zubiaga et al. (2016) delved into the analysis of conversational threads on social media to gain insights into how rumors propagate and how individuals respond to them, shedding light on the dynamics of misinformation propagation.

These approaches evolved to better serve journalists and fact-checkers. The focus shifted from classification to fact-checking and information retrieval, aiming to assist journalists in source verification. This transition led to the development of tools to meet their specific needs (Schlichtkrull et al., 2023). The availability of datasets like FEVER (Thorne et al., 2018), MultiFC (Augenstein et al., 2019), and X-Fact (Gupta and Srikumar, 2021) has been instrumental in enabling researchers to experiment with and develop novel methods for evidence retrieval and rumour verification (Nasir et al., 2021; Lee et al., 2020; Lewis et al., 2020).

While there have been notable studies in the broader field of misinformation and fact verification, there's a notable gap when it comes to a systematic analysis of the textual content of fake news spreaders. Much of the existing research has predominantly focused on the detection of misinformation sources, fact-checking, or the development of classification algorithms to distinguish true from false content. However, there is limited in-depth work that methodically dissects the text generated by those actively involved in spreading fake news that utilises state-of-the-art models (Ghanem et al., 2020; Rangel et al., 2020) and where the language analysis is not supplementary to the network and graph analysis (Aswani et al., 2019).

In this work we seek to methodically analyse the textual content generated by those responsible for spreading fake news. The primary objective is to gain a deeper understanding of the characteristics, strategies, and linguistic patterns employed by these actors in disseminating misleading or false information. Unlike traditional fact-checking, our work does not intend to verify or debunk specific claims but rather aims at understanding the textual content shared by individuals or groups behind the spread of fake news, thereby providing further insights into their content dissemination strategies.

3 Data

For our analysis, we exclusively focus on Twitter users, particularly tweets in the English language. Our goal was to extract a diverse tweet corpus for both users regularly spreading fake news or news from questionable sources, and users sharing content from verified sources. In the following we describe our data collection methodology stemming from various sources.

3.1 Data Collection

In total, we draw upon three diverse data sources to extract relevant tweets from user account sharing trusted and untrusted sources. Moreover, we extract tweets from legacy-verified Twitter accounts as a control group.

3.1.1 Media Bias Fact Check (MBFC)

Our first corpus is extracted from a list of known conspiracy sites provided by "Media Bias Fact Check" (*MBFC*). This source is commonly used in the study of fake news (Nakov and Da San Martino, 2020; Cinelli et al., 2020). For this dataset, we extracted tweets that share URLs from known untrusted sites[1] and then sample users based on the frequency of sharing these links. In particular, we considered only those users in the 75 percentile in terms of number of links shared. In order to gather enough information, all user accounts that were not older than 30 days were excluded from the analysis. Subsequently, all posts made by the sampled users during September 2021 were collected, which aligns with the date when the *MBFC* lists were last updated prior to conducting this experiment. User accounts were then further filtered based on their activity, only keeping those users posting more

[1] https://mediabiasfactcheck.com/conspiracy/

frequently than the median daily posts. Finally, to ensure a diverse representation, users were sampled based on their number of followers by maintaining the original distribution and thus encompassing both popular and less popular accounts. This final sample represents the ***MBFC-untrusted*** subset.

The above methodology is mirrored to collect users that share links form trusted news-sources according to MBFC[2] resulting in the ***MBFC-trusted*** subset.

3.1.2 FakeNewsNet (FNN)

The FakeNewsNet dataset, referred to as *FNN* (Shu et al., 2018), contains two subsets: (1) tweets related to news content, e.g. tweets revolving around US politics and tweets; and (2) tweets related to social context, e.g. tweets talking about celebrities. Tweets in each groups are further classified as either untrusted or trusted. For the purpose of this study, we concentrate solely on the politics-related subset, as it exhibits a closer alignment with the majority of the links found within the *MBFC* lists. To extract relevant users, we initially scrape all tweets in the dataset and randomly sample users. Finally, all tweets posted by the selected users from September 2021 are retrieved. Only the accounts that have at least 100 posts were considered to create the ***FNN-untrusted*** and ***FNN-trusted*** subsets.

3.1.3 Profiling Fake News Spreaders (PAN)

The English subset of the *PAN 2020: Profiling Fake News Spreaders* task (*PAN*) (Rangel et al., 2020) is dataset that comprises a total of 50,000 English tweets obtained from 500 users, with each user contributing 100 tweets. These users are categorized as either trusted news spreaders (***PAN-trusted***) or untrusted news spreaders (***PAN-untrusted***). In the interest of privacy, no additional user-specific information, such as author descriptions or popularity metrics, is disclosed. Despite its relatively modest size and the limitation on the extraction of additional user details, the *PAN* dataset is considered robust and reliable. Its construction involved manual checks, and it underwent thorough scrutiny by multiple individuals, primarily due to its relevance in a competitive context. This rigorous validation process enhances the dataset's trustworthiness and accuracy.

		Tweets	Users	Size	TTR	#emoji
MBFC	untrusted	1,703,896	1,489	136	0.018	0.24
	trusted	1,676,615	1,535	132	0.021	0.26
FNN	untrusted	246,107	430	122	0.036	0.19
	trusted	351,857	476	124	0.030	0.13
PAN	untrusted	25,000	250	88	0.138	0.02
	trusted	25,000	250	88	0.149	0.13
Verified users		178,324	803	103	0.048	0.26
Total		4,206,799	5,233	123	0.014	0.24

Table 1: Number of tweets and users present in each dataset studied. The average size of the tweet (number of characters), along with the Type Token Ratio (TTR) and average emoji presence, are also reported.

3.1.4 Control (Verified users)

In order to have a control group to compare in our experiments, we sampled tweets from legacy-verified accounts for which the authenticity is known. This dataset was compiled by sampling verified users and collecting their tweets during the same time period as the previous datasets. Our aim was to select users whose characteristics align closely with the distribution patterns observed in the *FNN* and *MBFC* datasets.

3.2 Statistics and Descriptive Analysis

By considering these diverse data sources, we aim to comprehensively examine and understand the dynamics of untrusted news spreaders on the Twitter platform. Our analysis encompasses a total of 4,206,799 tweets contributed by 5,233 users, as presented in Table 1. In addition to the number of tweets and users, we also investigate the average length of tweet and average emoji usage. We did not identify a clear pattern between the *trusted* and *untrusted* subsets as far as these metrics are concerned.

Looking into the lexical characteristics of each dataset, distinctions between the *untrusted* and *trusted* subsets become more apparent. For instance, when assessing lexical diversity using the Type Token Ratio (TTR), we observe that *untrusted* users, with the exception of the *FNN* dataset, tend to employ a less diverse vocabulary which is consistent with previous research (Horne and Adali, 2017). Our analysis based on the average presence of emojis in each tweet reveals no consistent pattern, despite prior research suggesting higher emoji usage among untrusted news spreaders (Er and Yilmaz, 2023). For example, while the untrusted subset of the MBFC dataset exhibits higher emoji usage, the opposite holds true for the FNN dataset.

[2] https://mediabiasfactcheck.com/pro-science/, https://mediabiasfactcheck.com/center/

MBFC		FNN		PAN		Verified
untrusted	trusted	untrusted	trusted	untrusted	trusted	
news	tigray	biden	music	trump	film	game
biden	jisoo	people	hit	realdonaldtrump	kobe	thunderstorm
vaccine	ethiopia	say	play	new	season	football
covid	indiedev	trump	househunters	instyle	styles	season
border	tigraygenocide	ebay	dance	webtalk	spoilers	good
passport	brexit	marijuana	trump	post	promo	thank
australia	bts	prohibition	biden	impeachment	date	collision
mandate	dior	covid	september	publish	trailer	direction

Table 2: Top eight terms in each dataset according to lexical specificity.

Lexical specificity. To gain an overall understanding of the prevalent topics within our corpora, we employ lexical specificity (Lafon, 1980). Lexical specificity is a word-level metric that indicates the importance of each word in a subcorpus. In particular, for this analysis we use the formulation outlined in Camacho-Collados et al. (2016), and extract the top terms in each dataset. Table 2 displays the top ten lemmas[3] in each dataset based on their lexical specificity scores.

Notably, due to the same time period during data collection, a significant overlap exists between the *MBFC* and *FNN* datasets, particularly within their ’untrusted’ subsets. Terms such as ’biden’ and ’vaccine’ are common across both. Additionally, a discernible trend emerges, indicating that ’untrusted’ subsets across datasets often feature more controversial and divisive topics. This is evident in the presence of terms like ’covid,’ ’prohibition,’ and ’impeachment,’ in contrast to the ’trusted’ subsets, which exhibit more generic and neutral terms such as ’bts,’ ’music,’ and ’film.’ This distinction becomes even more pronounced when examining the top terms in the *Verified* dataset, which include terms like ’game,’ ’football,’ and ’love.’

4 Methodology

Our goal is to analyse various content-related features from the extracted posts in Section 3. To capture the nuanced language features present in the data, we employ a range of pre-trained language models designed for social media usage. Our primary focus encompasses sentiment and affection analysis, topic classification, and the identification of hate speech in textual content, features are frequently employed in the study of misinformation propagation (Vicario et al., 2019; Verma et al., 2020), aiming to uncover emotionally charged language and controversial topics.

All the language models used are built upon the RoBERTa architecture (Liu et al., 2019) and trained on social media corpora, making them well-suited for analysing Twitter data. More specifically:

Sentiment Analysis. The model *twitter-roberta-base-sentiment-latest* (Loureiro et al., 2022) is used to extract the sentiment polarity where each tweet is classified as *negative, neutral, or positive*. This model has been fine-tuned for sentiment analysis using the dataset provided in the *Sentiment Analysis in Twitter* task of Semeval 2017 (Rosenthal et al., 2019). By analysing the sentiment expressed in social media content, we can gain insights into information being shared (Baishya et al., 2021). Specifically, presence of exaggerated positive sentiment or negative sentiment in response to fake news can serve as indicators of misinformation (Alonso et al., 2021).

Emotion Analysis. We leverage *twitter-roberta-base-emotion-multilabel-latest* (Camacho-Collados et al., 2022) to assign one or more emotions to each tweet. This model is trained using data from the ’Affect in Tweets’ Semeval 2018 task (Mohammad et al., 2018), covering 11 different emotions. Similar to sentiment analysis, the presence of specific emotions has been used to analyse the spread of rumours and misinformation, with negative emotions potentially contributing to the spread of misinformation (Vosoughi et al., 2018; Weeks, 2015).

Hate Speech Detection. We use the *twitter-roberta-base-hate-multiclass* hate speech detection model (Antypas and Camacho-Collados, 2023), which is trained on a combination of 13 different hate speech Twitter datasets and is capable of identifying hate speech from seven target groups. The inclusion of hate speech detection as a feature is motivated by previous research indicating a positive correlation between the presence of hate speech and misinformation (Inwood and Zappavigna, 2023).

[3] Lemmatization was done using *spaCy* https://spacy.io/.

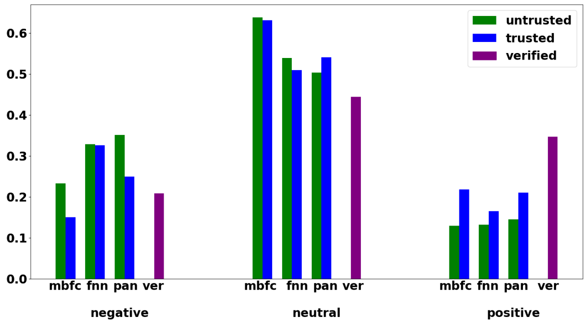


Figure 1: Sentiment distribution in each dataset for *trusted* and *untrusted* users in Twitter.

Topic Classification. We use *tweet-topic-21-multi* (Antypas et al., 2022), a multi-label classification model fine-tuned on a Twitter topic classification dataset. This model assigns one or more topics to each tweet from a list of 19 topics. Our hypothesis is that there may be a significant difference between the topics discussed by *untrusted* news spreaders and regular users, e.g. untrusted news spreaders potentially engaging in discussions related to sensitive topics at a higher volume.

All the specialised models described above are perform in line of the state of the art for each of the tasks in the social media context[4] and they enable us to delve deeper into the complex linguistic nuances within the social media data. Nonetheless, as we describe in the Limitations section, they all have a degree of error that needs to be considered when making conclusions.

5 Analysis

We consider each pair of collected datasets (*untrusted* and *trusted*), along with the *Verified* control dataset. Our examination involves a comparison of the tweets within each dataset individually, as well as their aggregation for each user. This holistic approach enables us to explore a variety of perspectives and insights across the datasets and their combined impact.

5.1 Textual Analysis

Table 3 displays the aggregated results for the sentiment, emotion, hate speech and topic analysis. For each feature we consider each user independently by taking their mean value and then aggregate the results of users belonging in the same subset. Even though differences between *untrusted* and *trusted* subsets exist, it is challenging to identify trends that are consistent across the datasets. In the following sections we investigate each characteristic individually.

5.1.1 Sentiment

When evaluating the presence of sentiment in tweets, a noticeable trend emerges: tweets associated with *untrusted* news spreaders tend to exhibit a higher degree of negativity compared to those posted by other users. The distribution of sentiment across the datasets is displayed in Figure 1. In the case of the *FNN* dataset, however, this difference is almost not negligible. Finally, even though there is more negativity in *untrusted* users, the distributions among negative, neutral and positive tweets are very similar in all cases except for the verified users that tend to be more positive overall.

5.1.2 Emotion

Similarly to the findings in sentiment analysis, the analysis of affect reveals a consistent pattern where untrusted news spreaders tend to gravitate toward more negative emotions. Figure 2a provides insight into the distribution of the 11 emotions present across across all subsets.

A clear contrast emerges, with tweets attributed to *trusted* users generally displaying greater joy and featuring a lesser presence of anger and disgust, in stark contrast to the tweets originating from *untrusted* users. This trend remains consistent even when evaluating the per-user aggregation. Finally, similarly to the sentiment distribution patterns, there are no noticeable differences when analysing the overall emotion distribution and, in this case, it also related to that of verified users.

5.1.3 Hate Speech

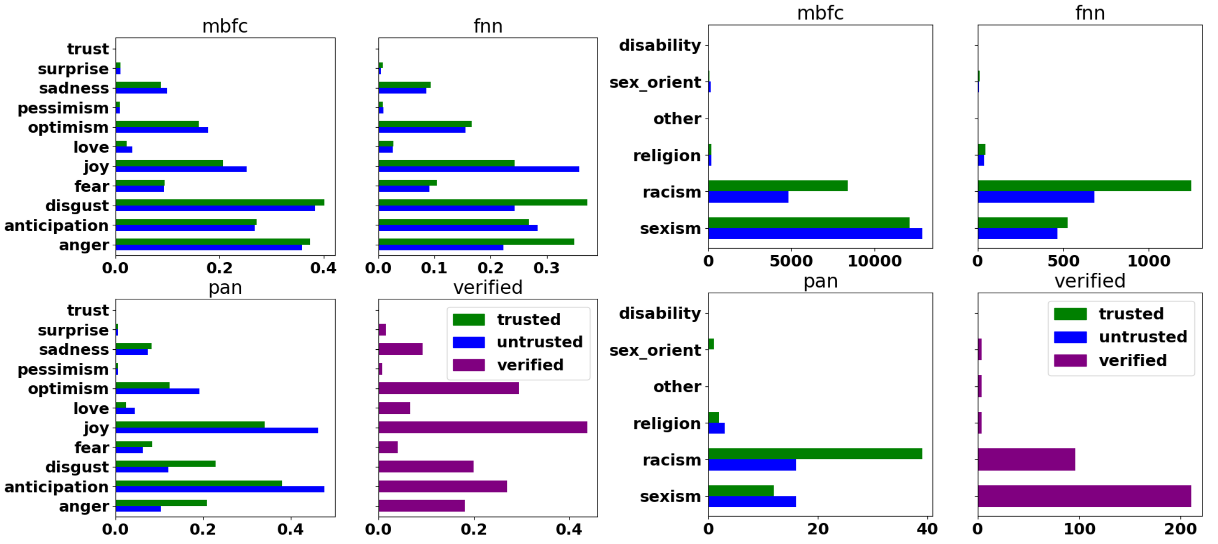
When examining hate speech, a feature that often coexists with misinformation (Inwood and Zappavigna, 2023), such as Holocaust denial and the Great Replacement theory, it does not appear to be a prominent feature in the collected datasets. Our analysis indicates an absence of hate speech, with 99% of all tweets being devoid of it.

There does appear to be a variance in the types of hate speech across the subsets (as displayed in Figure 2b). *Untrusted* subsets exhibit a higher inclination towards racism, while in the trusted subset, sexism appears to be more prevalent. However, given the limited number of instances, it is prudent to exercise caution when drawing extensive conclusions based on this data.

[4] Sentiment Analysis: 73.7% Recall, Emotion Analysis: 80% F1-macro, Hate Speech: 94% Accuracy, Topic Classification: 59% F1-macro – please refer to the individual references for more details.

		MBFC		PAN		FNN		verified
		untrusted	trusted	untrusted	trusted	untrusted	trusted	
Senti.	negative	0.35 ± 0.12	0.32 ± 0.15	0.16 ± 0.14	0.24 ± 0.16	0.35 ± 0.19	0.37 ± 0.19	0.2 ± 0.11
	neutral	0.52 ± 0.1	0.5 ± 0.11	0.63 ± 0.18	0.64 ± 0.15	0.48 ± 0.18	0.47 ± 0.17	0.44 ± 0.15
	positive	0.13 ± 0.09	0.17 ± 0.12	0.22 ± 0.17	0.13 ± 0.09	0.17 ± 0.15	0.16 ± 0.14	0.37 ± 0.16
Emotion	anger	0.39 ± 0.14	0.34 ± 0.17	0.1 ± 0.11	0.21 ± 0.17	0.33 ± 0.24	0.37 ± 0.23	0.17 ± 0.11
	anticipation	0.25 ± 0.1	0.26 ± 0.12	0.48 ± 0.2	0.38 ± 0.18	0.28 ± 0.19	0.26 ± 0.19	0.28 ± 0.12
	disgust	0.42 ± 0.15	0.37 ± 0.17	0.12 ± 0.12	0.23 ± 0.18	0.35 ± 0.23	0.39 ± 0.23	0.18 ± 0.11
	fear	0.1 ± 0.05	0.09 ± 0.08	0.06 ± 0.08	0.08 ± 0.07	0.09 ± 0.09	0.1 ± 0.12	0.04 ± 0.05
	joy	0.2 ± 0.12	0.26 ± 0.17	0.46 ± 0.22	0.34 ± 0.22	0.26 ± 0.2	0.24 ± 0.19	0.46 ± 0.16
	love	0.02 ± 0.03	0.03 ± 0.05	0.04 ± 0.06	0.02 ± 0.03	0.03 ± 0.05	0.03 ± 0.05	0.07 ± 0.07
	optimism	0.16 ± 0.1	0.19 ± 0.11	0.19 ± 0.14	0.12 ± 0.1	0.18 ± 0.14	0.18 ± 0.14	0.31 ± 0.14
	pessimism	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.02	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
	sadness	0.09 ± 0.04	0.1 ± 0.05	0.07 ± 0.06	0.08 ± 0.05	0.1 ± 0.07	0.1 ± 0.09	0.09 ± 0.05
	surprise	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01
	trust	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Hate	disability	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	not_hate	0.99 ± 0.01	0.99 ± 0.01	1.0 ± 0.01	1.0 ± 0.01	0.99 ± 0.01	0.99 ± 0.02	1.0 ± 0.0
	other	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
	racism	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.02	0.01 ± 0.01	0.01 ± 0.02	0.01 ± 0.01
	sexism	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.0
	religion	0.0 ± 0.0	0.0 ± 0.0	0.03 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
	sex_orient	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Topic	arts	0.01 ± 0.03	0.02 ± 0.03	0.04 ± 0.1	0.02 ± 0.02	0.02 ± 0.05	0.02 ± 0.08	0.01 ± 0.03
	business	0.06 ± 0.11	0.05 ± 0.1	0.07 ± 0.11	0.06 ± 0.1	0.05 ± 0.11	0.04 ± 0.1	0.02 ± 0.05
	celebrity	0.05 ± 0.04	0.07 ± 0.08	0.24 ± 0.21	0.28 ± 0.28	0.05 ± 0.05	0.05 ± 0.08	0.07 ± 0.07
	diaries	0.07 ± 0.07	0.09 ± 0.09	0.08 ± 0.1	0.04 ± 0.07	0.08 ± 0.1	0.08 ± 0.11	0.15 ± 0.1
	family	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.02	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
	fashion	0.0 ± 0.01	0.01 ± 0.02	0.06 ± 0.13	0.05 ± 0.1	0.01 ± 0.01	0.01 ± 0.02	0.01 ± 0.01
	film	0.03 ± 0.03	0.04 ± 0.05	0.23 ± 0.26	0.16 ± 0.18	0.04 ± 0.06	0.05 ± 0.09	0.06 ± 0.07
	fitness	0.11 ± 0.11	0.06 ± 0.07	0.02 ± 0.04	0.02 ± 0.03	0.05 ± 0.07	0.06 ± 0.07	0.02 ± 0.04
	food	0.01 ± 0.02	0.02 ± 0.03	0.02 ± 0.05	0.01 ± 0.02	0.02 ± 0.04	0.02 ± 0.03	0.03 ± 0.03
	gaming	0.0 ± 0.02	0.0 ± 0.02	0.01 ± 0.02	0.01 ± 0.03	0.0 ± 0.02	0.0 ± 0.02	0.01 ± 0.04
	learning	0.02 ± 0.03	0.03 ± 0.04	0.01 ± 0.02	0.02 ± 0.02	0.03 ± 0.06	0.03 ± 0.07	0.02 ± 0.04
	music	0.02 ± 0.03	0.03 ± 0.07	0.09 ± 0.14	0.08 ± 0.13	0.04 ± 0.13	0.04 ± 0.11	0.04 ± 0.06
	news	0.76 ± 0.2	0.67 ± 0.27	0.31 ± 0.27	0.51 ± 0.29	0.65 ± 0.3	0.68 ± 0.27	0.24 ± 0.22
	hobbies	0.01 ± 0.03	0.01 ± 0.02	0.01 ± 0.02	0.0 ± 0.01	0.01 ± 0.04	0.01 ± 0.03	0.01 ± 0.01
	relations	0.01 ± 0.01	0.01 ± 0.02	0.02 ± 0.03	0.02 ± 0.03	0.01 ± 0.01	0.01 ± 0.02	0.01 ± 0.02
	science	0.05 ± 0.07	0.04 ± 0.08	0.04 ± 0.11	0.05 ± 0.09	0.04 ± 0.06	0.04 ± 0.08	0.02 ± 0.04
	sports	0.03 ± 0.05	0.06 ± 0.13	0.08 ± 0.15	0.08 ± 0.11	0.08 ± 0.15	0.05 ± 0.1	0.35 ± 0.31
	travel	0.01 ± 0.01	0.01 ± 0.02	0.02 ± 0.07	0.01 ± 0.01	0.02 ± 0.04	0.01 ± 0.02	0.02 ± 0.05
	youth	0.02 ± 0.02	0.02 ± 0.03	0.01 ± 0.01	0.01 ± 0.02	0.02 ± 0.06	0.01 ± 0.02	0.01 ± 0.02

Table 3: Average presence of each feature (i.e., sentiment analysis, emotion analysis, hate speech, and topic classification) per user along with standard deviations.



(a) Emotion distribution in each dataset.

(b) Hateful entries in each dataset.

Figure 2: Emotion & Hate speech results of *trusted* and *untrusted* users in Twitter.

5.1.4 Topics

Regarding the topics that *untrusted* news spreaders and regular users typically discuss, the results appear to suggest a similar distribution of topics (Figure 3). *Untrusted* news spreaders appear to engage more extensively in posting tweets related to news and social issues, which are those related to politics, among others. This suggests that these accounts may be more socially active, and can create the illusion of a larger representation than that of the general population.

Conversely, there is no discernible distinction in the case of the remaining popular topics, with variations existing among the datasets. For instance, the topic "celebrity_&_pop_culture" is more prevalent in the *Panuntrusted* dataset but less common in the other untrusted subsets. Again here we can observe more differences with respect to verified users, where *sports* and *diaries_&_daily_life* topics are much more prominent.

5.2 Spreader Detection Analysis

Recognising that a significant portion of our datasets relies on weak labels, with the distinction between users propagating untrusted news and those who do not being based on heuristics based on the number of posts shared from untrusted sources, we perform a robustness analysis on the *Pan20* dataset which includes train and test splits. To this end, we train a classifier capable of discerning between the *trusted* and *untrusted* classes and compare the results with our approach.

The train/test split originally utilised in the competition is retained, consisting of 300 users for training and 200 users for testing. We assess the performance of two classifiers: (1) A classifier based on the best-performing models as presented in the competition (Buda and Bolonyai, 2020; Pizarro, 2020), utilising an XGBoost classifier (Chen and Guestrin, 2016). This model is trained using TFIDF features and a combination of word and character n-grams; and (2) A pre-trained Longformer (Beltagy et al., 2020), which is further fine-tuned using the *PAN* dataset. We leverage the implementation provided by Hugging Face (Wolf et al., 2020) for the fine-tuning of the Longformer[5]. Hyper-parameter tuning, including batch size, epochs number, and learning rate, is conducted using Ray Tune (Liaw et al., 2018).

The results reveal that the XGBoost model (*XGB*) surpasses the Longformer classifier, achieving a 74% macro F1 score compared to the Longformer's 70%. One possible explanation for this outcome lies in the unstructured nature of Twitter text, which presents an added challenge to the language model. The Longformer, not explicitly trained on social media corpus data, may face limitations in handling this specific type of text.

When examining the results of the XGB classifier in the *PAN* dataset, we observe an almost identical trend when compared with our initial results. For example, when looking the sentiment distribution of user accounts using the XGB classifier

[5] https://huggingface.co/allenai/longformer-base-4096

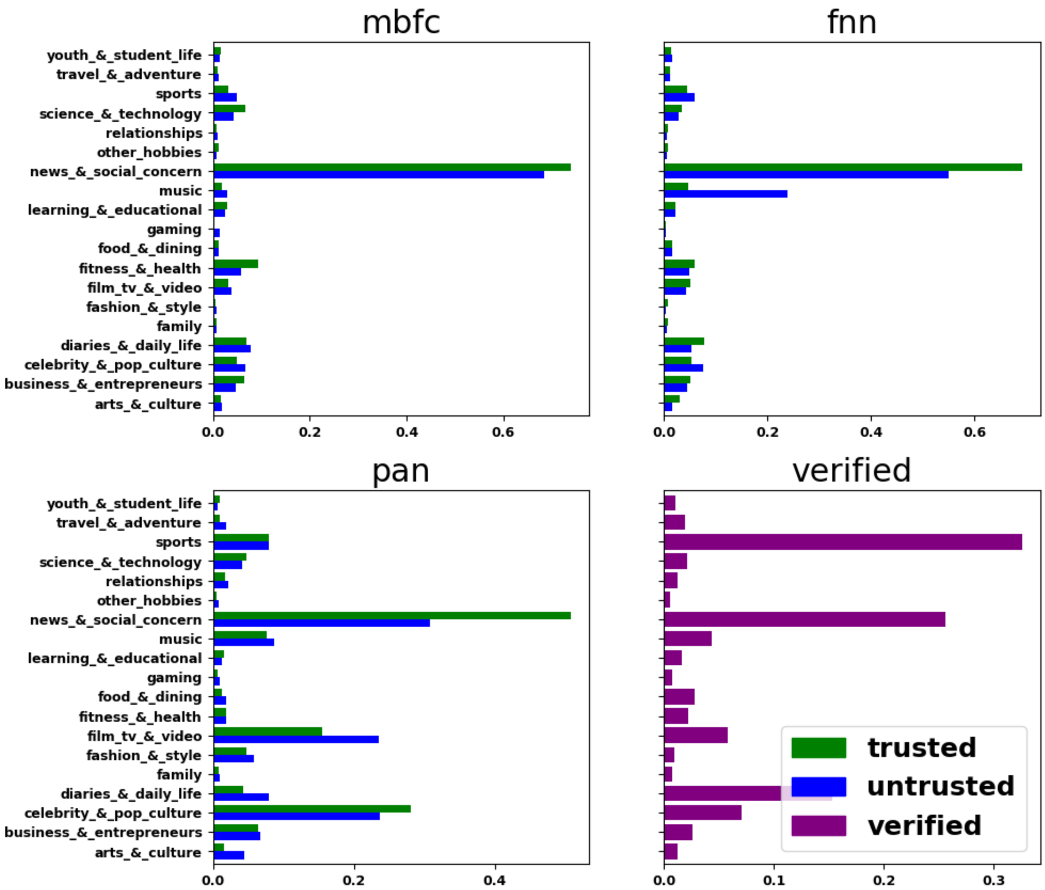


Figure 3: Topic distribution in each dataset for *trusted* and *untrusted* users in Twitter.

and our initial distinction of *trusted* and *untrusted* users, only minimal differences can be observed (*Pan-untrusted*: 23% Negative and 13% Positive; *XGB-untrusted*: 24% and 13%). Our experiment indicates that even though developing a classifier to identify *untrusted* users may not be the optimal approach, it can still be used as a proxy to derive useful information and identify patterns that can be used to reveal malicious actors. Additional results for all tasks regarding the performance of the *XGB* model and the differences with our approach follow a similar trend and can be found in Appendix C.

As a final experiment, we attempt to enhance our XGB classifier by integrating the features already extracted. Our results reveal that while the incorporation of new features generally results in only marginal variations in the model's performance, the addition of certain features, especially sentiment, holds the potential to notably improve its effectiveness. This suggests that careful selection and integration of specific features can yield incremental but meaningful gains in the classifier's performance, an exploration that we leave for future work as it falls out of the scope of this work.

6 Conclusions

This paper's comparative analysis aims to delve into the dynamics of misinformation dissemination in the digital age by examining the distinctions between untrusted news spreaders and other users. To this end, we have compiled a substantial sample of untrusted news spreaders and the general content shared of these users in Twitter. Using this large corpus stemming from three diverse datasets (MBFC, FNN and PAN), we have analysed the disparities in their language usage.

The initial exploration of traits associated with untrusted news spreaders, including the presence of hate speech, did not necessarily reveal the distinctions we anticipated. Other language features such as sentiment and emotional content indicate the existence of relatively small language differences between the two groups of users. These differences provide valuable insights that can inform the development of systems designed to identify and counteract malicious accounts. In particular, our results suggest that misinformation mitigation efforts should be focused on the specific content shared, rather than in profiling individual accounts.

7 Limitations

While we strived to derive insights from a large dataset using state-of-the-art classifiers and a robust analytical setup, we acknowledge the presence of factors that constrain the depth of our findings. For example, the focus on English-language content, potentially limiting the scope of global social media interactions and perspectives. Additionally, the exclusive use of Twitter data might not fully represent the dynamics on other social media platforms. While verified accounts are employed as a control group, it should be noted that they may not serve as a perfect control due to factors like their popularity, potential biases, or unique behaviours. Furthermore, the extraction of users relies on heuristics, introducing some degree of noise and potential inaccuracies in the data. Finally, we made use of automatic models based on transformers. While these have been tested extensively in prior work, there are inherent limitations in these models, as well as possible unwanted biases. All these limitations should be considered when interpreting our results and conclusions.

8 Ethical Statement

In our study involving user-generated content from social media, we ensured user privacy in several ways. First, we replaced all user mentions in the texts with placeholders and removing user IDs. Moreover, all the data utilised in our research is sourced from publicly available information or collected using the official Twitter API. Finally, all the information is provided in an aggregated fashion, without reporting sensitive information from individual users.

While our dataset and methodology have the potential for analysing individual behaviours, our primary objective is to offer researchers a valuable tool for the analysis and aggregation of social media content.

References

2023. Facebook MAU worldwide 2023 | Statista — https. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/. [Accessed 8-10-2023].

M. A. V. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares. 2021. Sentiment analysis for fake news detection. *Electronics*, 10:1348.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400.

Reema Aswani, Arpan Kumar Kar, and P Vigneswara Ilavarasan. 2019. Experience: managing misinformation in social media—insights for policymakers from twitter analytics. *Journal of Data and Information Quality (JDIQ)*, 12(1):1–18.

I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. W. Hansen, C. Hansen, and J. G. Simonsen. 2019. Multifc: a real-world multi-domain dataset for evidence-based fact checking of claims. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conferen*.

D. Baishya, J. J. Deka, G. Dey, and P. K. Singh. 2021. Safer: sentiment analysis-based fake review detection in e-commerce using deep learning. *SN Computer Science*, 2.

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

Jakab Buda and Flora Bolonyai. 2020. An ensemble model using n-grams and statistical features to identify fake news spreaders on twitter. In *CLEF (Working Notes)*.

Jose Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez-Cámara. 2022. Tweetnlp: Cutting-edge natural language processing for social media. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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

Sagar Deshmukh. 2015. Analysis of whatsapp users and its usage worldwide. *International Journal of Scientific and Research Publications*, 5(8):1–3.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055.

M. F. Er and Y. B. Yilmaz. 2023. Which emotions of social media users lead to dissemination of fake news: sentiment analysis towards covid-19 vaccine. *Journal of Advanced Research in Natural and Applied Sciences*, 9:107–126.

Sarah Evanega, Mark Lynas, Jordan Adams, Karinne Smolenyak, and Cision Global Insights. 2020. Coronavirus misinformation: quantifying sources and themes in the covid-19 'infodemic'. *JMIR Preprints*, 19(10):2020.

Bilal Ghanem, Simone Paolo Ponzetto, and Paolo Rosso. 2020. Factweet: profiling fake news twitter accounts. In *Statistical Language and Speech Processing: 8th International Conference, SLSP 2020, Cardiff, UK, October 14–16, 2020, Proceedings 8*, pages 35–45. Springer.

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

Abdullah Hamid, Nasrullah Shiekh, Naina Said, Kashif Ahmad, Asma Gul, Laiq Hassan, and Ala Al-Fuqaha. 2020. Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case. *arXiv preprint arXiv:2012.07517*.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

O. Inwood and M. Zappavigna. 2023. Conspiracy theories and white supremacy on youtube: exploring affiliation and legitimation strategies in youtube comments. *Social Media + Society*, 9:205630512211504.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Eric Lazarski, Mahmood Al-Khassaweneh, and Cynthia Howard. 2021. Using nlp for fact checking: A survey. *Designs*, 5(3):42.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-Tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? *ACL 2020*, page 36.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Dhiraj Murthy. 2018. *Twitter*. Polity Press Cambridge.

Preslav Nakov and Giovanni Da San Martino. 2020. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19.

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. 2021. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007.

Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11):4944–4957.

Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2023. Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 65–74.

Juan Pizarro. 2020. Using N-grams to detect Fake News Spreaders on Twitter—Notebook for PAN at CLEF 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Francisco Rangel, Anastasia Giachanou, Bilal Hisham Hasan Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR workshop proceedings*, volume 2696, pages 1–18. Sun SITE Central Europe.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. *arXiv preprint arXiv:2304.14238*.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Qi Su, Mingyu Wan, Xiaoqian Liu, Chu-Ren Huang, et al. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1-2):1–13.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Shreya Verma, Aditya Paul, Sharat S Kariyannavar, and Rahul Katarya. 2020. Understanding the applications of natural language processing on covid-19 data. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1157–1162. IEEE.

Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22.

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

Brian E Weeks. 2015. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of communication*, 65(4):699–719.

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

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

A Computational Resources

An *NVIDIA GeForce RTX 4090* GPU was utilised for the experiments conducted:

- 18 hours for the inference process of (sentiment, emotion, topic) on the *MBFC*, *FNN*, and *Verified* datasets.
- 6 hours for the training of the Longformer model (Section *Spreader Detection Analysis*.

B Model Categories

B.1 Emotion Categories

The *twitter-roberta-base-emotion-multilabel* model classifies each entry in one or more of the following classes: *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust*.

Features	F1	Accuracy
text	74	74
text-s	**75**	**76**
text-e	72	72
text-t	69	69
text-st	71	71
text-se	73	73
text-et	68	69
text-set	72	72

Table 4: Comparative results of F1 scores and accuracy for various feature combinations using the XGB classifier on the *PAN* dataset. *s: sentiment, e: emotion, t: topic*

B.2 Hate Speech Categories.

The *twitter-roberta-base-hate-multiclass* model classifies each entry in one of the following classes: *not_hate, sexism, racism, religion, other, sexual_orientation, disability*

B.3 Topic Classification Categories

The *tweet-topic-21-multi* model assigns each tweet one or more topics from the following list: *arts_&_culture, business_&_entrepreneurs, celebrity_&_pop_culture, diaries_&_daily_life, family, fashion_&_style, film_tv_&_video, fitness_&_health, food_&_dining, gaming, learning_&_educational, music, news_&_social_concern, other_hobbies, relationships, science_&_technology, sports, travel_&_adventure, youth_&_student_life*

C Spreader Detection: *XGB*

Table 4 highlights the performance of each feature set, with the 'text and sentiment (text-s)' combination achieving the highest F1 score of 75 and accuracy of 76, suggesting it is the most effective combination for this analysis.

Figures 6b, 4, and 6a illustrate that the discrepancies in the distribution of the examined features, sentiment, emotion and hate speech respectively[6], between the XGB model's predictions and the *PAN* dataset are negligible, indicating that they exhibit comparable trends.

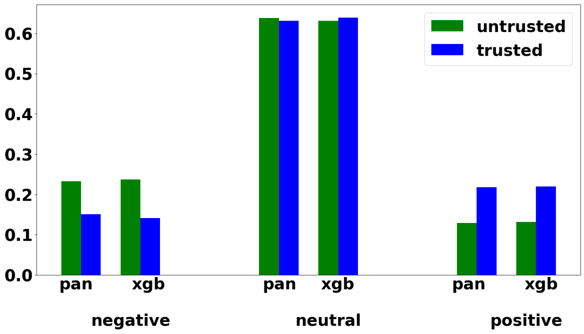


Figure 4: Sentiment comparison between *PAN* dataset and the *XGB* models' predictions.

[6] Results related to the topic distribution can be found in Appendix C.

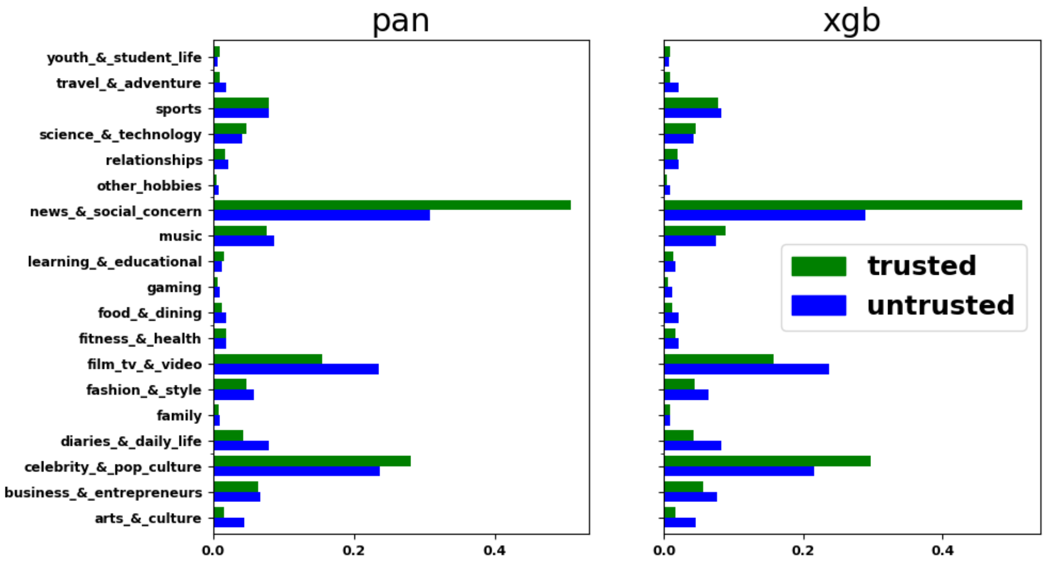
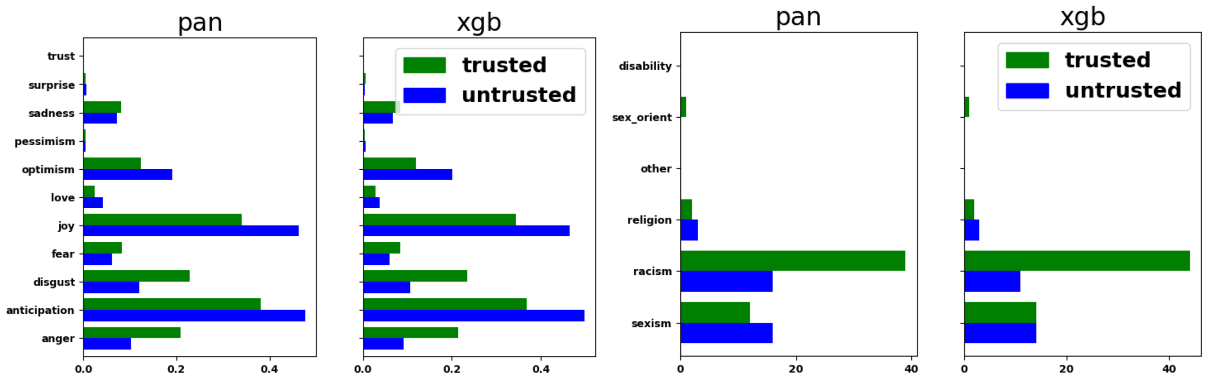


Figure 5: Topic distribution comparison between the *PAN* dataset and the predictions from the *XGB* model.



(a) Emotion distribution comparison between the *PAN* dataset and the predictions from the *XGB* model.

(b) Hate speech distribution comparison between the *PAN* dataset and the predictions from the *XGB* model.

Figure 6: Emotion & Hate speech results for *PAN* dataset and *XGB* model.