MiDe22: An Annotated Multi-Event Tweet Dataset for Misinformation Detection

Cagri Toraman[1]*, Oguzhan Ozcelik[2,3], Furkan Şahinuç[4]*, Fazli Can[2]
[1]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
[2]Department of Computer Engineering, Bilkent University, Ankara, Turkey [3]Aselsan, Ankara, Turkey
[4]Ubiquitous Knowledge Processing Lab (UKP Lab), Technical University of Darmstadt
ctoraman@ceng.metu.edu.tr
oguzhan.ozcelik@bilkent.edu.tr
furkan.sahinuc@tu-darmstadt.de
canf@cs.bilkent.edu.tr

Abstract

The rapid dissemination of misinformation through online social networks poses a pressing issue with harmful consequences jeopardizing human health, public safety, democracy, and the economy; therefore, urgent action is required to address this problem. In this study, we construct a new human-annotated dataset, called MiDe22, having 5,284 English and 5,064 Turkish tweets with their misinformation labels for several recent events between 2020 and 2022, including the Russia-Ukraine war, COVID-19 pandemic, and Refugees. The dataset includes user engagements with the tweets in terms of likes, replies, retweets, and quotes. We also provide a detailed data analysis with descriptive statistics and the experimental results of a benchmark evaluation for misinformation detection.

Keywords: Human-annotation, Misinformation detection, Multi-event dataset, Tweet

1. Introduction

With the growth of online social networks, people develop new behaviors and trends. An example is the amount of news consumed in these networks, and eventually the phrase “social media” is coined. However, considering their popularity and easy accessibility, it is inevitable to observe different kinds of content in social media platforms; e.g information manipulations, fake news, and misinformation/disinformation spread[1]. Twitter (rebranding to X since July 2023) is one of the platforms where misinformation can be widely spread as observed in the U.S. Elections (Grinberg et al., 2019), so that “fake news” became the Word of the Year in 2017 (CollinsDictionary, 2017).

Misinformation is spread in many domains including but not limited to health, politics, and disasters. Once misinformation is spread, the consequences can be devastating (Islam et al., 2020b; Reuters, 2022). For instance, many people died because of false rumors that claim that the cure for COVID-19 is drinking methanol (Islam et al., 2020b). Another example is that Ukraine sought an emergency order from the International Court of Justice due to the false claims of genocide against Russian speakers in Ukraine (Reuters, 2022). Considering the importance of misinformation spread in society and the ugly truth of unavoidable diffusion and beliefs, misinformation detection becomes a critical task that requires advanced methods and datasets.

A straightforward solution for misinformation is to avoid the spread in advance. However, people can be biased to change their beliefs even if corrections exist, and the attempts to correct falsehoods may not avoid its spread and even sometimes help its diffusion (Nyhan and Reifler, 2010). Moreover, targeted advertising to increase user engagement can help misinformation spread, which may be a source of revenue for social media platforms (Neumann et al., 2022).

We have four main observations on existing social media collections for misinformation detection. Although they mostly cover a limited number of topics (Ma et al., 2017), these topics remain too high-level to provide an opportunity to systematically examine which type of incidents trigger the misinformation spread. The availability of fine-grained event-specific information can play a significant role in capturing different user behaviors for detecting and preventing misinformation. Furthermore, the existing datasets focus on widely used languages such as English (D’Ulizia et al., 2021), while they are very limited for low-resource languages. Lastly, user engagements (like, reply, retweet, and quote) and media elements (image and video) in false tweets can be useful to analyze different types of information diffusion and detection methods (e.g. multimodal), but not all types are always included in the datasets.

In order to bridge these gaps, we present an annotated multi-event tweet dataset for **Mi**sinformation **De**tection under several recent

*Work partially done in Aselsan, Ankara, Turkey.

[1] We use misinformation as an umbrella term that refers to all instances where information have falsehoods.

LREC-COLING 2024, pages 11283–11295
20-25 May, 2024.

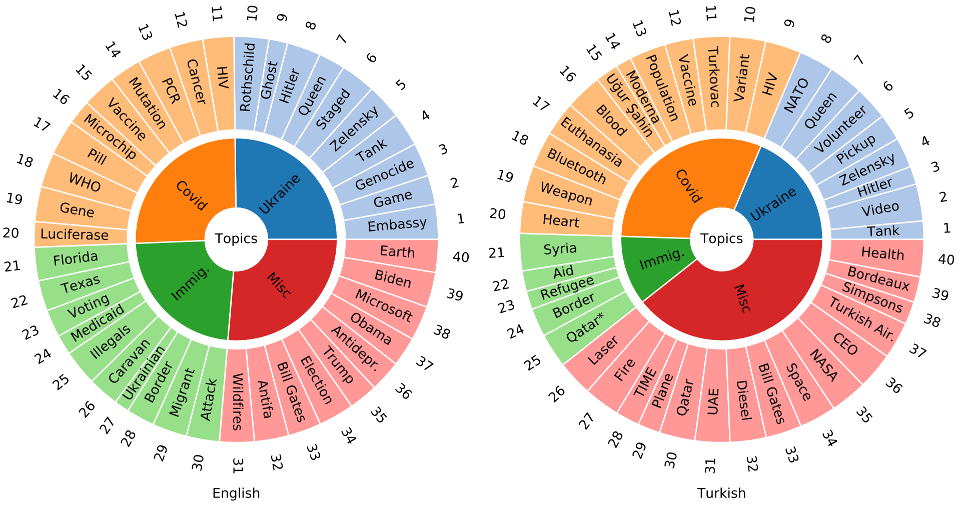


Figure 1: The topics (inner circle) and events (outer circle) in MiDe22 for English (left) and Turkish (right). The areas are proportional to the number of tweets they have.

events from 2020 to 2022, called MiDe22, including English and Turkish tweets with four types of user engagements and they are likes, replies, retweets, and quotes.

1.1. Dataset Contents

The MiDe22 dataset[2] consists of three parts: (i) Topics and Events, (ii) Tweets, and (iii) Engagements. Each part exists for both English (MiDe22-EN) and Turkish (MiDe22-TR).

Topics and Events. We consider the issues occupying the world's agenda in recent years as the topics of our dataset. Then, we extract the significant events with the highest spread of misinformation. Figure 1 presents an overview of the structure of our dataset. The inner circle indicates the COVID-19 pandemic, the 2022 War between Russia and Ukraine, Refugees (Immigration), and Miscellaneous events that are not categorized under the previous topics. Overall, these topics contain 40 newsworthy events in the outer circles of the figure. We also provide the event titles along with their topics online[2].

Note that we prefer well-known recent events for both languages. The reason is that some misinformation events can be global and observed in several countries, such as "COVID-19 vaccines contain Human Immunodeficiency Virus (HIV)". These common events can provide an opportunity to inspect how misinformation is spread in different languages. On the other hand, there are local events that have influence in specific regions. The details on the events are given in Section 3.1.

Tweets. The dataset has tweets related to the events. The crawling process is explained in Section 3.1. Each tweet is labeled according to three classes: False information, True information, and Other. The Other class includes tweets that cannot be categorized under false and true information. The annotation process is explained in Section 3.2.

User Engagements and Media. We provide the user engagements with all tweets. Separate engagement splits are provided in the types of like, reply, retweet and quote. We also provide media elements in our dataset, i.e. image and video if they exist in the tweets.

1.2. Contributions

Our contribution involves the development of a novel tweet dataset for misinformation detection in two languages with various topics and user engagements. The languages are a widely used language: English, and a low-resource language: Turkish. The topics of the dataset cover several recent events, such as the 2022 Russia-Ukraine War and the COVID-19 pandemic. The dataset includes the user engagements with all tweets in terms of likes, replies, retweets, and quotes. It can be used in many studies such as misinformation, event, and topic detection. Additionally, we conduct experiments to provide initial baseline scores from different model families, e.g., bag-of-words, neural, and transformer-based models. Apart from demonstrating the quality and utility of our dataset, these baselines also provide a benchmark for researchers to compare against and further enhance their developments. The variety of baseline models is rich enough to perform statistical tests and interpret the results properly.

[2] The dataset and all other related documents can be accessed at https://github.com/metunlp/MiDe22

Dataset Name	Langs.	Domain	Topics	Date of Data	Engagements	Size	Labels
LIAR (Wang, 2017)	En	Statements	MISC	2007-2016	None	12.8k	Annotated
FakeNewsNet (Shu et al., 2020)	En	News, tweets	MISC	n/a	None	23.1k, 1.9m	Query
CoAID (Cui and Lee, 2020)	En	News, tweets	C19	2019-2020	Reply	4.2k, 160k	Query
COVIDLies (Hossain et al., 2020b)	En	Tweets	C19	2020	None	6.7k	Annotated
CMU-MisCOV19 (Memon and Carley, 2020)	En	Tweets	C19	2020	None	4.5k	Annotated
MM-COVID (Li et al., 2020)	6 langs.	Tweets	C19	n/a	Reply, retweet	105.3k	Query
VaccineLies (Weinzierl and Harabagiu, 2022)	En	Tweets	C19, HPV	2019-2021	None	14.6k	Annotated
MuMin (Nielsen and McConville, 2022)	41 langs.	Tweets	MISC	n/a	Reply, retweet	21.5m	Query
MR2 (Hu et al., 2023)	En, Zh	Tweets, Weibo	MISC	2017-2022	Reply, retweet	14.7k	Annotated
MiDe22 (this study)	En, Tr	Tweets	RUW, C19, IMM, MISC	2020-2022	Reply, retweet, like, quote	10.3k	Annotated

Table 1: **Related misinformation studies**. RUW stands for Russia-Ukraine War, C19 for COVID-19, IMM for Immigration and Refugees, HPV for Human Papilloma Virus, and MISC for Miscellaneous. The last column shows if tweets are annotated by humans, or labeled by the output of queries to Twitter API. Size is given in terms of number of tweets.

2. Related Work

In this section, we provide a brief review of the existing literature and explore the methods used for the analysis and detection of misinformation, the available datasets for research purposes, and the various interventions implemented to combat the spread of misinformation.

2.1. Misinformation Analysis

Misinformation analysis is the process of identifying, evaluating, and understanding the spread and impact of false, misleading, or inaccurate information. Misinformation modeling covers temporal and patterns of information diffusion to analyze spread (Shin et al., 2018; Rosenfeld et al., 2020), and also analysis of misinformation spreads during important events such as the 2016 U.S. Election (Grinberg et al., 2019), the COVID-19 Pandemic (Ferrara et al., 2020), and the 2020 BLM Movement (Toraman et al., 2022a).

2.2. Misinformation Detection

Misinformation detection is a challenging task when the dynamics subject to misinformation spread are considered. The task is also studied as fake news detection (Zhou and Zafarani, 2020), rumor detection (Zubiaga et al., 2018), and fact/claim verification (Bekoulis et al., 2021; Guo et al., 2022).

There are two important aspects of misinformation detection. First, the task mostly depends on supervised learning with a labeled dataset. Second, existing studies rely on different feature types for automated misinformation detection (Wu et al., 2016). Text contents are represented in a vector or embedding space by natural language processing (Oshikawa et al., 2020) and the task is formulated as classification or regression mostly solved by deep learning models (Islam et al., 2020a). The features extracted from user profiles can be used to detect the spreaders (Lee et al., 2011). Besides contents, there are efforts to extract features from the network structure such as network diffusion models (Kwon and Cha, 2014; Shu et al., 2019a) and graph neural networks (Mehta et al., 2022). Lastly, external knowledge sources (Shi and Weninger, 2016; Toraman et al., 2022b) and the social context among publishers, news, and users (Shu et al., 2019b) can be integrated to the learning phase.

Rather than identifying the content with misinformation, there are efforts to detect the user accounts that would spread undesirable content such as spamming and misinformation. Social honeypot (Lee et al., 2011) is a method to identify such users by attracting them to engage with a fake account, called honeypot. There are also bots producing computer-generated content to promote misinformation (Himelein-Wachowiak et al., 2021).

2.3. Misinformation Datasets

There are several efforts in the literature to construct a dataset for misinformation detection. The LIAR dataset (Wang, 2017) includes short statements from different backgrounds, annotated by PolitiFact API. News and related tweets for fact-checked events are composed in a dataset in (Shu et al., 2020). Recently, global events and their repercussions in social media lead to the emergence of new misinformation datasets. For instance, Memon and Carley (2020) annotate tweets according to misinformation categories such as fake treatments for COVID-19. In (Li et al., 2020), news sources are investigated for fake news in different languages. Hossain et al. (2020b) retrieve common misconceptions about COVID-19, and label tweets according to their stances against misconceptions. Weinzierl and Harabagiu (2022) compose the vaccine version of the same dataset. Other datasets include COVID-19 healthcare misinformation (Cui and Lee, 2020), and large-scale multimodal misinformation (Nielsen and McConville, 2022). (Hu et al., 2023) curate annotated multimodal social media dataset for two widely-spoken languages (English and Chinese), providing reply and retweet engagements. Lastly, there are very limited datasets for low-resource languages (Hossain et al., 2020a; Lucas et al., 2022) but do not exist for Turkish.

Classes	Sample Sentence
True	No, WHO's Director-General Didn't Say COVID Vaccines Are 'Being Used To Kill Children'.
False	The director-general of the WHO and I quote: "countries are using the vaccine to kill children".
Other	Africa moving toward control of COVID-19: WHO director.

Table 2: Sample sentences from MiDe22. The event number is EN18.

2.4. Misinformation Intervention and Generative AI

Misinformation intervention is the task of reducing the negative effects of spread in advance. One way to fight against misinformation is to spread true information by cascade modeling (Budak et al., 2011). Other methods include detecting credible information (Morstatter et al., 2014), cost-aware intervention (Thirumuruganathan et al., 2021), and crowdsourcing (Twitter, 2022). However, with the recent success of transformer-based generative models, such as ChatGPT[3], it becomes more difficult for a human reader to assess and interfere with the credibility of the news source (Hsu and Thompson, 2023). Recent studies (Zellers et al., 2019; Spitale et al., 2023) reveal that social media users cannot distinguish manipulative contents generated by Generative AI (Brown et al., 2020) and humans.

2.5. Our Differences

In Table 1, we summarize notable datasets in the literature and compare them with our dataset. We aim to provide a resource for misinformation detection and analysis, rather than intervention. Different from existing works, our study covers several recent events for misinformation analysis, including the 2022 Russia-Ukraine War, providing human-annotated tweets and user engagements on Twitter.

3. Dataset Construction

3.1. Data Crawling

There are 40 events under four topics per language (English and Turkish). We manually browsed fact-checking platforms (PolitiFact.com, EuVsDisinfo.eu, UsaToday.com/FactCheck for English, and Teyit.org for Turkish), and manually selected all events related to our topics at the beginning of April 2022[4]. The events range from September 10th, 2020 to March 21st, 2022 in English, and October 5th, 2020 to March 11th, 2022 in Turkish. To find relevant tweets for events, we used a predetermined set of keywords for each event. At this point, we emphasize that the main criteria of keyword selection is to reach the critical mass in terms of the number of tweets for a given event. Therefore, there is not any bias stemming from keywords towards True and False labels. The details of tweet crawling and query structure are given in Appendix A.1. We collected tweets via Twitter API's Academic Research Access[5].

Each event is represented by 11 attributes: Event's language, id, topic, title, URL for evidence, the keywords for querying tweets, the date when evidence is provided, the start date of querying tweets, the end date of querying tweets, the keywords used while querying tweets for the Other class, sample tweet ID(s) in this event. The tweets range from September 19th, 2020 to April 5th, 2022 in English, and September 15th, 2020 to April 5th, 2022 in Turkish.

We excluded retweets to avoid duplicates. We used Dice similarity (Schütze et al., 2008), and applied a similarity threshold (85%) between a newly collected tweet and previous tweets. If it exceeded the threshold, then we skipped that tweet and collected another tweet. We did not set a limit on tweet length, since misinformation can be spread by a few or no words using media objects. We kept the original contents, and provided links to the images and external URLs in tweets. We also collected all user engagements returned to our queries in the types of like, reply, retweet and quote.

3.2. Data Annotation and Statistical Authentication

Each tweet in the dataset is labeled according to three classes: True information, False information, and Other. The True class includes tweets with the correct information regarding the corresponding event. The False class includes tweets with misinformation on the corresponding event. The Other class includes tweets that cannot be categorized under false and true information. In general, these tweets include opinions or information related to the events, which cannot be directly judged as True or False. Sample sentences for each class label are given for the EN18 event in Table 2. The

[3] https://chat.openai.com

[4] Some of the events were later filtered out due to the insufficient number of tweets.

[5] Note that legacy Twitter API accesses (e.g., Academic Research Access) were deprecated in April 2023. New API accesses are provided on https://developer.twitter.com/en/products/twitter-api

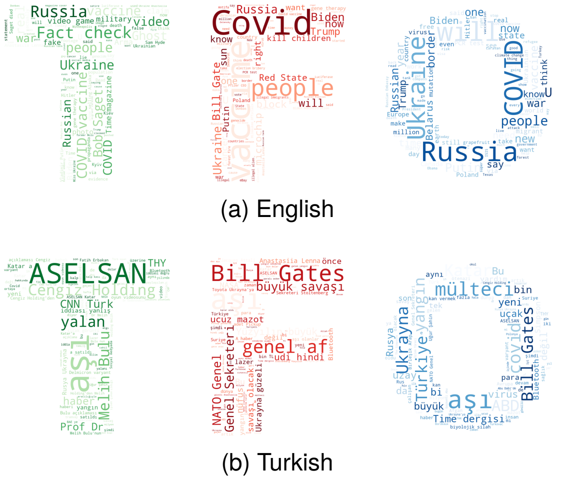
Statistics		MiDe22-EN			MiDe22-TR		
		True	False	Other	True	False	Other
Tweets		727	1,729	2,828	669	1,732	2,663
User Engagements	Like	11,662	8,587	33,086	16,594	24,076	30,446
	Reply	853	1,065	3,291	1,316	1,528	2,677
	Retweet	2,839	3,127	9,106	3,055	5,333	6,442
	Quote	339	451	2,673	682	858	1,649
Average	Like	16.04±168.61	4.97±36.65	11.70±153.55	24.80±122.33	13.90±85.52	11.43±64.16
	Reply	1.17±11.95	0.62±3.82	1.16±11.51	1.97±11.66	0.88±4.81	1.01±17.50
	Retweet	3.91±43.35	1.81±15.69	3.22±39.11	4.57±27.14	3.08±18.96	2.42±16.26
	Quote	0.47±4.55	0.26±1.56	0.95±23.33	1.02±5.62	0.50±5.64	0.62±11.00
	Image	0.10±0.30	0.08±0.27	0.13±0.34	0.17±0.37	0.22±0.42	0.17±0.37
	Video	0.01±0.08	0.03±0.17	0.03±0.18	0.07±0.25	0.05±0.22	0.05±0.22

Table 3: The main statistics of our dataset for English (EN) and Turkish (TR). The mean and standard deviation among tweets for each attribute are given.

event is about the speech of T. A. Ghebreyesus, the Director-General of World Health Organization (WHO), during the opening of the WHO Academy in Lyon. Ghebreyesus stumbled over his words, first mispronouncing the word "children" led some people to claim he said "kill children".

We assigned five annotators who were computer engineering undergraduate students. We developed an annotation tool based on INCEpTION (Klie et al., 2018) for ease of labeling. Before the annotation process started, all annotators were given a tutorial about the task. Explicit definitions of True and False tweets were provided along with the corresponding examples. We try to mitigate any bias such as political leanings and beliefs during annotation tutorials. Annotation guidelines are detailed in the online repository[2]. The annotations are designed so that each tweet is labeled by two annotators, and tweets are distributed randomly and almost evenly among annotators. If there was no agreement between two annotators, then a different annotator was assigned to label. We applied the majority voting in that case. If there was still no agreement among three annotators, then we removed it from the dataset by assuming that the tweet was problematic.

Since each tweet is annotated by at least two annotators, we calculate Krippendorf's alpha-reliability (Krippendorff, 1970) to measure inter-annotator agreement (IAA). The resulting alpha coefficients are 0.785 and 0.791 in English and Turkish, respectively. Regarding the study of Landis and Koch (1977), our dataset has substantial agreement among annotators. Furthermore, the IAA scores of MiDe22 are higher than or similar to those of existing datasets (Nakamura et al., 2019; Pérez-Rosas et al., 2017; Wang, 2017; Nguyen et al., 2020).



(a) English

(b) Turkish

Figure 2: Word clouds for most frequently observed keywords in the (a) English and (b) Turkish datasets for True, False, and Other. Collocations are calculated within a window size of two consecutive words.

4. Data Analysis

4.1. Quantitative Analysis

Our dataset is available in English and Turkish. The tweet counts together with the average numbers of user engagements (like, reply, retweet, and quote) are listed in Table 3. The average number of all types of user engagements per true tweet is higher than false tweets in both languages. On the other hand, there is no significant difference in the average number of images and videos. Nevertheless, false tweets in Turkish have more images, while false tweets in English have slightly more videos, compared to true ones.

4.2. Content Analysis

Figure 2a and 2b show the word clouds for each class (true, false, and other). Although the dataset

	Classes	Emoji and Hashtags with the ratio of frequencies				
MiDe22-EN	True	😂 0.37	❎ 0.10	🤦‍♂️ 0.09	🤣 0.09	👇 0.07
	False	😂 0.24	👇 0.15	🤔 0.11	🤣 0.08	🚨 0.08
	Other	😂 0.07	🤣 0.07	🔥 0.05	🤔 0.05	🇺🇦 0.04
	True	#Ukraine 0.16	#FactCheck 0.10	#Russia 0.09	#FakeNewsAlert 0.07	#FactsMatter 0.06
	False	#Ukraine 0.12	#Poland 0.06	#Zelensky 0.05	#COVID19 0.05	#Russia 0.04
	Other	#Ukraine 0.15	#Poland 0.06	#Zelensky 0.05	#COVID19 0.05	#Russia 0.04
MiDe22-TR	True	✖ 0.41	😂 0.15	✅ 0.14	😀 0.09	🔍 0.08
	False	😂 0.18	👇 0.10	🤣 0.08	🔴 0.07	🤔 0.06
	Other	😂 0.17	🔴 0.10	👇 0.08	🤣 0.08	🔥 0.06
	True	#Ukrayna 0.10	#sondakika 0.05	#Rusya 0.05	#cnnturk 0.05	#haber 0.04
	False	#Ukrayna 0.06	#UkraineRussiaWar 0.06	#SONDAKİKA 0.04	#Ukraine 0.03	#worldwar3 0.03
	Other	#Ukrayna 0.06	#Rusya 0.05	#uzay 0.03	#mülteci 0.03	#SONDAKİKA 0.03

Table 4: **Top-5 most frequent emoji and hashtags for each class (row) with their frequency ratios.** The ratio is the number of emoji/hashtag divided by the number of tweets with that emoji/hashtag.

includes several topics, COVID-related keywords are observed more in false tweets for both languages (e.g. "aşı" translated as "vaccine"), and also political figures (e.g. Biden, Trump, and Putin). On the contrary, we observe fact-checking keywords in true tweets (e.g. "yalan" translated as "lie").

We also provide the top five most frequently observed emojis and hashtags in Table 4. When smiling and laughing emojis are discarded, true tweets include mostly cross signs that would represent false information. False tweets contain the pointing down emoji that would point a message to readers, and also a thinking emoji that would emphasize the false information to readers. In terms of hashtags, we find that most of the hashtags are related to the 2022 Russia-Ukraine War. In English, there are fact-checking hashtags in true tweets (e.g. #FakeNewsAlert), while a similar kind of hashtag is not observed in Turkish.

4.3. Temporal Analysis

We provide the distribution of the tweets for each topic in Figure 3. Most of the events gain popularity rapidly, reach their peak, and fall from the grace after a while. Most of the distributions exhibit a bimodal Gaussian shape, meaning that there is a second peak point (local or global) for the distribution. There is a stimulus that makes the event regain its popularity. In our early analyses, we find that the dates of the turning point in distributions coincide with the dates of fact-checking news.

When we examine the distribution of events according to tweet posting date, we observe both similar and different patterns among topics. For instance, COVID-19 events can last up to six months (Figure 3b), since it is a long-term incident. A similar pattern also exists in Turkish with events lasting more than four months (Figure 3e). The Russia-Ukraine War is a fresh topic, covering a relatively shorter time period. However, Figures 3a and 3d show that there are different events that could lead to the spread of misinformation in this short period of time. Overall, we argue that detection algorithms can be developed based on the event's life span, e.g. user engagements for the long-term and context for short-term events.

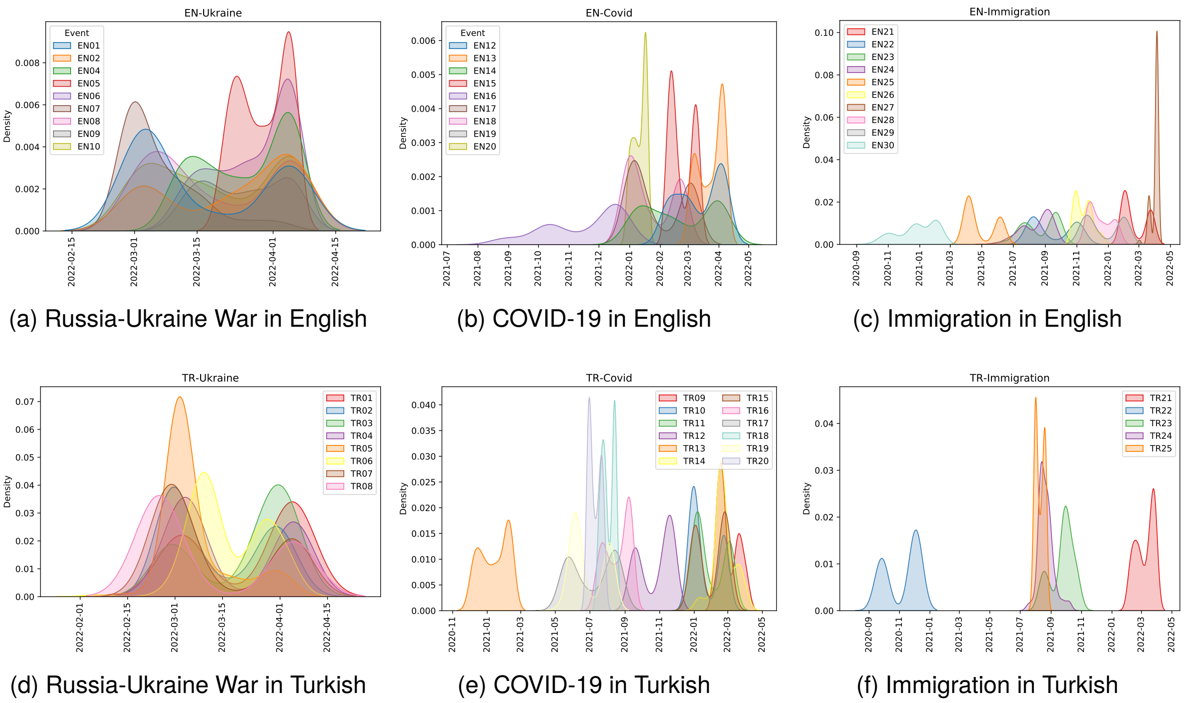
5. Experiments

In order to understand if the constructed dataset is adequate in terms of task difficulty, we target misinformation detection using only tweet text. We leave utilization of user engagements for future work. We implement total of eight benchmark models (see Table 5) from the following model families:

Bag-of-Words: We consider conventional machine learning classifiers based on the bag-of-words model since tweets can include specific terms and phrases used for reporting manipulated news (e.g. "Did you know?") and correcting falsehood (e.g. "FactCheck"). We implement a linear Support Vector Machine (SVM) (Vapnik, 1999) with TD-IDF vectors of each tweet using scikit-learn (Pedregosa et al., 2011). SVM is trained with a stopping criterion, i.e., 1e-3 tolerance. The remaining parameters are selected default.

Neural Models: We implement Long Short-term Memory (LSTM) (Hochreiter and Schmidhuber, 1996) and Bi-directional Long Short-Term Memory (BiLSTM) (Graves and Schmidhuber, 2005) with PyTorch (Paszke et al., 2019). The embedding size is 125, and there are 50 units in each layer. After the LSTM layers, there is a dense layer with perceptrons of the same size of units. Next, there is a dropout layer with a probability of 0.5. We use the sigmoid activation function. They are trained during 20 epochs, where we set a learning rate of 1e-3 with a batch size of 16.

Transformer-based Language Models: We use BERT base uncased (Devlin et al., 2019) and DeBERTa (He et al., 2021) pretrained with English corpus, BERTurk uncased base model (Schweter, 2020) for Turkish corpus, and mBERT base un-



(a) Russia-Ukraine War in English

(b) COVID-19 in English

(c) Immigration in English

(d) Russia-Ukraine War in Turkish

(e) COVID-19 in Turkish

(f) Immigration in Turkish

Figure 3: **Temporal distribution of tweets by topics**. The y-axis represents the density of tweet counts. The x-axis represents the date that tweets are shared. The events EN03 and EN11 are neglected due to the shorter time range.

cased (Devlin et al., 2019) and XLM-R base (Conneau et al., 2020) for multilingual corpus, by HuggingFace (Wolf et al., 2020). We use cross-entropy loss and set a learning rate of 5e-5 with a batch size of 16 during 10 epochs. The number of tokens is set to 128 with padding and truncation, where each tweet is an input sequence.

We apply stratified (in terms of both classes and events) 5-fold cross-validation to get an average performance score for robustness. The training of SVM, LSTM, and BiLSTM is performed on Intel Core i9-10900X 3.70GHz CPU with 20 cores with 128 GB memory. The fine-tuning of large language models is performed on a single NVIDIA RTX A4000.

5.1. Experimental Results

We report the performance of models in Table 5. We observe that Transformer-based language models outperform conventional methods (SVM, LSTM, and BiLSTM) in both languages. However, SVM has a better performance than LSTM and BiLSTM. The reason could be the distribution of words in the false and true tweets. SVM is trained on a Bag-of-Words model where features represent the importance score of individual words.

Among language models, DeBERTa has the highest performance in English, while a multilingual model, XLM-R, in Turkish. This observation can show the generalization capability of multilingual models for low-resource languages, such as Turkish, in misinformation detection.

The performance of misinformation detection in terms of F1-Score can be observed differently in other annotated datasets: 90.07 in (Weinzierl and Harabagiu, 2022), 50.20 in (Hossain et al., 2020b), and 83.95 in our study (no F1 score reported in (Wang, 2017) and no detection score in (Memon and Carley, 2020)). We argue that the performance depends on datasets since misinformation detection is a dynamic task where context changes rapidly. The context can be integrated into the learning phase via knowledge sources (Pan et al., 2018; Toraman et al., 2022b) to adapt to the dynamic nature of the misinformation task.

6. Discussion

6.1. Possible Use Cases

The dataset can be used for several tasks in natural language processing, information retrieval, and computer vision. Misinformation detection with textual features (Su et al., 2020) or visual features (Cao et al., 2020) is the primary objective of this dataset. Multimodal approaches with both textual and visual features have also promising results (Khattar et al., 2019). Other opportunities include the analysis of information diffusion (Shin et al.,

Model	MiDe22-EN			MiDe22-TR		
	Precision	Recall	F1	Precision	Recall	F1
SVM	79.29±0.9	79.00±0.8	78.33±0.9	78.78±1.1	78.60±1.1	78.20±1.1
LSTM	72.71±2.2	72.94±2.0	72.24±1.8	72.59±0.9	72.71±0.9	72.23±0.8
BiLSTM	73.31±0.7	73.61±0.8	73.30±0.7	74.16±0.7	74.21±0.8	73.77±0.9
BERT	82.44±0.8	82.33±1.0	82.35±0.9	78.63±1.3	78.39±1.4	78.46±1.4
DeBERTa	**84.04±0.7**	**83.94±0.7**	**83.95±0.7**	76.03±2.2	75.14±2.3	75.30±2.2
mBERT	78.87±2.9	78.37±4.1	78.21±4.3	78.63±1.0	78.20±1.2	78.26±1.2
XLM-R	79.19±3.0	78.94±3.0	79.01±3.0	**83.18±1.1**	**82.76±1.1**	**82.82±1.1**
BERTurk	78.31±1.5	78.36±1.6	78.30±1.6	82.89±1.1	82.52±1.1	82.58±1.1

Table 5: **The results of benchmark models for Misinformation Detection on MiDe22**. The average score of five folds with standard deviation is reported in terms of weighted precision, recall, and F1 scores. The best scores for each dataset and metric are given in bold.

2018) and bot accounts using tweet conversations (Çetinkaya et al., 2020) since the dataset has the user engagements for all tweets. Furthermore, the efforts to detect the events of social media posts, i.e. event or topic detection (Şahinuç and Toraman, 2021), can benefit due to the variety of events in the dataset.

6.2. Difficulties Encountered

Finding relevant tweets to events was a challenging task. We run different queries with different keywords to fetch the highest number of relevant tweets. Although we used Twitter Academic Access API, we could not increase the number of tweets relevant to events. Another difficulty was the guidance of annotators in this dataset. We tried to guide the annotators to be as objective and unbiased as possible by providing a guidelines document and a dedicated live video seminar, where we explained the events, claims and evidences, annotation tool, and example annotations.

7. Conclusion

We curate a multi-event tweet dataset for misinformation detection that has novelty in terms of the variety of languages (English and Turkish), topics (various topics and 40 events per language), and engagements (like, reply, retweet, and quote). We further analyze the dataset and provide benchmark experiments including the performances of state-of-the-art models. We publish the dataset and the files related to the dataset curation for transparency. They provide new opportunities for researchers from different backgrounds including but not limited to natural language processing, social network analysis, and computer vision.

In future work, we plan to develop new models on our dataset for various tasks such as multimodal detection and adversarial attacks for misinformation. Cross-lingual misinformation spread is another opportunity since our dataset covers two languages with overlapping events. We can also extend our study to other social media platforms for cross-platform misinformation detection.

8. Limitations

We acknowledge a set of limitations in this study. First, creating a misinformation dataset is more difficult than other types of tweet datasets due to the regulations of social media platforms. Making the dataset balanced in terms of labels can be therefore challenging. Second, we decided on the events included in the dataset manually by browsing the fact-checking platforms such as PolitiFact.com and Teyit.org. Furthermore, human annotation is a costly and laborious process.

In this study, we labor five annotators to label tweets due to budget and time limitations. The annotators were given careful guidelines on the topics and definitions of class labels. However, the dataset can still reflect their personal biases and interpretations to some extent. Recent advances in generative AI can be also integrated to generate label annotations (Zhu et al., 2023). Lastly, our study focuses on the English and Turkish languages only, which might reflect the cultural biases exposed by newsletters and fact-checkers. There could be different instances for the same topics in other languages.

9. Ethical Concerns

We consider the ethical concerns regarding the stakeholders in misinformation detection (Neumann et al., 2022). First, all sources of information (tweet author) should be treated equally. We collected the tweets returned to our API queries without discriminating or selecting authors. Second, subjects of information (the subject in tweet content) should be represented fairly and accurately.

Language	English
ID	EN2
Topic	Ukraine
Title	Viral clip shows ’Arma 3’ video game not war between Russia and Ukraine
Evidence URL	https://www.usatoday.com/story/news/factcheck/2022/02/21/.../6879521001/
Query keywords	arma 3 russia ukraine
Evidence date	2022-02-21
Query start date	2021-12-21
Query end date	2022-04-06
Other keywords	russia ukraine war video
Sample tweet(s)	1499460925253832707
Query-1	arma 3 russia ukraine lang:en (has:media OR has:geo) -is:retweet 2021-12-21 2022-04-06 100
Query-2	arma 3 russia ukraine lang:en -is:retweet 2021-12-21 2022-04-06 100
Query-3 for Other	russia ukraine war video lang:en (has:media OR has:geo) -is:retweet 2021-12-21 2022-04-06 50
Query-4 for Other	russia ukraine war video lang:en -is:retweet 2021-12-21 2022-04-06 50

Table 6: An example event in the dataset. A part of URL is cropped due to space constraints.

Since tweets may include false claims about the subject of information, we included true tweets that refute the claims as well. Third, all seekers of information (the audience of tweet authors) should obtain relevant and high-quality information. Distributive justice is out of context since our focus is to detect misinformation not to distribute tweets to the audience. Lastly, individuals/organizations should generate fair evidence with testimonial justice. We assigned annotators to label the data according to a guidelines document that includes the details of events, claims, and corrections with sources of evidence[6].

We obtain an internal IRB approval for our misinformation detection dataset study, which includes the approval of two reviewers.

In order to provide transparency (Bender et al., 2021; Baeza-Yates, 2022), we publish the files related to data crawling and annotation: The queries and details of the events, the annotation guidelines document, video seminar recording for directing annotators, and the details of the annotation tool[2].

A. Appendix

A.1. Tweet Crawling

An example query for crawling tweets from Twitter API for a specific event is given in Table 6. We manually determined the events and query keywords by browsing events in fact checking web pages. The motivation for using a different keyword set for the Other class is that we might not find irrelevant tweets or tweets with no information with the query keywords prepared for the True and False classes. The start and end dates of querying tweets are selected before and after two months of the evidence date, unless restricted by the crawling date (2022-04-06).

We run four consecutive queries for each event. We first collected tweets with media object (image, video, or GIF) and geographic location tags (Query-1). If the number of such tweets was not enough to fulfill the number of target tweets (50 tweets per class, total of 100 tweets for the True and False classes), then we collected tweets without media objects and geographic location tags (Query-2). After running queries for the True and False classes, we collected tweets for the Other class with the same approach by first searching for media objects (Query-3) and then regular tweets (Query-4).

We set the highest number of tweets to be collected for each class (true, false, other) to 50 tweets to provide a balance among classes and limit the total number of tweets to be annotated.

A.2. Event List

There are four topics in the dataset. The topics are the 2022 Russia-Ukraine War, COVID-19 pandemic, Refugees (Immigration), and Miscellaneous. There are 40 events under four topics for both languages. The list of events along with their topics are published online[2].

A.3. User Engagements

The detailed list of tweets and user engagements (like, retweet, reply, and quote) per event are published online[2].

A.4. Annotation Tool

The details of the annotation tool are published online[2].

[6] We relied on trust-worthy fact-checking platforms as sources of evidence: https://www.politifact.com, https://euvsdisinfo.eu, and https://eu.usatoday.com/news/factcheck for English, and https://teyit.org for Turkish.

B. Bibliographical References

Ricardo Baeza-Yates. 2022. Ethical challenges in ai. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1–2.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1):1–35.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tom Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 665–674, New York, NY, USA. Association for Computing Machinery.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161.

Yusuf Mücahit Çetinkaya, İsmail Hakkı Toroslu, and Hasan Davulcu. 2020. Developing a Twitter bot that can join a discussion using state-of-the-art architectures. *Social Network Analysis and Mining*, 10(1):1–21.

CollinsDictionary. 2017. Collins 2017 word of the year shortlist. (Accessed Oct 19, 2023).

Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Furkan Şahinuç and Cagri Toraman. 2021. Tweet length matters: A comparative analysis on topic detection in microblogs. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*, page 471–478, Berlin, Heidelberg. Springer-Verlag.

Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

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

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: A survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Emilio Ferrara, Stefano Cresci, and Luca Luceri. 2020. Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science*, 3(2):271–277.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S presidential election. *Science*, 363(6425):374–378.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing.

McKenzie Himelein-Wachowiak et al. 2021. Bots and misinformation spread on social media: Implications for COVID-19. *J Med Internet Res*, 23(5):e26933.

Sepp Hochreiter and Jürgen Schmidhuber. 1996. LSTM can solve hard long time lag problems. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, pages 473–479. MIT Press.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020a. BanFakeNews: A dataset for detecting fake news in Bangla. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2862–2871, Marseille, France. European Language Resources Association.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020b. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

T. Hsu and S. A. Thompson. 2023. Disinformation researchers raise alarms about a.i. chatbots. (Accessed: 19 Oct 2023).

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2901–2912, New York, NY, USA. Association for Computing Machinery.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020a. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):1–20.

Md Saiful Islam et al. 2020b. COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621 – 1629.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Sejeong Kwon and Meeyoung Cha. 2014. Modeling bursty temporal pattern of rumors. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):650–651.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Kyumin Lee, Brian Eoff, and James Caverlee. 2011. Seven months with the devils: A long-term study of content polluters on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 185–192.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Toward a multilingual and multimodal data repository for COVID-19 disinformation. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4325–4330.

Jason Lucas, Limeng Cui, Thai Le, and Dongwon Lee. 2022. Detecting false claims in low-resource regions: A case study of Caribbean islands. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 95–102, Dublin, Ireland. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Nikhil Mehta, Maria Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380, Dublin, Ireland. Association for Computational Linguistics.

Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing COVID-19 misinformation communities using a novel Twitter dataset. *arXiv preprint arXiv:2008.00791*.

Fred Morstatter, Nichola Lubold, Heather Pon-Barry, Jürgen Pfeffer, and Huan Liu. 2014. Finding eyewitness tweets during crises. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 23–27, Baltimore, MD, USA. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. 2022. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1504–1515, New York, NY, USA. Association for Computing Machinery.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1165–1174, New York, NY, USA. Association for Computing Machinery.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

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

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International Semantic Web Conference*, pages 669–683. Springer.

Adam Paszke et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Reuters. 2022. Russian no show at U.N. court hearing on ukrainian 'genocide'. (Accessed Oct 19, 2023).

Nir Rosenfeld, Aron Szanto, and David C. Parkes. 2020. A kernel of truth: Determining rumor veracity on Twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*, WWW '20, page 1018–1028, New York, NY, USA. Association for Computing Machinery.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133.

Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287.

Kai Shu, H Russell Bernard, and Huan Liu. 2019a. Studying fake news via network analysis: detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, pages 43–65. Springer.

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, Suhang Wang, and Huan Liu. 2019b. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

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.

Saravanan Thirumuruganathan, Michael Simpson, and Laks V.S. Lakshmanan. 2021. To intervene or not to intervene: Cost based intervention for combating fake news. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2300–2309. Association for Computing Machinery.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022a. Blacklivesmatter 2020: An analysis of deleted and suspended users in Twitter. In *14th ACM Web Science Conference 2022*, WebSci '22, page 290–295, New York, NY, USA. Association for Computing Machinery.

Cagri Toraman, Oguzhan Ozcelik, Furkan Şahinuç, and Umitcan Sahin. 2022b. ARC-NLP at checkthat!-2022: Contradiction for harmful tweet detection. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 722–739.

Twitter. 2022. Birdwatch is a collaborative way to add helpful context to tweets and keep people better informed. (Accessed Oct 19, 2023).

Vladimir Vapnik. 1999. *The nature of statistical learning theory*. Springer Science & Business Media.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2022. VaccineLies: A natural language resource for learning to recognize misinformation about the COVID-19 and HPV vaccines. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6967–6975, Marseille, France. European Language Resources Association.

Thomas Wolf et al. 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.

Liang Wu, Fred Morstatter, Xia Hu, and Huan Liu. 2016. Mining misinformation in social media. In *Big Data in Complex and Social Networks*, pages 135–162. CRC Press.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2).