MISINFORMATION DETECTION BY LEVERAGING USER COMMUNITIES ON SOCIAL MEDIA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Oğuzhan Özçelik
May 2024

Misinformation detection by leveraging user communities on social media
By Oğuzhan Özçelik
May 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Fazlı Can(Advisor)

Özgür Ulusoy

İsmail Sengör Altıngövde

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

MISINFORMATION DETECTION BY LEVERAGING USER COMMUNITIES ON SOCIAL MEDIA

Oğuzhan Özçelik
M.S. in Computer Engineering
Advisor: Fazlı Can
May 2024

Social media platforms have become a primary source of accessing information. However, the spread of misinformation is inevitable due to the ease of creating and sharing malicious content, including fake news. Social media users on such platforms (e.g., Twitter) often find themselves exposed to similar viewpoints and tend to avoid contrasting opinions, particularly when connected within a community. To investigate this problem, we examine the presence of user communities and leverage them as a tool to detect misinformation on social media. In this thesis, we first collect tweets together with user engagements relevant to recent events between 2020 and 2022. We then construct a human-annotated social media dataset having 5,284 English and 5,064 Turkish tweets with their veracity labels. After the data construction process, we leverage the presence of user communities for misinformation detection on social media. For this purpose, we propose a text similarity-based method that utilizes user-follower interactions within a social network to identify misinformation content. Our method first extracts important textual features of social media posts using contrastive learning. We then measure the similarity for each social media post, based on its relevance to each user in the community. Next, we train a classifier to assess the truthfulness of social media posts using these similarity scores. We evaluate our approach on three social media datasets and compare our method with the state-of-the-art approaches. The experimental results show that contrastive learning and user communities can effectively enhance the detection of misinformation on social media. Our model can identify misinformation content by achieving a consistently high weighted F1 score of over 90% across all datasets, even employing only a small number of users in communities.

Keywords: misinformation detection, social networks, social media analysis, text similarity, user communities, transformer-based language models.

ÖZET

SOSYAL MEDYADAKİ KULLANICI TOPLULUKLARINDAN YARARLANILARAK YANLIŞ BİLGİLERİN TESPİTİ

Oğuzhan Özçelik
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Fazlı Can
Mayıs 2024

Sosyal medya platformları bilgiye erişmek için en önemli kaynaklardandır. Ancak, sahte haberler dahil olmak üzere kötü amaçlı içeriklerin kolayca oluşturulup paylaşılması nedeniyle yanlış bilginin yayılması kaçınılmazdır. Bu platformlardaki (Twitter) kullanıcılar özellikle bir topluluk içindeyken benzer bakış açılarına maruz kalmakta ve karşıt görüşlerden kaçınmaktadır. Bu fikirden yola çıkarak, kullanıcı topluluklarını inceleyip onları sosyal medyada yanlış bilgi tespiti için bir araç olarak kullanıyoruz. Bu tezde, ilk olarak 2020 ve 2022 yılları arasındaki güncel olaylarla ilgili paylaşılan tweetleri ve kullanıcı etkileşimlerini toplayıp 5.284 İngilizce ve 5.064 Türkçe tweet içeren doğruluk derecelerine göre etiketlenmiş bir sosyal medya veri seti oluşturuyoruz. Ardından, bir sosyal ağ içinde kullanıcı-takipçi etkileşimlerini kullanan metin benzerliği tabanlı bir yanlış bilgi tespiti yöntemi öneriyoruz. Yöntemimiz öncelikle, karşılaştırmalı öğrenme kullanarak tweetlerin önemli metinsel özelliklerini çıkarmaktadır. Ardından, her bir gönderinin tespit edilen topluluklardaki her kullanıcıya olan yakınlık düzeyine göre benzerliğini ölçmektedir. Son olarak, bu benzerlik puanlarını kullanarak tweetlerin doğruluğunu değerlendirmek için bir sınıflandırıcı eğitiyoruz. Yaklaşımımızı üç farklı sosyal medya veri setinde en son yaklaşımlarla karşılaştırıyoruz. Deneysel sonuçlar, karşılaştırmalı öğrenmenin ve kullanıcı topluluklarından yararlanmanın sosyal medyada yanlış bilgi tespitini etkili bir şekilde artırabileceğini göstermektedir. Ek olarak, modelimiz topluluklardaki az sayıda kullanıcıyı kullanarak bile, tüm veri setlerinde %90'ın üzerinde ağırlıklı F1 skoru elde ederek yanlış bilgi içeriklerini etkili bir şekilde tespit edebilmektedir.

Anahtar sözcükler: yanlış bilgi tespiti, sosyal ağlar, sosyal medya analizi, metin benzerliği, kullanıcı toplulukları, dönüştürücü tabanlı dil modelleri.

Acknowledgement

Firstly, I would like to express my deepest gratitude to my advisor, Prof. Fazlı Can, for his deep expertise, generous help, and insightful guidance throughout my thesis journey. His support went beyond academic advice, enriching my understanding of art and literature as well. I will always use his advice and feedback throughout my life.

I would also like to thank the rest of my thesis committee: Prof. Özgür Ulusoy and Prof. İsmail Sengör Altıngövde for their time and valuable feedback.

Also, I would like to thank my current and former colleagues at ASELSAN Inc., especially Dr. Veysel Yücesoy, who is always willing to help and guide, Asst. Prof. Çağrı Toraman, Ümitcan Şahin, Fazıl Altınel, Berkan Kılıç, Ahmet Kağan Kaya, Ceyhun Emre Öztürk, and Uğur Sedef for their participation in valuable discussions in the NLP field and academic life.

My sincerest thanks also go to my girlfriend Serap for all her precious love and support; my lovely cat, Cezmi, who always encourages me to feed him and never hesitates to scratch me; and my nephew, Aren, who brought us great joy by joining us two years ago.

Lastly, I would like to thank my family. Your presence in my life has been a source of comfort and encouragement. Thank you for being there for me, no matter what. Without them, none of my accomplishments would exist.

As I embark on this new chapter, I want to share a quote from my favorite philosopher Oruç Aruoba (may he rest in peace): "Those who cannot find their own way walk all paths in vain." I am confident that my passion for NLP will lead me to a fulfilling path. With unwavering determination and a commitment to lifelong learning, I look forward to making a meaningful impact.

Contents

List of Figures

List of Tables

List of Publications

This thesis includes content from following publications:

1. **Ozcelik, O.**, Toraman, C., and Can, F., Detecting Misinformation on Social Media Using Community Insights and Contrastive Learning. Submitted to: *ACM Transactions on Intelligent Systems and Technology* (under review).

2. **Ozcelik, O.**, Toraman, C., and Can, F., SiMiD: Similarity-based Misinformation Detection via Communities on Social Media Posts. In *Proceedings of the 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-8). IEEE. 2023, November.

3. Toraman, C., **Ozcelik, O.**, Şahinuç, F., and Can, F. (2024). MiDe22: An Annotated Multi-Event Tweet Dataset for Misinformation Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)* (to appear)

Chapter 1

Introduction

The use of social media has made it easier for people to access information. However, users encounter social media posts that may contain incorrect or malicious information. When we consider the widespread use of social media, misinformation dissemination is inevitable as it is easy to create and share false information, including rumors and fake news. The spread of misinformation can affect trust in scientific and financial information [3], as well as in public health information [4]. Unverified events can also lead to poor decisions and have a lasting effect on people's reasoning [5]. For example, during the SARS outbreak in China from 2002 to 2003, fear of acquiring the illness led to stigma against Asian people [6]. Moreover, false rumors that consuming methanol is the cure for COVID-19 have caused many deaths [7]. Therefore, misinformation detection on social media is a significantly emergent problem to mitigate its spread and effects.

Misinformation detection is a complex problem by nature since malicious users play a critical role with a deliberate intent to mislead and deceive readers by enhancing the social influence of information [8, 9]. This complexity makes the misinformation detection problem challenging to develop advanced methods capable of accurately identifying patterns in content and social networks to reduce the negative effect of misinformation diffusion on social media [9]. An example of the diffusion of false information through a sample user community on social

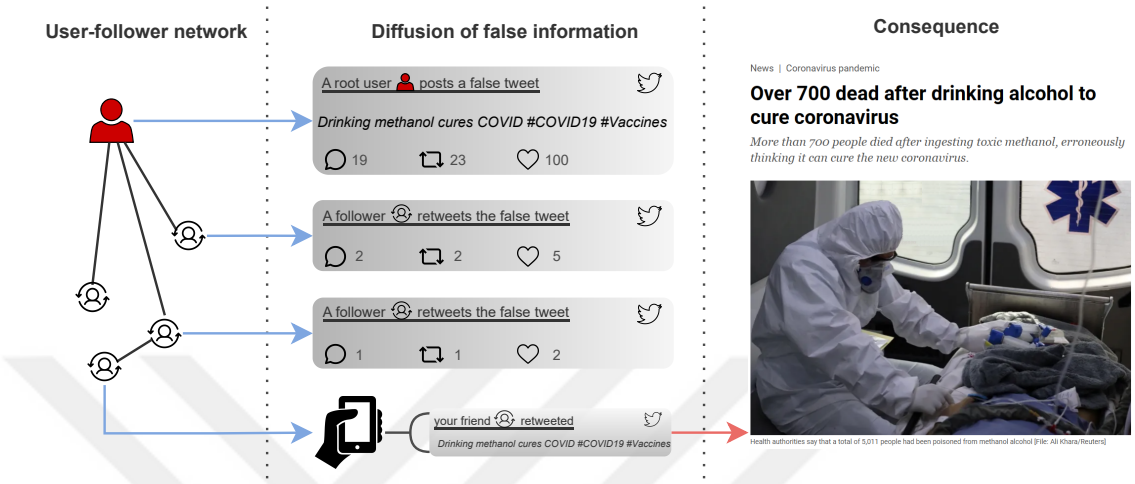


Figure 1.1: The illustration of the misinformation diffusion in a user community.

media is illustrated in Figure 1.1.

Terms such as "fake news", "rumor detection", "disinformation," and "misinformation" are often used interchangeably yet have important distinctions. While "fake news" typically refers to fabricated stories presented as news articles by manipulating information [10], "rumors" generally originate from unknown sources and spread by word-of-mouth [3]. The difference between disinformation and misinformation lies in the intention of spreader. "Disinformation" is actually a subset of "misinformation", that is false information shared to intentionally mislead readers [11]. On the other hand, misinformation is false information that is not necessarily intentionally shared to mislead or cause harm [12]. In other words, people who share misinformation on social media may not be aware that they contribute to the spread of false information. In this thesis, we specifically focus on misinformation detection. This scope acknowledges both the intentional and unintentional nature of inaccurate information while recognizing its potential for harm. By adopting this focus, we aim to investigate mechanisms for automated misinformation detection.

One of the factors contributing to misinformation spread is the existence of echo chambers [13]. The term "echo chamber" can be explained based on the group polarization theory [14], which is the group of users who reinforce each

other's beliefs and ideas resulting in a more extreme stance of community. Echo chambers can be found on various online media, including forums [15] and microblogging platforms [16, 17]. Vicario et al. [18] show that information diffusion within communities plays a crucial role in increasing polarization. Social media users may have the tendency to be exposed to or select fallacious content since they are in an echo chamber or a community supporting their opinions [18]. Another similar term is "filter bubble", where users only encounter information that confirms their existing views and beliefs due to their previous online activities [19].

One approach to interpreting user communities involves utilizing online social networks (OSNs). OSNs can be conceptualized as extensive graphs, where nodes represent users and edges depict their connections. These networks exhibit complex patterns that capture the essence of human interactions. In such networks, social network clustering algorithms [20, 2] can be used to identify cohesive user groups, often referred to as communities. These communities represent more than just clusters; they embody concentrated hubs of shared interests, beliefs, and often, shared information. While fostering specialized knowledge and community-driven dialogue, these communities also present significant challenges. They can exacerbate polarization [18] and contribute to misinformation spread [21]. Furthermore, individuals may perceive their viewpoints to be in the minority and thus choose not to express them although their views are against the misinformation. Based on the spiral of silence theory [22], individuals may perceive their views as minority opinions, leading to self-silencing and amplifying the dominant narrative, potentially even misinformation. Therefore, we anticipate that leveraging information obtained from user communities can aid in detecting misinformation.

Previous approaches to solving the misinformation detection problem can be categorized into several methodologies. Machine learning models are employed to engineer features based on the content and context of social media posts [23, 24, 25, 26]. Propagation-based approaches investigating the propagation patterns, e.g., replies and comments, are utilized with hybrid models [27, 28, 29, 30, 31, 32]. Multimodal approaches employ multiple modalities [33], e.g., images and texts, and hybrid methods utilize combined models [34], e.g., BERT [35] and GRU

[36]. Weakly supervised methods using pseudo-labeled data are proposed for fake news detection [37, 38]. Different from the previous studies, we identify user communities by user-follower social networks and assess the veracity (i.e., real or fake) of social media posts using contextual embeddings obtained from a transformer-based language model that is trained with contrastive loss [39].

In this thesis, we first curate a misinformation detection dataset containing English and Turkish tweets from recent events between 2020 and 2022. Next, we propose a misinformation detection framework, namely **Si**milarity-based **Mi**sinformation **D**etection (SiMiD), which employs textual information from user communities in social networks. We use Twitter (rebranding to X since July 2023) as a social media platform, and tweets as textual content in our experiments. Our task is binary detection of falsehoods in tweets in terms of true and false classes. Specifically, SiMiD consists of three major components:

1. We first construct a user-follower social network as a subpart of a real-world dataset to model user communities. In this network, each node is represented by the textual features of the root user's social media posts or the average of its parent nodes' social media posts. We then cluster this network to obtain multiple communities. Note that, these communities are not necessarily echo chambers, i.e., we are not interested in the detection of echo chambers [40], but rather obtaining user communities that may represent the properties of echo chambers.

2. After we obtain user communities, we remove the users in such communities from the datasets to prevent data leakage while training the text encoder. We fine-tune a Sentence-BERT (SBERT) [41] model with pairs of tweets according to veracity labels via the contrastive learning objective to obtain a text encoder. We give pseudo-labels for each tweet pair according to entailments and contradictions between their veracity labels.

3. We extract text embedding of social media posts and node embeddings in social networks using our text encoder. By using text embedding, we obtain similarity scores between input sentences and social network nodes. Finally,

we train a Support Vector Machine (SVM) model based on these similarity scores to detect the veracity of social media posts.

1.1 Research Questions

In this thesis, we investigate the following research questions:

RQ1. Can we empirically observe any relation between the detected user communities and the veracity of shared social media posts?
Motivation: We anticipate cohesive user communities share similar viewpoints and thus contribute to the spread of misinformation.

RQ2. Can we improve the performance of automated misinformation detection by leveraging user communities in a similarity-based approach?
Motivation: We expect using text similarity in user-follower networks can improve the misinformation detection performance.

RQ3. What are the contributions of contrastive learning and user communities to our misinformation detection model?
Motivation: We anticipate each component of the proposed model contributes to the full model in an ablation study.

RQ4. How does oversampling with pseudo-labeling affect the contribution of contrastive learning for misinformation detection datasets?
Motivation: Pseudo-labeling makes strong assumptions (see Section 5.3); therefore, it may not have a positive impact on the performance of the proposed method.

RQ5. How does the change in the number of individuals within the communities affect the contribution of user communities for misinformation detection?
Motivation: We do not expect that all users in the social network have the same influence on the model. Employing only a few gatekeepers/brokers may still produce promising results.

1.2 Contributions

The contributions of this thesis can be summarized in six folds: We

- Provide a detailed literature review on misinformation detection in terms of recent studies under several model families and social media datasets
- Construct a novel annotated tweet dataset for misinformation detection in English and Turkish languages
- Analyze emerging user communities on social media by using the truthfulness of the users who share true or false content.
- Propose a novel approach for misinformation detection by using retweet and follower engagements in social networks with Transformer-based language models
- Conduct extensive misinformation detection experiments on three real-world datasets with eight baselines from different model families.
- Investigate oversampling for misinformation detection datasets with a contrastive learning objective of Transformer-based language models
- Discuss the effect of the number of pseudo-labeled pairs of tweets and the number of individuals in user communities on the performance of misinformation detection. For instance, we find that even using a small number of individuals in user communities, our proposed model achieves competitive results.

1.3 Practical Implications

This thesis provides the following practical implications:

- Practical tools and other frameworks can be designed by using or extending our models to other languages in order to mitigate misinformation spread on social media.
- Regarding the effectiveness of social networks in misinformation detection, platforms and organizers can benefit from our experiments and analyses to investigate the dynamic of information flow within user communities.
- A brief survey in terms of recent misinformation detection methods and datasets.
- Opening new perspectives for the NLP community to leverage and explore contrastive learning and pseudo-labeling approaches, which can include optimizing contrastive learning and defining pseudo-labeling strategies across diverse NLP tasks.

The rest of this thesis is organized as follows, in Section 2, we review previous studies for misinformation detection, which provides a brief survey about the recent state of automated misinformation detection in terms of models. Section 3 describes the processes of the MiDe22 dataset in terms of construction, tweet crawling, and annotation. Furthermore, we review existing datasets for misinformation detection. In Section 4, we explain how to prepare social networks and identify user communities. We describe the proposed text similarity-based model in Section 5. Section 6 describes the datasets used in our experiments, and reports the experimental setup and results. We provide an ablation study with three different perspectives which are the contributions of each component to the proposed model, the effect of the number of pseudo-label pairs, and the effect of the number of individuals within communities in Section 7. We discuss the limitations of this thesis in Section 8. Finally, Section 9 concludes the paper with possible future work. We also make our implementations publicly available and provide the necessary details for the reproducibility of experimental results.[1]

[1]https://github.com/ogozcelik/simid-misinformation-detection

Chapter 2

Related Work

In this chapter, we review previous methods for automated misinformation detection. We describe prior works in Table 2.1 by indicating their important attributes. Readers may refer to survey papers for more details on misinformation detection terminology, methods, and datasets [8, 9, 42].

Since previous methods employ various features such as textual content [43], propagation networks [30], user engagement [31], and images [33] jointly; we decide to split related work into model families instead of feature types. Therefore, we review automated misinformation detection studies in the following aspects: machine learning-based (MLB), propagation-based neural network (PNN), multimodal and hybrid (MH), and weakly supervised (WS) solutions.

2.1 Machine Learning-based Solution

Machine learning-based automated misinformation detection methods can employ various numerical features such as length of words [23], profile information of social media users [44, 24, 25], and number of followers/following of users [23, 25]. Castillo et al. [23] utilize decision tree classifiers (DTC) on content and user features, and user engagement information including retweets. The event location

Table 2.1: Previous misinformation detection models. Columns descriptions are as follows. Backbone: Utilized model architecture, Text: Textual content type used as an input to the backbone, User: Utilization of user profile information, Engage: Social engagement types (Retweet/Repost (RT) and Comment/Reply (C)), Network: Utilized social network (user-user or news-user) or propagation types (comment or retweet), Code: Availability of source code.

Family	Model	Study	Backbone	Text	User	Engage	Network	Code
MLB	DTC	[23]	Decision Tree	Tweets	✓	RT	✗	✗
	SVM-RBF	[44]	SVM	Tweets	✓	RT	✗	✗
	SVM-DTST	[24]	SVM	Tweets	✓	C, RT	✗	✗
	RFC	[25]	Random Forest	Tweet	✓	RT	user-user	✗
	CIMTdetect	[26]	SVM	News	✗	RT	news-user	✓
PNN	CSI	[27]	LSTM	News	✓	C, RT	news-user	✗
	Rumor-RvNN	[45]	GRU	Tweet	✗	C	comment	✓
	CRNN	[28]	CNN, RNN	✗	✓	RT	retweet	✗
	GCAN	[30]	CNN, GNN, GRU	Tweet	✓	RT	user-user	✓
	Bi-GCN	[46]	GNN	Tweet	✗	C	comment	✓
	DUCK	[31]	BERT, GAT	Tweet	✓	C, RT	comment	✓
MH	dEFEND	[29]	GRU	News	✗	C	✗	✓
	SAFE	[33]	CNN	News	✗	✗	✗	✓
	BERT-BiGRU	[34]	BERT, BiGRU	Tweet	✗	✗	✗	✗
WS	TDSL	[37]	CNN	Tweet	✗	✗	✗	✓
	SSL	[38]	BiLSTM	News	✗	✗	✗	✗
	MetaAdapt	[43]	RoBERTa	Tweet	✗	✗	✗	✓
	SiMiD	this study	SBERT, SVM	Tweet	✗	RT	user-user	✓

of rumors and the platform where users share posts are fed into an SVM model with the RBF kernel [44]. Content, user, and diffusion-based features are also given to an SVM model considering the time intervals of the event, i.e., the feature is calculated in a time series including the slopes between each time interval [24]. Predictive features such as syntactical, user profiles, and network features are used with a random forest model for misinformation detection [25]. To detect fake news, user community information obtained from news-user networks together with the content information are fed into SVM [26]. They obtain the belonging user community of each user and use this as a categorical feature for a machine learning model. Similarly, Qureshi et al. [47] utilize news-user networks and obtain user-based and network-based categorical features to train various machine learning models for deception detection. These last two models have similar motivations when compared to our proposed model due to utilizing community information from social networks. Differently, we utilize user communities to assess the truthfulness of social media posts by obtaining text similarities and comparing the sentence embeddings. For this purpose, we also fine-tune a large

language model [41] using contrastive learning and pseudo-labeling.

2.2 Propagation-based Neural Network Methods

Propagation refers to the spread of information through a network such as social media. Propagation-based methods for misinformation detection, therefore, investigate dissemination patterns of information and how users play a role in [9]. This can be achieved by employing social media interactions, such as comments and reposts [48], into graph learning algorithms [30, 46, 31]. By representing information as interconnected nodes and edges, which are graphs, these solutions can characterize the spread of misinformation and offer new perspectives on information diffusion, enabling the identification of key nodes, influential entities, and propagation paths. Ruchansky et al. [27] employ LSTM networks to detect fake news using news articles, user comments propagation, and user credibility scores. Liu and Wu [28] propose CNN and RNN-based architecture that uses retweet propagation by employing solely user profiles as features. Lu and Li [30] propose a graph-based network (GCAN) consisting of source tweet encoding with CNN and GRU-based retweet propagation representation. Tian et al. [31] employ Transformer-based models with a graph learning mechanism for propagation trees.

2.3 Multimodal and Hybrid Methods

Multimodal and hybrid approaches include but are not limited to the combination of social media features (e.g., image, propagation, user features) and different techniques, leveraging the strengths of individual methods to enhance overall accuracy in the detection of misinformation. User comments and news articles are used in a sentence-comment co-attention sub-network (known as dEFEND) to

capture check-worthy sentences and comments for explainable fake news detection [29]. Textual and visual instances are used to construct models that learn from both feature modalities by using similarities between them [33]. Therefore, they propose A CNN framework encoding the content of fake news and the corresponding image to predict veracity. A hybrid method employing the BERT [35] model with BiGRU is proposed to solve the misinformation detection, particularly on the COVID-19 domain [34]. In this method, COVID-19-related social media posts are first fed into the BERT model and then the output is given to BiGRU to obtain the veracity of posts.

2.4 Weakly Supervised Misinformation Detection Methods

Weakly supervised methods [49] aim to construct predictive models through the utilization of weak supervision. These methods incorporate labeled and unlabeled, or pseudo-labeled data. Dong et al. [37] propose two-path semi-supervised learning via CNNs, where one path is for supervised learning and the other one is for unsupervised learning. Thus, the supervised path operates with a limited dataset, while the unsupervised path utilizes a substantial amount of unlabeled data. The confidence network layer is utilized in a semi-supervised deep learning network [38]. This layer automatically generates pseudo-labeled data and refines the self-learning process to enhance the accuracy of the neural network by providing more confident labels. Yue et al. [43] propose a meta-learning method for detecting misinformation in few-shot scenarios across different domains. They train the model on various source tasks and calculate similarity scores to the target task, i.e. misinformation content on COVID-19. Then, using these similarity scores, they rescale the meta gradients to adaptively learn from the source tasks.

Chapter 3

Data Curation

In this thesis, we construct a social media dataset for Misinformation Detection using the events between 2020 and 2022, namely, MiDe22 [50]. This dataset includes tweets from two languages, which are English (MiDe22-EN) and Turkish (MiDe22-TR).

3.1 Contributions of MiDe22 Dataset

The contributions of this dataset can be summarized in three folds:

- To the best of our knowledge, MiDe22 is the first misinformation detection dataset curated and published as an open-source repository[2] in Turkish language. Although there are studies investigating Turkish misinformation and fake news detection [51, 52], their constructed datasets are not shared and publicly available.
- Different from the benchmark misinformation detection datasets such as

[2]https://github.com/ogozcelik/MiDe22

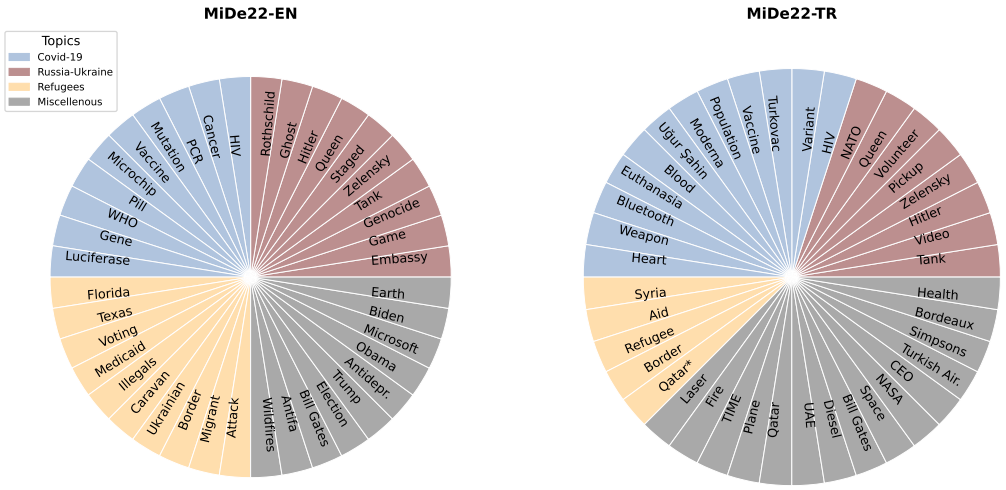


Figure 3.1: The distribution of events of each topic for the MiDe22 dataset.

Twitter15 and Twitter16 [53], MiDe22 consists of several recent events between 2020 and 2022. These topics include but are not limited to COVID-19, the Russia-Ukraine war, and Refugees.

- There are misinformation detection datasets including both news articles and tweets. However, such datasets [54, 55] are annotated only for news articles while relevant tweets for these articles remained as unlabeled. In MiDe22, there are 10,348 human-annotated tweets both in Turkish and English in total.

3.2 Description of MiDe22

The contents of the MiDe22 dataset are composed of four parts: topics, events, tweets, and user engagements.

Topics are worldwide issues having the possibility of spread of misinformation. We consider three main topics, which are "COVID-19", "Refugees", and "Russia-Ukraine war". Furthermore, we include diverse events that are not categorized under the previous topics and named such events under the "Miscellaneous" topic.

Table 3.1: Sample sentences for each veracity class from `MiDe22`.

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.

Events are specific news or rumors that are able to spread on social media. For instance, "COVID-19 vaccines do not contain HIV" is a verified event under the "COVID-19" topic. The distribution of events for each topic is described in Figure 3.1. For a detailed explanation of the events and relevant fact-checking articles, readers may refer to the open-source repository of MiDe22.[2]

Tweets are social media posts shared by users about an event. Tweets may claim false or true statements about an event. In addition, tweets may not make any statement about the truth or falsehood of an event. For this purpose, we consider three labels in order to annotate tweets, which are "True", "False", and "Other". Example tweets for each veracity class for an event are given in Table 3.1. The sample event is about the speech of T. A. Ghebreyesus, the Director-General of the 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". Considering this fact-checked information, we can identify the tweets of true and false classes. However, the tweet for the "Other" class is not related to the given event and, therefore cannot be verified by the annotators as true or false. We obtain tweets having similar keywords for each event, such as "COVID-19" and "WHO Director", as seen for the tweet in the "Other" class in Table 3.1. We anticipate that this strategy can make the dataset more challenging for developing competitive misinformation detection models.

User Engagements are social media interactions and relevant propagations for a social media post. We use all Twitter-specific user engagement types, which are retweet, reply, quote, and like. The retweet is resharing another user account's tweet with your followers by giving credit to the original author. The reply is responding to a specific tweet, i.e., starting a conversation with the author of original post. The quote is resharing another user account's tweet with your own

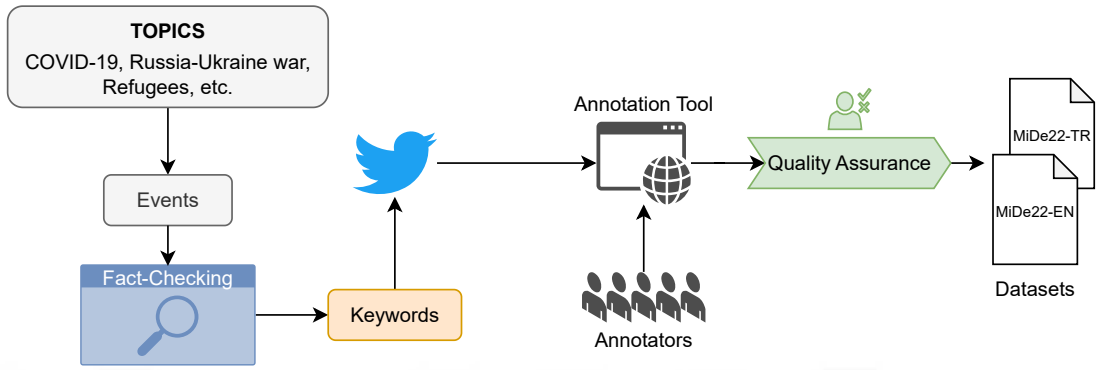


Figure 3.2: The construction process of MiDe22 dataset.

added comment by providing further context. The like expresses approval or enjoyment of a tweet without needing to write anything. We collect attributions of all user accounts that interacted with tweets using at least one of the aforementioned interaction types. The statistics of the number of user engagements are discussed in the following sections in Table 3.3.

3.3 Annotation Process

The MiDe22 is a human-annotated misinformation detection dataset. The overall process including, fact-checking, tweet crawling, implementing an annotation tool, annotating data with annotators, and quality assurance is illustrated in Figure 3.2. We first define main topics which are world-wide issues as described in Section 3.2. In addition to the main topics, to ensure the domain diversity of the dataset we keep a topic named "Miscellenous" for the events not related to the main topics. After that, we manually browsed fact-checking sites[3] to select relevant events for each topic. From these fact-checking sites, we obtain verified events and keep their verification pages with relevant content in order to further provide annotators during the tweet annotation process. An example of a fact-checked event is given in Figure 3.3. For each event we also select keywords, the beginning and end date of the event for the tweet crawling process via Twitter

[3]politifact.com, euvsdisinfo.eu, usatoday.com/news/factcheck/ for English, and teyit.org for Turkish

Figure 3.3: An example of a fact-checked event by Teyit.org platform. The event is related to the debunked claim that the COVID-19 vaccine causes myocarditis.

API's Academic Research Access.[4] An example of an event with the relevant attributes for crawling is given in Table 3.2.

After the tweet crawling process, tweets are assigned among five annotators. The annotators are undergraduate students from the computer engineering department at Bilkent University. For the annotation process, we distribute the tweets randomly to the annotators in a balanced manner according to tweet topics. We developed an annotation tool based on INCEpTION [56] for ease of labeling. Afterward, we organize a tutorial meeting with the annotators to plan the annotation process such as the annotation scheme, deadlines, and the annotation tool. They were provided with explicit definitions of true and false tweets with the corresponding examples. We pay attention to mitigate biases such as

[4]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

Table 3.2: An example event query with relevant attributes to use for tweet crawling.

Attribute	**Description**	**Sample**
Event No	The ID of event	TR13
Topic	Topic name	COVID-19
Event	Desciption of the event	Bill Gates'in aşılarla dünya nüfusunu azaltmayı amaçladığı iddiası
Link	Fact-checking site for the verification	https://teyit.org/analiz-bill-gatesin-asilarla-dunya-nufusunu-azaltmayi-amacladigi-iddiasi
Keywords	Keywords for crawling	bill gates (nüfus OR nufus)
Check Date	When the event is fact-checked	10.12.2020
Start Date	When the event begins	10.10.2020
End Date	When it is possibly end	10.02.2021
Other Keywords	Additional keywords	bill gates
Sample Tweets	A Twitter ID for a sample tweet	1355892316234510343

political inclinations and beliefs during our tutorial meeting. We distribute the tweets so that each tweet is annotated by at least two different annotators. In cases where there is a disagreement between two annotators, a different annotator is tasked with labeling the tweet. We employed a majority voting approach in such instances. If disagreement persisted among three annotators, the tweet was excluded from the dataset under the assumption that it posed a problem. The details for the interface of the annotation tool is given in Appendix B.

Given that each tweet undergoes annotation by a minimum of two annotators, we employ Krippendorf's alpha-reliability [57] to assess inter-annotator agreement (IAA) scores. The resultant alpha coefficients are 0.785 and 0.791 for MiDe22-EN and MiDe22-TR respectively. Following the study of Landis and Koch [58], our dataset demonstrates a substantial agreement among annotators. Additionally, the IAA scores achieved by MiDe22 surpass or align closely with those reported in existing datasets [59, 60, 61].

3.4 Statistics of MiDe22

The main statistics of MiDe22 are given in Table 3.3. We present the number of tweets, user engagements (likes, replies, retweets, and quotes), and the average user interactions with standard deviations across veracity labels (true, false, and

Table 3.3: The main statistics of the MiDe22 dataset for English (EN) and Turkish (TR). The mean and standard deviation among tweets for each attribute are given.

Attributes		**MiDe22-EN**			**MiDe22-TR**		
	Labels	**True**	**False**	**Other**	**True**	**False**	**Other**
	Tweets	727	1,729	2,828	669	1,732	2,663
User Engagement	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

other). The tweets are imbalanced in terms of veracities. In other words, true, false, and other labeled tweets constitute approximately 15%, 30%, and 55% of the entire dataset, respectively. We observe higher numbers for the "like" and "retweet" user interactions, this is possibly due to people being more likely to use these social media interaction types. This can be supported by the lower numbers of the reply and quote interaction types. In both languages, the mean user engagement per true tweet is higher than false tweets. Furthermore, we observe higher standard deviations for the "like" engagement. In this thesis, we use the English split of the MiDe22 dataset to compare our proposed model to other English benchmark datasets. We plan to study the Turkish split of the MiDe22 in future work.

3.5 Previous Misinformation Detection Datasets

In this section, we review existing misinformation detection datasets in the social media domain and explain the motivation behind choosing the datasets used in the experiments of this thesis.

Table 3.4: The statistics of tweet datasets which can be used for misinformation detection. The veracity labels are denoted as **T**rue, **F**alse, **N**ot related, and **U**nverified. For the "Size" column we only provide the size of the English part of the datasets if they are multi-lingual. IAA refers to inter-annotator agreement scores (e.g., Cohen's kappa) if available otherwise denoted as n/a (not available). PA, TV, and UE indicate public availability, topic variety, and the presence of user engagements (retweet, like, etc.) of the datasets, respectively, and (†) denotes if the retweet interactions are available in the UE column. If a given repository has annotated news articles (denoted by ∗ in the Size column) instead of tweets, we provide the size of annotated news articles, e.g., FakeNewsNet.

Dataset	Study	Labels	Size	IAA	PA	TV	UE	Annotated by
Twitter15	[62]	T, F, N, U	1,490	n/a	✓	✓	✓†	authors
Twitter16	[63]	T, F, N, U	818	n/a	✓	✓	✓†	authors
PHEME	[54]	T, F, U	2,402	0.62	✓	✓	✓†	journalists
FakeNewsNet	[55]	T, F	*23,196	n/a	✓	✓	✓†	journalists
CoAID	[64]	T, F	958	n/a	✓	✗	✓	query
COVIDLies	[65]	T, F, N	6,761	0.69	✓	✗	✗	domain experts
ReCOVery	[66]	T, F	*2,029	n/a	✓	✗	✗	not given
VaccineLies	[67]	T, F	14,642	0.67	✓	✗	✗	researchers
MiDe22	[50]	T, F, N	5,284	0.79	✓	✓	✓†	guided annotators

3.5.1 Review on Existing Datasets

There are efforts to curate datasets that can be used for stance detection [68], rumor classification [69], and fact verification (i.e., misinformation detection) [70]. We list annotated tweet datasets that can be used specifically for misinformation and rumor detection in Table 3.4. Twitter15 [62] is a rumor detection dataset composed of 540 rumor stories posted until March 2015. For each rumor story, the authors collected relevant tweets considering binary labels, i.e., rumor or non-rumor. Twitter16 [63] has a similar motivation to Twitter15 and includes more recent tweets than Twitter15. Twitter15 and Twitter16 have a specific version curated by Ma et al. [53]. The new version contains retweet information (propagation) and multi-class labels such as true, false, non-rumor, and unverified. Zubiaga et al. [54] curate the PHEME dataset in German and English with recent rumors from 2016. Similar to the aforementioned datasets they collect relevant tweets from the rumor threads including the retweets. FakeNewsNet [55] has news articles and relevant tweets with social engagements. Although it is a multi-dimensional dataset with annotated news articles, the relevant tweets are

not human-annotated.

After 2019, the outbreak of the COVID pandemic not only impacted the healthcare sector but also witnessed an increase in the dissemination of rumors and misinformation on social media, so-called "infodemic" [71, 72]. Consequently, researchers have initiated the production of COVID-specific datasets to delve into this phenomenon [66, 64, 65, 73, 67]. ReCOVery [66] contains annotated news articles about COVID-19 in terms of reliability and thousands of relevant tweets that are not labeled. CoAID [64] has COVID-related claims, news articles, and relevant tweets between 2019 and 2020 with ground truth labels which are decided if the corresponding tweet shares the fact-checked news articles. COVIDLies [65] is a stance detection dataset where tweets-misconceptions pairs with an annotated stance, which can be used to identify the veracity of claims. Similar to the COVIDLies dataset, Weinzierl and Harabagiu [67] curate the VaccineLies dataset by adding Twitter conversation as misconceptions in addition to readily available Wikipedia webpages. COVIDLies and VaccineLies are mainly motivated by stance detection studies. Lastly, MiDe22 [50] is a misinformation detection dataset where related tweets considering the fact-checked rumors between 2020-2022 were crawled and human-annotated.

3.5.2 Motivation Behind Choosing Datasets

Although there are publicly available social media datasets for misinformation detection tasks, many of them [64, 55, 66] remain as a repository (i.e., not curated via human-annotation), cover specific topics, e.g., COVID-19 and vaccines, [65, 67], or do not contain any user engagements [66] (see Table 3.4). In this thesis, we pay attention to choosing social media datasets with several perspectives (i.e., multi-dimensional dataset), such as publicly availability, domain/topic variety for better generalization, and presence of social media user engagements (retweet, like, etc.) to utilize our approach and benchmark analyzes with baselines. Therefore, we first choose two benchmark datasets (based on the comprehensive survey [42]): Twitter15 [62] and Twitter16 [63]. Since these two datasets

contain textual data from 2015-2016, we also include our recent dataset, MiDe22 [50], which ensures the requirements. Overall, these three real-world datasets contain both old (from 2015-2016) and timely (from 2020-2022) topics, and retweet interactions of users with social media posts. The datasets encompass tweets with human-annotated veracity labels, including but not limited to claim categories beyond "True" and "False". The statistics and details of these three datasets in terms of tweet and user counts are discussed in Section 6.

Chapter 4

Community Detection Using Real-World Datasets

In this chapter, we first explain how to construct social networks. Next, we investigate the detection and analysis of user communities. As indicated earlier our first research question investigates the relation between detected user communities and the veracity of shared social media posts.

4.1 Construction of the Social Networks

For **RQ1**, we anticipate that cohesive user communities share similar viewpoints and may contribute to misinformation spread. To investigate such communities and obtain information for further use for our text similarity-based approach, we construct social networks for each dataset. The main statistics of three social networks are given in Table 4.1. We provide the numbers of different types of individuals, which are true and false root users, and engagers. True and false users are root users who post veracity-labeled tweets in the corresponding dataset. Engagers are users who retweet and follow these root users, i.e., they do not have labeled tweets. We also report the average number of edges per node and

Table 4.1: User-follower social network statistics of each dataset. We report the Top 5 communities detected by the Louvain algorithm [2] with the ratios of each community in the graph and the number of veracity-labeled individuals of each community together with the engager count. The "+ Remaining" row shows the statistics outside the Top 5 communities in the social network. Column notations are, T: Ratio of true-labeled users in root users, F: Ratio of false-labeled users in root users, R: Root users who shared posts E: Engager who retweets and follows root users, and S: Size of the community in the corresponding social network.

Datasets	Twitter15					Twitter16					MiDe22				
Communities	T	F	R	E	S	T	F	R	E	S	T	F	R	E	S
Community A	0.80	0.20	15	236	0.21	0.11	0.89	9	152	0.21	0.88	0.12	43	451	0.50
Community B	0.00	1.00	5	116	0.10	1.00	0.00	6	105	0.14	0.19	0.81	43	373	0.42
Community C	1.00	0.00	7	111	0.10	0.83	0.17	6	85	0.12	1.00	0.00	3	40	0.04
Community D	0.25	0.75	4	94	0.08	0.50	0.50	4	79	0.11	0.00	0.00	0	19	0.02
Community E	0.50	0.50	4	90	0.08	0.33	0.67	3	71	0.09	0.00	1.00	1	2	0.01
+ Remaining	0.32	0.68	25	499	0.43	0.42	0.58	12	247	0.33	0.10	0.90	10	1	0.01
In Total	0.50	0.50	60	1146	1.00	0.50	0.50	40	739	1.00	0.50	0.50	100	886	100.00
Edge per node	82	42	62	3	-	74	51	63	5	-	20	9	14	1	-
Modularity	0.72					0.70					0.51				

modularity to show connectivity among users and user communities, respectively. For instance, users in Twitter15 have more retweeting and following interactions than users in MiDe22. Social networks with high modularity have dense connections between the nodes within user communities but sparse connections between nodes in different communities. We observe that Twitter15 and Twitter16 are more modular social networks than MiDe22.

To obtain user engagements, followers of each user across three datasets are crawled via Twitter API's Academic Research Access[4]. Following the retrieval of this user data, we randomly select root users together with their retweeters and followers, i.e., engagers. In the selection of root users from each dataset, we take into account the total number of accessed root users (provided in Table 6.1), targeting a proportion of 20%. The motivation behind choosing 20% of the root users is similar to preparing a validation set. For instance, the Twitter16 dataset has 200 accessed users in Table 6.1; therefore, we randomly choose 40 root users, where 20 have true-labeled and the other half have false-labeled tweets. The size of the social network should be large enough to obtain sufficient information and not affect the remaining data size so that we can properly train the models.

We then construct user-follower social networks using the selected root users and their engagers. In these social networks, each node represents a Twitter user (either root or engager), and each edge occurs if a user follows and retweets at least one of the tweets of the other user. We include the follower/following engagement in addition to retweets, if a user follows and retweets another user this leads to a potential spread of misinformation with a strong connection between these users.

We use Gephi [74] to visualize and infer communities of social networks. For the visualization of our social network, the Force Atlas layout is used with the default parameters in Gephi. The visualization depicts how users are connected considering the veracity of their social media posts. For instance, users who share true-labeled social media posts may follow or retweet other users who share false-labeled posts.

4.2 Detection and Emprical Analysis of the Communities

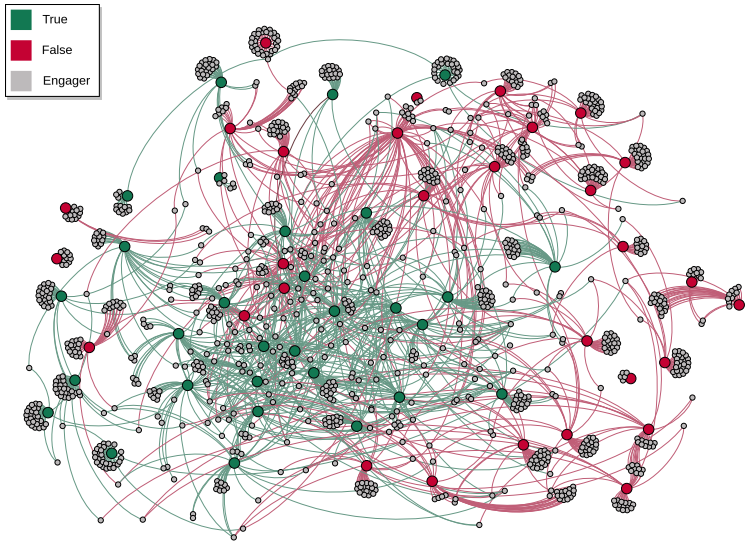
Upon constructing the user-follower networks detailed in the previous section, we proceed to implement a community detection algorithm. For this purpose, we use the Louvain algorithm [2] for community detection in social networks. This algorithm has modularity optimization which considers within-community and between-community connections and can uncover communities at different levels of granularity.

We report the top 5 detected communities with the corresponding ratio of the users and the number of veracity-labeled users in the communities in Table 4.1. Since the remaining communities are less representative with insufficient size, we empirically employ the five largest user communities.

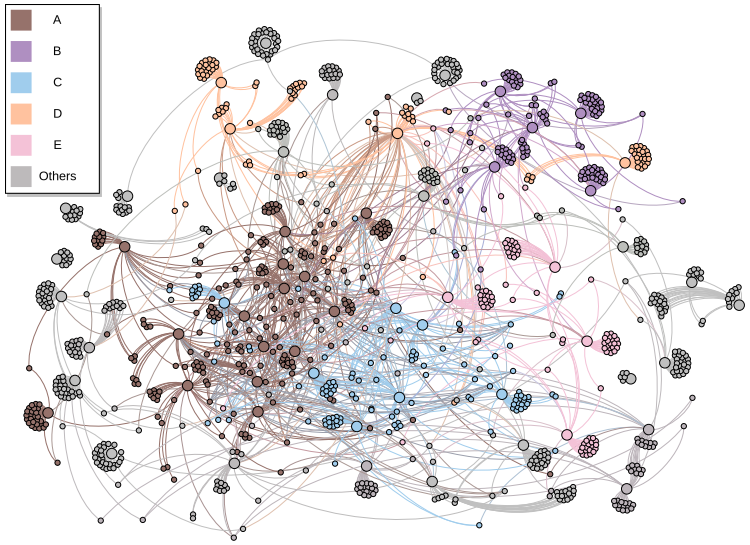
Answer to RQ1:

We observe a tendency for the top three communities in each dataset to align with specific veracity label groups. For instance, in Table 4.1, community A users (that represents the most crowded community for each network) are composed of 80%, 11%, and 88% of true-labeled root users, and 20%, 89%, and 12% of false-labeled root users for Twitter15, Twitter16, and MiDe22 social networks, respectively. From this, Twitter15 and MiDe22 form a true-labeled group of users, whereas Twitter16 forms a false-labeled group of users for community A. Considering the following communities (B and C), it is seen that users with a specific veracity label (i.e., true or false) are more dominant. For example, community B of the Twitter15 social network consists of only five false-labeled users, which covers 100% of community B. Another observation is that the MiDe22 dataset is almost composed of two main communities (which cover 0.92 of the whole social network; that is 0.50 + 0.42). We can say that the MiDe22 dataset forms two main communities, where they contain community-centric information in terms of truth or falsity. Different from the top three communities, communities D and E either comprise a small number of users or lack dominant distribution among specific veracity labels within each dataset. Considering the decreasing representativity after the top five communities, we exclude the other detected sparse communities and provide their overall statistics in the "+ Remaining" row in Table 4.1.

The aforementioned observations can directly map to visual representations of each social network. In Figures 4.1, 4.2, and 4.3, social networks for each dataset are colored according to the veracity labels of the users or the users' belonging community. Furthermore, when we examine social networks across the columns (e.g., Figures 4.1a and 4.1b or Figures 4.3a and 4.3b), we can observe the red or green-colored users residing in the same user communities as previously discussed using Table 4.1. For example, in Figures 4.3a and 4.3b, we observe that green-colored users who share true tweets are in the brown-colored community that is the crowdest in MiDe22 (i.e., Community A). Furthermore, the red-colored users who share false tweets mostly reside in the purple-colored community (Community B). We also observe more user connections (i.e., edges) in Twitter15 compared to Twitter16 and MiDe22 from Figures 4.1a, 4.2a, and 4.3a.

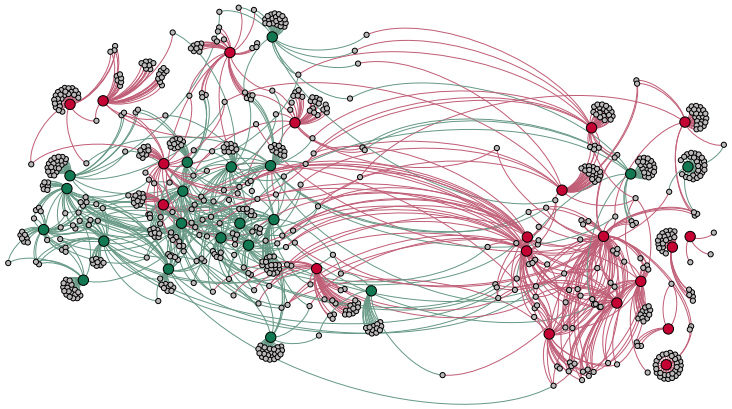


(a) Twitter15 veracity-labeled

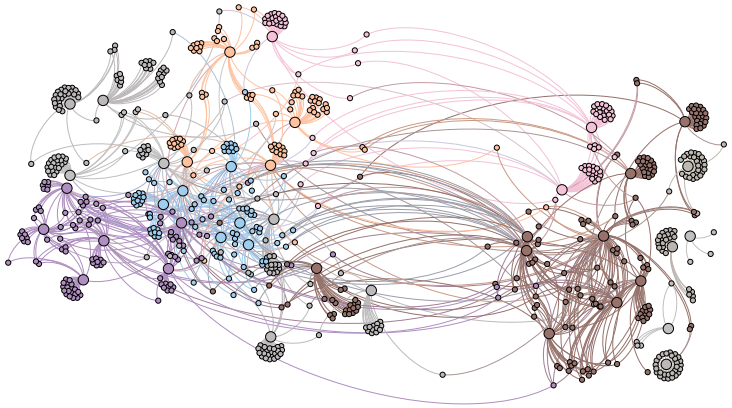


(b) Twitter15 community-labeled

Figure 4.1: The visualizations of the user-follower social networks for Twitter15. Each node represents a Twitter user, and the edges between nodes show that there are retweets and follower interactions. In (a), each node is colored if the user has a veracity-labeled tweet, otherwise gray-colored and named "Engager". In (b), each node is colored according to the community it is in, as indicated in Table 4.1). Nodes not belonging to the top five communities are colored gray. The root users are shown as double-sized engager users. This figure is taken from [1].

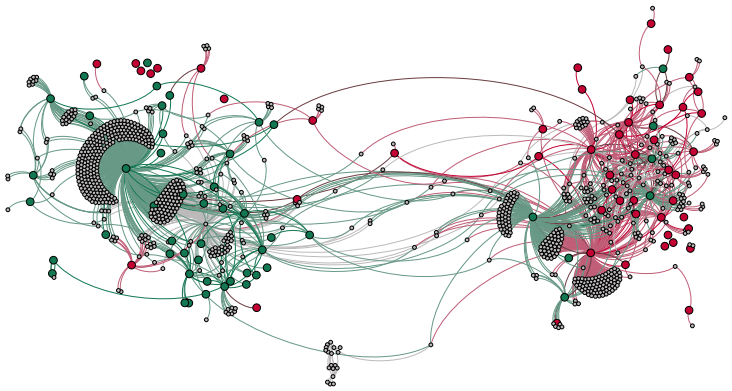


(a) Twitter16 veracity-labeled



(b) Twitter16 community-labeled

Figure 4.2: The visualizations of the user-follower social networks for Twitter16.



(a) MiDe22 veracity-labeled

(b) MiDe22 community-labeled

Figure 4.3: The visualizations of the user-follower social networks for MiDe22.

This makes it difficult for the communities in Twitter15 to separate from each other, as seen in Figure 4.1b. We conclude that there might be many broker or gatekeeper users who could control or mediate the flow of information between different communities of the network (examined in Section 7.3). Armed with these initial explorations, we employ the detected communities in a similarity-based misinformation detection model, explained in detail in the following section.

Chapter 5

Proposed Method: Similarity-based Misinformation Detection

Three stages of the proposed method are illustrated in Figure 5.1, which are (1) Community Construction, (2) Contrastive Learning via SBERT, and (3) Similarity-based Classification. The block at the top is an overview of our proposed method and the three blocks at the bottom represent the three stages. In the first stage, we construct communities by obtaining necessary user information from the datasets. For the second stage, we generate tweet pairs by giving pseudo-labels for each pair. Using these tweet pairs we fine-tune an SBERT model via contrastive learning. After these two stages, we end up with a social network having multiple user communities and a text encoder. In the last stage, we use these two modules to obtain similarities between input tweet text and all users' tweet texts in user communities. Finally, we use these similarity scores to train an SVM model for binary classification. We first introduce the notations, where the definition of each symbol is given in Table 5.1, and then explain each stage of the proposed model.

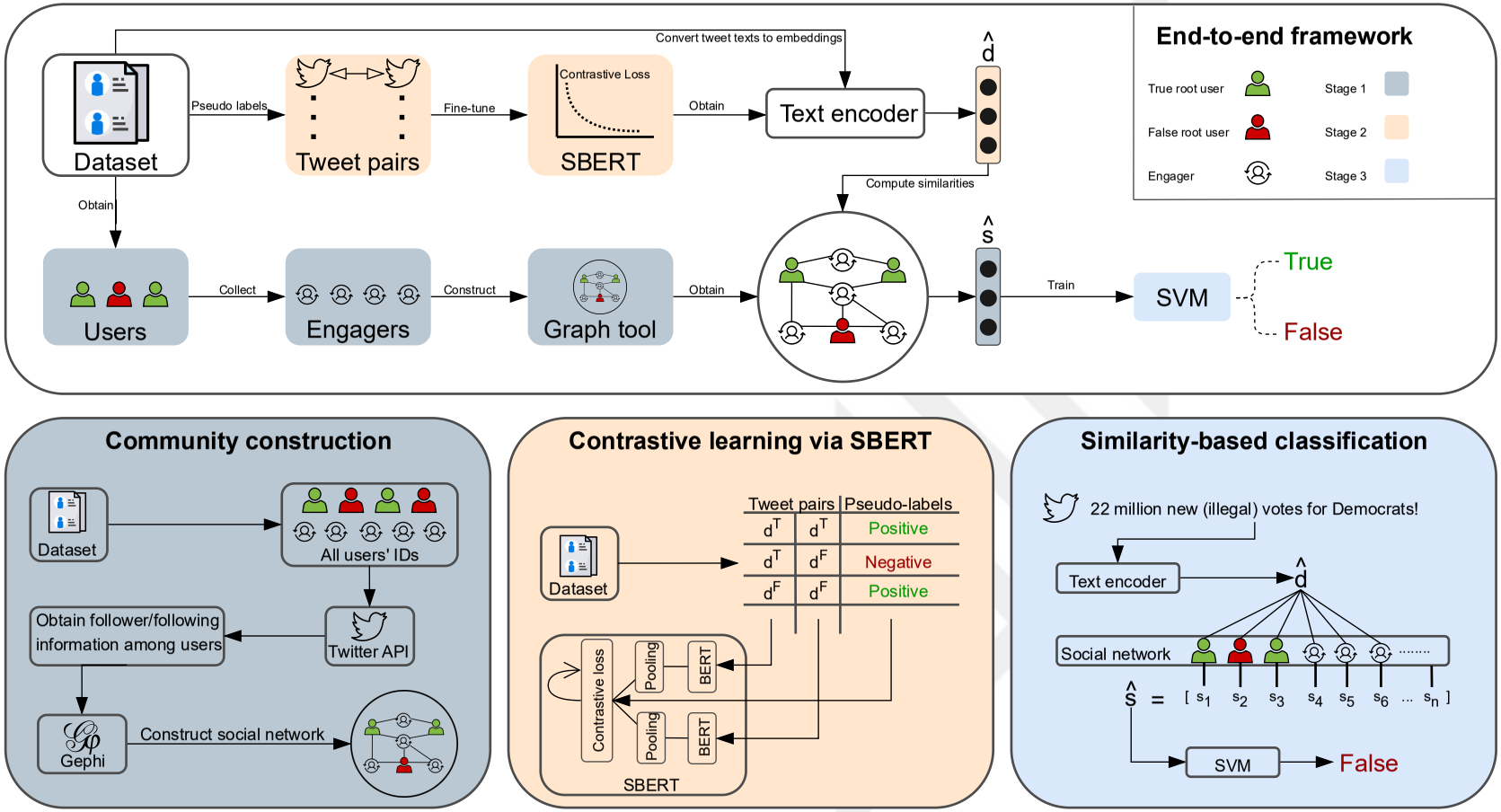


Figure 5.1: The illustration of the proposed method, SiMiD. The upper block represents an overview of the end-to-end framework. The three blocks at the bottom show three stages of the model in detail: community construction, contrastive learning via SBERT, and similarity-based classification. These stages are colored similarly at the end-to-end framework at the top.

Table 5.1: Description of the symbols

Textual content-related		User community-related	
D	Set of tweet texts	N	Constructed social network
R	Set of root users who publish tweets	C	Set of detected communities in N
E	Set of engager users	n	Nodes of N
D_r	Set of published tweets by a root user r	l	Links/edges of N
D_e	Set of engaged tweets by an engager user e	c_i	i$^{\text{th}}$ community in C
d_i	i$^{\text{th}}$ tweet text in D	$\hat{c_i}$	Vector representation of i$^{\text{th}}$ community
$\hat{d_i}$	Embedding vector of i$^{\text{th}}$ tweet in D	r_{ij}	j$^{\text{th}}$ root user in c_i
d^T	True-labeled tweet text	e_{ij}	j$^{\text{th}}$ engager user in c_i
d^F	False-labeled tweet text	$\hat{n_{r_{ij}}}$	Node embedding of root user r_{ij}
$\hat{s_i}$	Similarity vector d_i	$\hat{n_{e_{ij}}}$	Node embedding of engager user e_{ij}
$\hat{y_i}$	Predicted label for d_i		

5.1 Notations

We have a corpus $D = d_1, d_2, \ldots d_{|D|}$, where d is a tweet published by a root user r. Published tweets of a root user are denoted as D_r, where the number of shared tweets $|D_r|$ changes user by user. Tweets are labeled in terms of veracity as either true d^T or false d^F. We have a social network N composed of nodes n and edges l, where each n represents a root user r (i.e., whose tweets are labeled in the dataset) or an engager user e (i.e., whose tweets are not labeled but retweets a root user in the dataset), and l indicates a connection (i.e., edge) among users. Let an engager e_i, who follows root user r_j and retweets r_j's social media posts, there occur undirected edge l_{ij} between these users. In N, we obtain multiple communities $C = c_1, c_2, \ldots c_{|C|}$ by using social network partitioning algorithms, as explained in Section 4. Each c consists of root users r and engager users e. We construct N using users from the root users set $R = r_1, r_2, \ldots r_{|R|}$ and engager users set $E = e_1, e_2, \ldots e_{|E|}$.

5.2 Stage 1: Community Construction

To model the relationship between users in a community, we construct a social network N as explained in Section 4. In Figure 5.1, we first obtain users and their IDs to collect follower/following attributes of root r and engager e users

in each dataset D. After we obtain followers and followings of users r and e, we randomly choose 20% of the root users. After selecting root users, we add corresponding engager users to construct a social network using Gephi [74], a graph visualization and analysis tool.

In this social network, each node is represented with sentence embedding of tweets $\hat{d}$, where tweet embeddings are obtained from a text encoder under the contrastive learning stage in Figure 5.1 (detailed explanation of text encoder is in the following Section 5.3). If a node n is a root user r, its node feature $\hat{n_r}$ is the average of the text embeddings of that user's tweets (the number of tweets of a root user is denoted as D_r, where it differs from user to user) in the dataset. If a node n is an engager user e, its node feature $\hat{n_e}$ is the average of the tweet embeddings of neighbor root user nodes since there exist no labeled tweets from engager users in the dataset (the number of relevant tweets of an engager user is denoted as D_e, where it differs from engager to engager). The obtaining of node features from tweet embeddings for a single root user r and engager user e is formulated in Eq. 5.1:

$$\hat{n_r} = \frac{\sum_{1 \leq i \leq |D_r|} \hat{d}_i}{|D_r|} \qquad \hat{n_e} = \frac{\sum_{1 \leq i \leq |D_e|} \hat{d}_i}{|D_e|} \tag{5.1}$$

We filter out the tweets of root users in each N from the corresponding datasets D to ensure a fair training and testing process. In the following stages, we use these filtered datasets. Finally, we utilize the Louvain clustering algorithm [2] on the social network to obtain communities, $C = c_1, c_2, \ldots c_{|C|}$. In our proposed model, we employ user communities whose sizes are larger than 10% in Table 4.1. This maps to the top three, four, and two communities in Twitter15, Twitter16, and MiDe22, respectively.

5.3 Stage 2: Contrastive Learning via SBERT

We use SBERT [41] model that is trained on a large corpus composed of more than one billion similarity pairs[5]. We oversample (multiply the size by five) each dataset D to fine-tune this model since we observe insufficient performance with the existing number of instances in the datasets. For each true-labeled tweet d^T, we randomly choose five (five is decided as the default parameter that provides reasonable efficiency and effectiveness in our preliminary experiments) d^T and d^F. Similarly, for each false-labeled tweet d^F we randomly choose five d^T and five d^F to construct pairs. We employ pseudo-labeling for these pairs, where opposite label classes have negative pseudo-labels and the same label classes have positive pseudo-labels. For example, the binary pseudo-labels for pairs are $(d^T, d^F, 0)$, $(d^T, d^T, 1)$, $(d^F, d^F, 1)$, where 0 and 1 are negative and positive labels, respectively. An example of tweet-tweet pairs is given in Table 5.2. Next, we fine-tune the SBERT model using the contrastive learning method with the formulated contrastive loss $\mathcal{L}$ in Eq. 5.2:

$$\mathcal{L} = \frac{1}{2B} \sum_{i=1}^{B} \left[y_i \cdot \lambda_i^2 + (1 - y_i) \cdot \max(0, m - \lambda_i)^2 \right] \tag{5.2}$$

In this equation, B is the batch size, y_i is the label (either 0 or 1) of the pair i, λ_i is the distance between the text embeddings of the pair, and m is the margin hyper-parameter. The contrastive loss consists of two terms in the summation (Eq. 5.2). The motivation of the former is to penalize positive pairs (when $y_i = 1$), which reduces the distance between the positive paired sentences, i.e., $(d^F, d^F, 1)$ and $(d^T, d^T, 1)$. The latter penalizes negative pairs, i.e., $(d^T, d^F, 0)$ and $(d^F, d^T, 0)$, to encourage distances between these pairs to be larger than the specified margin m. We, then, obtain a fine-tuned text encoder with pseudo-labeled pairs, Encoder(d), where the model returns text embedding, $\hat{d}$, for a given input tweet, d.

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

5.4 Stage 3: Similarity-based Classification

After obtaining the social network and the communities (Section 5.2) and text encoder (Section 5.3), we compute similarities between input tweets and communities. For instance, we employ the top two communities c_1 and c_2 in the MiDe22 dataset as mentioned in Section 5.2. Let's have an input tweet d_i, j many root users r_{1j} and r_{2j}, and k many engagers e_{1k} and e_{2k} in user communities c_1 and c_2. We then obtain the similarity vectors of user communities between the input tweet d_i and nodes in the user communities. Finally, we concatenate these similarity vectors to obtain text similarities between input tweet and user communities $\hat{s_i}$. After calculating all similarity vectors s_i for each tweet d_i, we classify the veracity (true or false) of the tweet by utilizing an SVM model that is trained with similarity vectors as in Eq. 5.3:

$$
\begin{aligned}
d_i &= \text{“example tweet text”, where } d_i \in D \\
\hat{d_i} &= \text{Encoder}(d_i) \\
cos(\hat{a}, \hat{b}) &= \frac{\hat{a} \cdot \hat{b}}{\|\hat{a}\| \cdot \|\hat{b}\|}, \text{ cosine similarity of two vectors} \\
\hat{c_1} &= \langle cos(\hat{d_i}, \hat{n_{r_{11}}}), \ldots, cos(\hat{d_i}, \hat{n_{r_{1j}}}), cos(\hat{d_i}, \hat{n_{e_{11}}}), \\
&\quad \ldots, cos(\hat{d_i}, \hat{n_{e_{1k}}}) \rangle, \text{ where } r_{1j} \text{ and } e_{1k} \in c_1 \\
\hat{c_2} &= \langle cos(\hat{d_i}, \hat{n_{r_{21}}}), \ldots, cos(\hat{d_i}, \hat{n_{r_{2j}}}), cos(\hat{d_i}, \hat{n_{e_{21}}}), \\
&\quad \ldots, cos(\hat{d_i}, \hat{n_{e_{2k}}}) \rangle, \text{ where } r_{2j} \text{ and } e_{2k} \in c_2 \\
\hat{s_i} &= \langle \hat{c_1}, \hat{c_2} \rangle \\
\hat{y_i} &= \text{SVM}(\hat{s_i})
\end{aligned}
\tag{5.3}
$$

The pseudo-code for the overall proposed model is given in 1.

Algorithm 1 Pseudo-Code for the proposed model: SiMiD

Input: A tweet text d from the dataset D

Output: Predicted veracity y for the given tweet text d

procedure CommunityConstruction(D)
    $U \leftarrow$ ObtainUsers(D) ▷ Section 4
    $N \leftarrow$ ConstructSocialNetwork(U) ▷ Section 4
    $C \leftarrow$ DetectCommunities(N) ▷ Section 5.2
    return C
end procedure
procedure ContrastiveLearning(D)
    $tweet\_pairs \leftarrow$ ConstructPairs(D) ▷ Section 5.3
    $sbert\_model \leftarrow$ TrainSBERTModel($tweet\_pairs$) ▷ Section 5.3
    return $sbert\_model$
end procedure
procedure SimilarityBasedClassification($d, C, sbert\_model$)
    $\hat{d} \leftarrow$ Encoder($sbert\_model, d$) ▷ Section 5.3
    $similarity\_vector \leftarrow list()$
    for each $user$ **in** C **do**
        $s \leftarrow$ CalculateSimilarity($\hat{d}, user$) ▷ Section 5.4
        $similarity\_vector.insert(s)$
    end for
    $y \leftarrow$ Classifier($similarity\_vector$) ▷ Section 5.4
    return y
end procedure

function Main
    $D_{train}, D_{test} \leftarrow$ LoadDataset
    $C \leftarrow$ CommunityConstruction(D_{train})
    $sbert\_model \leftarrow$ ContrastiveLearning(D_{train})
    for each d **in** D_{test} **do**
        $y \leftarrow$ SimilarityBasedClassification($d, C, sbert\_model$)
    end for
end function

Table 5.2: Some examples of one, two, and three tweet pairs for the Twitter15 dataset. A candidate tweet is defined with a human-annotated label as "Fake". If the candidate tweet is paired with another tweet with the same label, then the pseudo-label becomes "Positive", otherwise "Negative".

Candidate Tweet = "justin bieber ringtone saves russian fisherman from potentially fatal bear attack" (Fake)		
Pair No.	**Tweet Pair <tweet$_i$, tweet$_j$>**	**Pseudo-label**
1	<Candidate, "#breaking: justin bieber thwarts bear attack!! (Fake)">	Positive
	<Candidate, "ex-marlboro man dies from smoking-related disease (Real)">	Negative
2	<Candidate, "#breaking: justin bieber thwarts bear attack!! (Fake)">	Positive
	<Candidate, "tupac was hiding for 18 years (Fake)">	Positive
	<Candidate, "ex-marlboro man dies from smoking-related disease (Real)">	Negative
	<Candidate, "paul walker's death: 'fast & furious 7' delayed but won't be scrapped (Real)">	Negative
3	<Candidate, "#breaking: justin bieber thwarts bear attack!! (Fake)">	Positive
	<Candidate, "tupac was hiding for 18 years(Fake)">	Positive
	<Candidate, "iphone 6 owners cail it's bending in their pockets (Fake)">	Positive
	<Candidate, "ex-marlboro man dies from smoking-related disease (Real)">	Negative
	<Candidate, "paul walker's death: 'fast & furious 7' delayed but won't be scrapped (Real)">	Negative
	<Candidate, "scientist finds puppy-sized spider in rainforest. sit. stay. forever. (Real)">	Negative

Chapter 6

Experimental Evaluation

In this chapter, we first revisit datasets used in the experiments and then provide the implemented baseline models from the literature. Finally, we describe the evaluation metrics and experimental setup and discuss the results.

We conduct our experiments on three datasets, namely, Twitter15, Twitter16, and MiDe22 (which are explained in detail in Section 3.5.2). We use the released versions of Twitter15 and Twitter16 from [53], and MiDe22 from [50]. These datasets consist of English tweets labeled based on their veracity. Twitter15 and Twitter16 are labeled in four classes, which are true information, false information, unverified information, and non-rumor, while MiDe22 is labeled in three classes: true information, false information, and other. For all datasets, we keep only true and false labeled tweets to ensure that our samples consist of factual social media posts. Twitter15 and Twitter16 are two benchmark datasets in misinformation detection studies [42]; however, they contain social media posts published until 2016. Therefore, we also include a recently published misinformation detection dataset, MiDe22, which is composed of tweets published between 2020 and 2022.

For all datasets, the retweet information of each tweet (i.e., who retweets the post) is publicly available. We obtain these provided user IDs for each tweet and

Table 6.1: The content and user statistics of datasets used in this study. Although all tweet texts are publicly available, some users' follower information cannot be accessed via Twitter API due to suspended and deleted accounts.

Datasets	Attribute	True	False	Total	Accessed*
Twitter15	Tweet	372	370	742	742
	Root user	214	260	443	362
	Engager	85,020	92,018	171,455	68,264
Twitter16	Tweet	207	205	412	412
	Root user	142	150	273	200
	Engager	50,646	49,322	98,000	31,832
MiDe22	Tweet	727	1,729	2,456	2,456
	Root user	626	1,563	2,175	404
	Engager	2,698	3,014	5,686	1,642

collect each user's follower information via Twitter API[4]. The main statistics of datasets are given in Table 6.1. We cannot access some users due to suspended and deleted accounts. We provide the statistics for the accessed users in the table. Twitter15 and Twitter16 have fewer instances of 742 and 412 tweets, respectively, when compared to the MiDe22 dataset. In contrast, MiDe22 contains fewer engagers, which is due to tweets in MiDe22 having less retweet interactions. On the other hand, Twitter15 and Twitter16 mostly include rumors therefore they have many retweet engagements. Moreover, Twitter15 and Twitter16 are balanced datasets in terms of class distributions. However, in MiDe22, the number of tweets labeled as false information is more than twice that of those labeled as true information, which causes data imbalance problems.

6.1 Baselines

We compare our approach with a set of baseline methods for misinformation detection. In this subsection, we first explain the rationale for choosing baseline methods. We then provide the list of baselines used in the experiments. Some baselines require additional data, such as reply threads of each social media post [24, 69, 29, 46, 31] or news content in addition to social media content [27, 26]. Since additional data are not shared publicly, we cannot include such baselines in

our experiments, e.g., SVM-DTST, CSI, Rumor-RvNN, CIMTdetect, dEFEND, and DUCK (see Table 2.1). Moreover, some studies lack model implementation, which makes reproducibility difficult for complex models, e.g., RFC [25], SSL [38]. We provide the baseline implementation details in the repository of this thesis.[1]

Overall, we consider to include baseline models from each aspect of our related work (Section 2): machine learning (DTC and SVM-RBF), propagation-based (CRNN), multimodal/hybrid (SAFE and BERT-BiGRU), and weakly supervised methods (MetaAdapt). Furthermore, we include state-of-the-art deep learning models, i.e., BiLSTM [75], BERT [35], and BERTweet [76], which produce very challenging scores in text classification problems [77]. Thus, we compare our proposed model with the following baseline models:

- **DTC** [23] is a decision tree-based model combining several social media features. Since the model employs various types of features, e.g., content, user, topic, and engagements, it provides a comprehensive perspective to mitigate the misinformation detection problem. We use available features from the datasets, which are content, user, and engagement features (i.e., text length, count of followers, retweet counts, number of hashtags, question marks, and URLs).

- **SVM-RBF** [44] is an SVM model with RBF kernel, finding the optimal hyperplane that separates different classes in a feature space, maximizing the margin between them. They use various types of social media features similar to [23], such as content, user, and engagements. Furthermore, they include the location of rumors and client programs used to post, such as web or mobile.

- **CRNN** [28] is a propagation-based fake news detection model. The architecture is constructed in a hybrid manner employing both CNN and GRU networks. For each source tweet, the propagation of retweets is given as a sequence to the CNN and GRU models, providing each user's social media profile information, such as number of followers, length of usernames, account verification, account age, and number of shared tweets.

- **SAFE \V** [33] is a similarity-based multi-modal fake news detection method that utilizes similarities between visual and textual features, named SAFE. In this study, we remove the visual feature part, as described in [33] and use only textual features (i.e., SAFE \V) since visual features are not in the scope of this study. The textual part of the model employs the Text-CNN [78] model with an additional fully connected layer.

- **BERT-BiGRU** [34] is a hybrid approach combining BERT [35] and Bidirectional GRU (BiGRU) [36] models to detect veracity of social media posts. In this approach, first, the BERT model is fine-tuned using the textual content of tweets. Next, token embeddings obtained from BERT are fed into the BiGRU model. Finally, a dense layer and sigmoid function are used to predict the veracity of social media posts.

- **MetaAdapt** [43] is a meta-learning-based approach for domain-adaptive few-shot misinformation detection. By training the initial model with multiple source tasks and adaptively learning from them based on similarity scores, MetaAdapt leverages limited target examples to effectively transfer knowledge from the source to the target domain. We use the PHEME [54] dataset and provide our validation splits as target examples, similar to the original implementation.

- **BiLSTM** [75] is a bidirectional recurrent neural network designed to capture contextual information from both past and future inputs of the sequence. Unlike traditional RNN architectures, BiLSTMs process input sequences in both forward and backward directions, allowing them to better understand and represent temporal dependencies in sequential data. We use one-hot encoding, i.e., converting each word in a given text into a unique numeric vector consisting of binary values, for input features during the training of BiLSTM. Tweet contents are given as an input sequence.

- **BERT** [35] is a bidirectional language model using self-attention [79] and pre-trained with the masked language modeling and next-sentence prediction tasks. Transformer-based language models capture the content of a sentence and the location of each word in the sentence, which provides

contextual information and long-range dependencies, based on the Transformer architecture [79]. We fine-tune this model on a downstream task, i.e., misinformation detection, with tweet texts and classify the truthfulness of tweets.

- **BERTweet** [76] has a similar architecture with BERT model. Differently, its pretraining corpus consists of tweet collections, and it uses a specialized tokenizer for text preprocessing. This improves its performance over BERT when the task includes noisy social media content [76, 80].

6.2 Evaluation Metrics

The evaluation metrics are accuracy, F1 scores on true and false classes, and weighted F1 scores over the distribution of classes. Accuracy is the ratio of the total number of correct predictions, i.e., True Positives (TP) and True Negatives (TN), and the total number of predictions, i.e., the sum of correct predictions with False Positives (FP) and False Negatives (FN). It is formulated in Eq. 6.1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

F1 score is the harmonic mean of precision $TP/(TP + FP)$ and recall $TP/(TP + FN)$ metrics, which can be calculated as in Eq. 6.2. We use per-class F1 scores for true and false classes and weighted F1 scores in terms of class distributions. Per-class F1 scores are calculated considering the candidate class as a positive class. For the true F1 score, the positive class is true-labeled tweets while the positive class is false-labeled tweets for the false F1 score. Furthermore, we use the weighted F1 score instead of the macro F1 score since it takes class imbalance into account. This is important, especially for the imbalanced MiDe22 dataset.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6.2}$$

6.3 Experimental Setup

We randomly choose approximately 20% of the unique root users[6] as described in detail in Sections 4.1 and 5.2. We then filter out the tweets of those root users from the datasets to ensure a fair training and testing process. We split the filtered datasets in a 5-fold stratified manner (in terms of the veracity classes) in 70% training, 10% validation, and 20% test splits, where community users are not included in any of the splits (see Section 4.1). We obtain the best models for the baselines and the proposed model according to the validation split and report the average scores of five test splits with standard deviations in Table 6.2. We also provide the statistical test results in the table, using a paired t-test at a 95% confidence interval in pairwise comparisons.

We use the Transformers library [81] to fine-tune BERT[7], BERTweet[8], and SBERT[5] models. We set the number of epochs as 10, batch size 16, dropout 0.1, sequence length 128, and learning rate 5e-5 for the fine-tuning of Transformers-based models. Differently, we train the BiLSTM model during 50 epochs using the TensorFlow library [82]. We use default parameters from scikit-learn[9] for the training of SVM model in the proposed method. We use default parameters specified in the corresponding studies for the implementation of baseline models. To obtain the best models, we evaluate the performance of models by calculating the weighted F1 scores on validation split after each 100 iterations. Regarding computing resources, the training is performed on a single NVIDIA RTX A4000. The computer used during the training process has 128 GB of memory. For the repeatability of our scores, other researchers may refer to the aforementioned hyper-parameters and our codebase.[1]

[6]We can select users from the pool of accessed users listed in Table 6.1. Due to account suspensions or deletions, some user accounts may be missing after crawling their information.

[7]https://huggingface.co/bert-base-uncased

[8]https://huggingface.co/vinai/bertweet-base

[9]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Table 6.2: Performance of baselines and the proposed method on test splits of three datasets in terms of four evaluation metrics. The best scores for each dataset and metric are given in bold. The symbol "*" indicates a statistically significant difference between the highest-performing method and others. For instance, in the MiDe22 dataset, the differences between the highest-performing method, SiMiD, and other methods, except BERTweet, are statistically significant in accuracy.

Datasets	Methods	Accuracy	True F1	False F1	Weighted F1
Twitter15	DTC	0.581±0.017*	0.534±0.040*	0.618±0.021*	0.577±0.019*
	SVM-RBF	0.649±0.042*	0.513±0.059*	0.724±0.039*	0.623±0.043*
	CRNN	0.520±0.014*	0.469±0.085*	0.548±0.069*	0.510±0.020*
	SAFE \V	0.897±0.033	0.891±0.037	0.903±0.029	0.897±0.033
	BERT-BiGRU	0.814±0.043*	0.792±0.063*	0.830±0.030*	0.812±0.046*
	MetaAdapt	0.645±0.041*	0.651±0.050*	0.630±0.071*	0.639±0.042*
	BiLSTM	0.880±0.058	0.858±0.083	0.895±0.044	0.877±0.062
	BERT	0.918±0.027	0.914±0.030	0.921±0.025	0.918±0.027
	BERTweet	0.912±0.025	0.911±0.025	0.912±0.027	0.912±0.025
	SiMiD (ours)	**0.931±0.020***	**0.928±0.019***	**0.933±0.020***	**0.931±0.020***
Twitter16	DTC	0.611±0.065*	0.579±0.059*	0.637±0.071*	0.609±0.063*
	SVM-RBF	0.635±0.035*	0.684±0.044*	0.560±0.053*	0.621±0.035*
	CRNN	0.564±0.105*	0.588±0.124*	0.487±0.237*	0.537±0.140*
	SAFE \V	0.890±0.026*	0.886±0.028*	0.894±0.025*	0.890±0.027*
	BERT-BiGRU	0.864±0.022*	0.859±0.023*	0.869±0.022*	0.864±0.022*
	MetaAdapt	0.705±0.061*	0.720±0.035*	0.671±0.116*	0.695±0.066*
	BiLSTM	0.726±0.048*	0.608±0.093*	0.789±0.030*	0.700±0.060*
	BERT	0.924±0.036	0.923±0.035	0.926±0.038	0.924±0.036
	BERTweet	0.922±0.023	0.918±0.024	0.925±0.022	0.921±0.023
	SiMiD (ours)	**0.937±0.031***	**0.935±0.032***	**0.940±0.030***	**0.937±0.031***
MiDe22	DTC	0.631±0.023*	0.268±0.020*	0.753±0.020*	0.618±0.017*
	SVM-RBF	0.756±0.012*	0.231±0.071*	0.855±0.006*	0.681±0.024*
	CRNN	0.685±0.020*	0.158±0.103*	0.804±0.022*	0.624±0.016*
	SAFE \V	0.850±0.015*	0.713±0.032*	0.899±0.010*	0.847±0.016*
	BERT-BiGRU	0.866±0.014*	0.758±0.019*	0.908±0.011*	0.866±0.013*
	MetaAdapt	0.642±0.060*	0.451±0.093*	0.718±0.087*	0.644±0.041*
	BiLSTM	0.842±0.011*	0.696±0.025*	0.893±0.008*	0.838±0.012*
	BERT	0.884±0.022*	0.793±0.032*	0.919±0.017*	0.884±0.021*
	BERTweet	0.886±0.013	0.779±0.030*	0.923±0.009	0.883±0.014*
	SiMiD (ours)	**0.908±0.022***	**0.832±0.040***	**0.937±0.015***	**0.908±0.022***

6.4 Experimental Results

We present the performance of our method and baselines for each dataset in Table 6.2.

Interpretations on model performances: Machine learning baseline models, i.e., DTC and SVM-RBF, have poor performance compared to other models. The feature representations for these models may not be sufficient to grasp the

complex structure of noisy tweet texts with different veracity labels and need further improvements. On the other hand, recurrence and Transformer-based text classification baselines, i.e., BiLSTM, BERT, and BERTweet achieve very challenging results in misinformation detection. The SiMiD model does not produce statistically significant results over these models in the Twitter15 and Twitter16 datasets. Yet, such models may not be fully adaptable for small-sized datasets since they have millions of parameters to tune; therefore, we can observe higher standard deviations. For example, BiLSTM and BERT result in ± 0.06 and ± 0.03 standard deviations on average for weighted F1 scores in Twitter15 and Twitter16, respectively. Propagation-based neural methods, e.g., CRNN, perform well in misinformation detection problems as described in Section 2. However, they cannot benefit if a tweet does not have many interactions. For instance, we have lower interactions for tweets in MiDe22 (see "Edge per node" row in Table 4.1). Following this statistic, the CRNN model cannot detect most of the true tweets in MiDe22 and produces a 0.158 F1 score. Multimodal methods such as SAFE employ many modalities, such as images and texts, therefore, they need tweets to include both textual and visual content. This can decrease the performance of misinformation detection for only textual content as expected. Yet the SAFE \V model produces challenging scores using the Text-CNN model similar to the family of BiLSTM. Furthermore, hybrid methods, e.g., BERT-BiGRU, incorporate state-of-the-art deep learning models such as BERT and BiGRU. Surprisingly, it produces lower scores compared to the vanilla BERT model. We anticipate that instead of using sentence embeddings, token embeddings obtained from BERT may not be adequate for the misinformation detection problem. Additional layers on the top of the BERT model employing token embeddings do not always produce challenging results [83]. Finally, we investigate a weakly supervised few-shot method, MetaAdapt, that uses tweet contents of a source dataset for the training of a Transformer-based model and tuning the parameter using few-shot samples from a target dataset. This model produces lower scores compared to supervised methods but yet promising results when compared to machine learning-based approaches. Since the source datasets are from different domains, one can further improve the performance of such methods by combining source datasets from various domains.

Interpretations on datasets: We conduct our experiments on benchmark misinformation detection datasets. Twitter15 and Twitter16 have relatively small datasets compared to MiDe22 (see Table 6.1). Such small datasets may make it challenging to fine-tune a complex text encoder due to overfitting [84]. We observe that our proposed model, including the fine-tuning process with oversampling and pseudo-labeling, results in the highest accuracy and F1 scores on the Twitter15 and Twitter16 datasets. Specifically, we improve the performance of automated misinformation detection compared to vanilla fine-tuning of the BERT and BERTweet models. Although SiMiD outperforms all baselines, it does not produce statistically significant scores for SAFE \V, BiLSTM, BERT, and BERTweet in the Twitter15 and Twitter16 datasets. We observe higher standard deviations for these two datasets. For example, BiLSTM in Twitter15 produces ± 0.062 for a 0.877 weighted F1 score, which indicates that weighted F1 scores over five test splits range between 0.815 and 0.939. We anticipate that complex models such as BiLSTM and BERT trained by relatively small Twitter15 and Twitter16 datasets are more prone to overfitting. On the contrary, BiLSTM in MiDe22 produces ± 0.012 standard deviations for a 0.838 weighted F1 score.

Interpretations on evaluation metrics: In Table 6.2, we provide accuracy, per-class F1 scores (i.e., true F1 and false F1), and weighted F1 score over the class distributions. We observe that machine learning models (i.e., DTC and SVM-RBF) have poor performance in the detection of true tweets. For instance, DTC and SVM-RBF have 0.268 and 0.231 true F1 scores in MiDe22, respectively. This outcome may be related to the generic hand-crafted feature representations, such as tweet lengths, number of followers, account age, etc. We anticipate that the performance of such models for the discrimination of true-labeled tweets from false-labeled ones can be improved with additional keyword-based features such as the inclusion of "fact", "fact-checking", and "rumor". For the detection of false tweets, strong baselines such as SAFE \V, BERT-BiGRU, BiLSTM, BERT, and BERTweet produce challenging scores compared to the SiMiD model. On the other hand, CRNN and MetaAdapt have room for improvement in false tweet detection since they produce 0.487 and 0.671 false F1 scores in Twitter16, respectively. For the accuracy and weighted F1 metrics, SiMiD outperforms most of

the baselines except SAFE \V, BiLSTM, BERT, and BERTweet models. Although these models achieve lower scores than our proposed model, SiMiD does not produce statistically significant scores, especially in the Twitter15 dataset.

Following our second research question, we explore enhancing automated misinformation detection by employing a similarity-based approach through contrastive learning and leveraging user communities. To do so, we compare our proposed method with strong misinformation detection baselines from each model family, such as machine learning, propagation-based neural networks, multimodal/hybrid methods, and weakly supervised approaches.

Answer to RQ2:

Overall, we show that leveraging user communities with a text similarity-based approach (SiMiD) can benefit automated misinformation detection on social media. SiMiD outperforms all baselines for all datasets in all metrics. Furthermore, it produces statistically competitive scores when compared to DTC, SVM-RBF, CRNN, and MetaAdapt models. The statistically lower performance of SiMiD in Twitter15 and Twitter16 is possibly related to the dataset's size. Due to the small size of samples in these datasets, constructing a well-represented community becomes a challenging task. Contrastively, we observe statistically significant results over all models in MiDe22 for accuracy and false F1 score. Considering the outcomes from Section 4.2, we anticipate that complex structures of Twitter15 and Twitter16 social networks (see Figures 4.1a and 4.2a) make it difficult to detect the veracity of these tweets. There may be some extra knowledge, e.g., user profile and timeline information, needed to better grasp the structure of the dataset. These promising results prove the effectiveness of SiMiD which leverages user communities with a text similarity-based approach for misinformation detection. To summarize, the SiMiD model produces robust scores considering the standard deviations ($\pm 2\%$) on the average of multiple experiments.

Chapter 7

Ablation Study

In this chapter, we explore our proposed model under several aspects following the research questions **RQ3**, **RQ4**, and **RQ5**. Specifically, we first investigate the contributions of each component to our proposed model, SiMiD. Next, we experiment with the number of pseudo-label pairs for contrastive learning and the number of individuals within communities.

7.1 Contributions of Contrastive Learning and Social Network to the Proposed Model

We examine the impact of the two major components in our model, namely contrastive learning (CL) and social network (SN), by measuring the performance without using them in Figure 7.1. To assess the effectiveness of the CL component, we use the SBERT model (mentioned in Section 5.3) without employing contrastive learning via fine-tuning ("-CL" in the figure). Likewise, to quantify the impact of the SN component, we remove the similarity-based classification stage that computes the similarity between tweets and users in the communities ("-SN" in the figure), and directly give the output of the SBERT model to the final classification layer in the framework (see Figure 5.1). Exploring our third

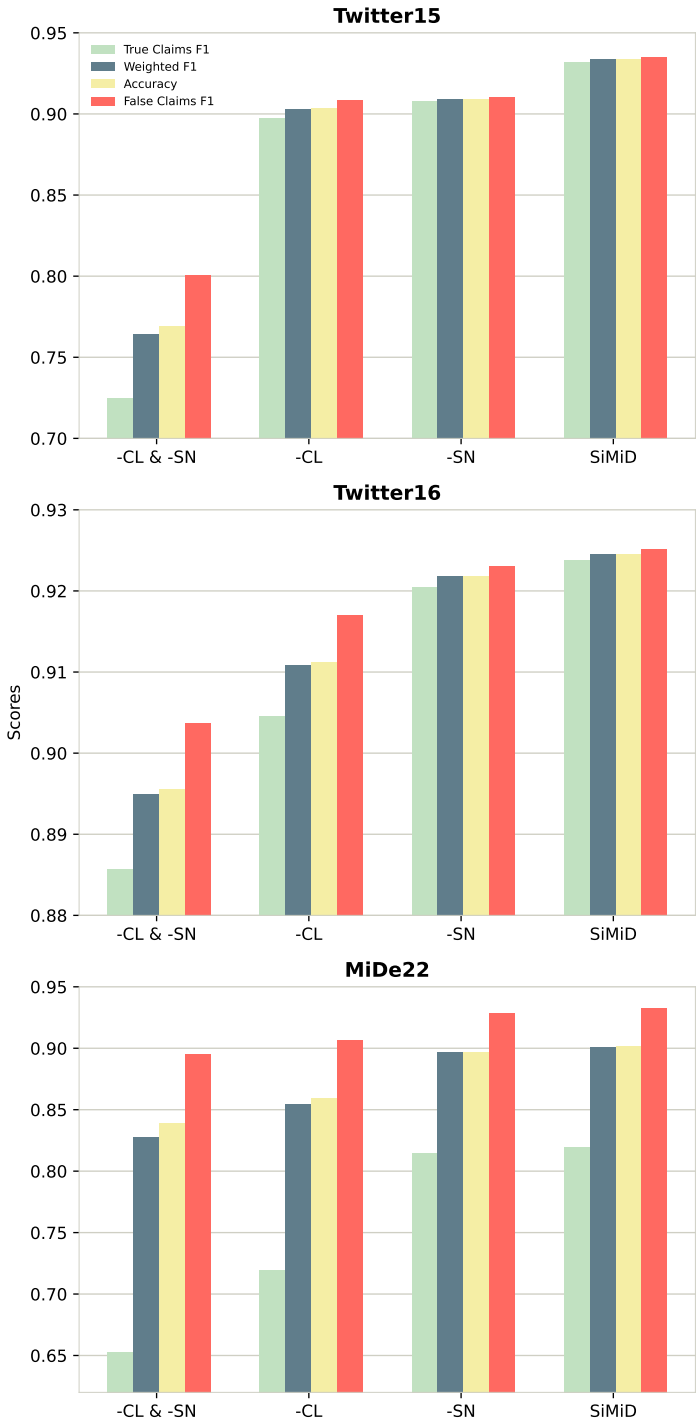


Figure 7.1: The model performance without contrastive learning (CL) and social networks (SN) for three datasets in terms of accuracy and F1. SiMiD represents all model components.

research question, we analyze the contributions of contrastive learning and user communities in misinformation detection. Different from the previous system (see Section 5), we include all users in the corresponding social networks. This is because we may overlook some important users in small communities, i.e., having less than 10% size. For example, we particularly investigate the effect of the number of individuals in the social networks in Section 7.3.

Answer to RQ3:

We observe CL and SN components contribute to the SiMiD model in all metrics for all datasets. Specifically, the use of CL and SN improves the misinformation detection performance in Twitter15 and Twitter16, observing the transition from "-CL&-SN" to "-CL" and "-SN" in Figure 7.1. Although CL and SN have also made a promising contribution to MiDe22, their performance can be improved by tuning the learning process, such as pairing inputs to generate train data. This lower contribution can be related to the size of MiDe22, whose size is at least four times greater than Twitter15 and Twitter16. Moreover, for the F1 scores of true classes, we observe a dramatic decrease when the CL and SN components are removed. Based on these observations we deduce that contrastive learning and social network components are essential to the detection of true labeled instances. We also see more robust scores (i.e., lower differences among each metric) with the full SiMiD model in Twitter15 and Twitter16. Considering the fewer instances of these small-sized datasets over-sampling seems to produce better learning of the specific misinformation detection task.

7.2 The Effect of the Number of Pseudo-label Pairs for Contrastive Learning

We investigate the effect of the number of pseudo-label pairs used in contrastive learning via SBERT (which is explained in detail in Section 5.3). SBERT can be fine-tuned using pairs or triplets in the following tasks: semantic textual similarity and natural language inference (NLI) [41]. Since we have categorical labels

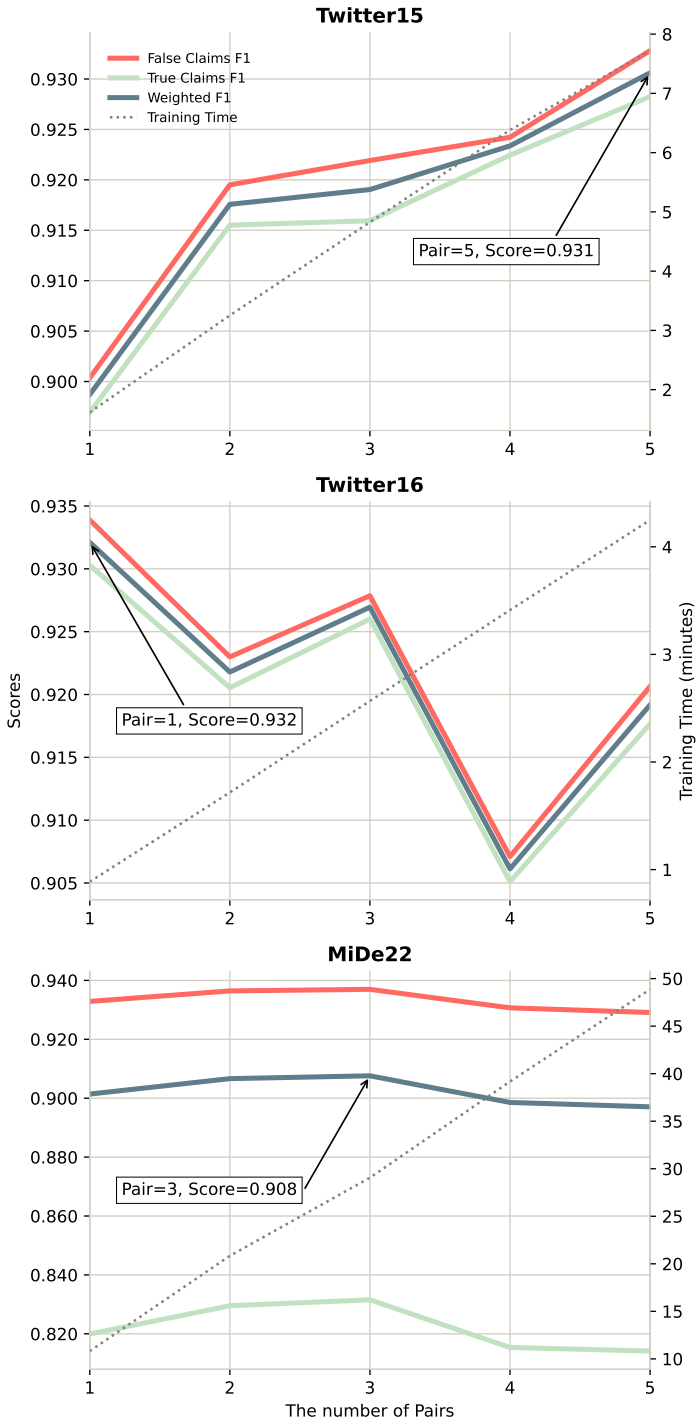


Figure 7.2: The model performance based on the number of pseudo-label for the fine-tuning of the SBERT model. We indicate the number of pseudo-label pairs on the x-axis, F1 scores on the left y-axis, and training times on the right y-axis. Rectangles highlight the best F1 scores and their corresponding pair numbers.

(i.e., true and false), we conduct natural language inference with entailments and contradictions between tweet pairs for pseudo-labeling, similar to the approaches in [85, 86]. In these approaches, the authors use NLI for the prediction of harmful tweet texts [85] and fact-checking. Differently, we aim to obtain a text encoder for further use in a similarity-based approach. In other words, we fine-tune the SBERT model with a set of tweet pairs and use these fine-tuned SBERT models as a text encoder in the SiMiD model (see Figure 5.1). Following the described process in Section 5.3, we construct tweet pairs, for each tweet in the dataset, by randomly selecting 1, 2, 3, 4, or 5 true and false labeled tweets from the same dataset. We repeat this process for each number of selected tweets, creating separate datasets for 1 to 5 pairs. For each pair, we give pseudo labels according to the opposition in the pair (i.e., positive label if both have the same veracity label). This method aims to diversify the dataset and make it suitable for contrastive learning. To do so, we can explore the effectiveness of the number of pairs on the performance of SiMiD. Addressing the fourth research question, we examine the impact of this oversampling strategy with pseudo-labeling on the efficacy of contrastive learning within misinformation detection datasets.

In Figure 7.2, we can see that with the increasing number of pseudo-label pairs the performance of our proposed model increases for the Twitter15 dataset. In other words, when we use just one pair for each sample, the score is the lowest compared to the greater number of pairs. We achieved the best score when using five different pairs for each sample in the Twitter15 dataset. MiDe22 dataset achieves the best score with three pair numbers, yet the performance decreases for four and five pairs. Surprisingly, the Twitter16 dataset shows a directly opposite trend line. It performs the best when using just one pair for each sample, and the worst for using four pairs. We argue that this may occur due to the randomness of the pairs. If the data is curated from social media posts from very distinct rumors, the model may not be able to grasp the semantic meaning of pseudo-labeled pairs. Another observation is that small-sized datasets (Twitter15 and Twitter16 compared to MiDe22) produce a higher variance for the change of pair number compared to MiDe22. We can see the difference between the best and the worst performance scores for Twitter15, Twitter16, and MiDe22 are ± 0.03,

±0.03, and ±0.01, respectively.

Answer to RQ4:

Overall, the choice of the number of pairs is not straightforward. That is, the performance may deteriorate by increasing the number of pairs (as in Twitter16); moreover, the drawback occurs due to high training times. We still arguably say that pseudo-labeled pair choice may be helpful for small-sized datasets, e.g., Twitter15. There may be no need for further fine-tuning or grid-search of the SBERT model for medium-sized datasets, such as MiDe22.

7.3 The Effect of the Number of Individuals within Communities

We analyze to assess the impact of the varying numbers of individuals within communities on performance. In our user community-based approach we employ all users in a community. However, if the are too many individuals in a community, the extracted information from the community can get complicated and the model may lose scalability. Therefore, we anticipate that using fewer users who are capable of carrying important information in a sense may still produce promising results. To investigate individuals, we find the gatekeepers in the social network based on their betweenness centrality [87], with a focus on individuals possessing higher betweenness centralities. The motivation behind using betweenness centrality is to find gatekeepers and bridges in the network who play a key role in the user-follower network [20]. A gatekeeper, in the context of betweenness centrality, can be seen as a person who controls access to information or resources by their strategic position in the network. They may not necessarily be the most connected or central individuals in terms of the number of direct connections, but they have a crucial role in mediating the interactions between different subgroups or communities. We assume that such influencer users can hold representative information about the communities. We analyze the model performance even when a small number of users are considered as representatives.

We detect users having the highest betweenness centralities using the metric in Equation 7.1. The shortest paths that pass through a particular user v (a single node in a social network) can identify users having the potential to influence the flow of information within the network. Betweenness centrality $C_b(v)$ for a single user v in network N is then calculated as follows:

$$C_b(v) = \sum_{s \neq v \neq t \in N} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7.1}$$

where σ_{st} is the number of shortest-path between two different users s and t, and $\sigma_{st}(v)$ is the number of shortest-path from s to t that v lies on.

In line with our final research question, we investigate how changes in the number of individuals within communities affect the overall contribution of user communities to the detection of misinformation. We rank 50 users based on their betweenness centralities scores, then employ a selection approach, where the ranked list is divided into groups of increasing size (5, 10, 15, etc.) as shown in the x-axis of Figure 7.3.

We find that even just using the top five users in each social network, we produce challenging results in all datasets. We observe a gradual increase in Twitter15 and Twitter16 with the increasing number of individuals. Surprisingly, employing only 20 users in Twitter16, we achieve the highest score. This outcome shows that the choice of users in terms of the betweenness centrality score instead of using all users in a community can affect the model performance drastically. On the other hand, we observe a slight increase in the model performance in the MiDe22 dataset. We anticipate that this slight increase is related to the average connection of users (see Table 4.1). The number of average connections of users in MiDe22 is lower than in Twitter15 and Twitter16, which can decrease the user representations in social networks. Due to insufficient connections, we deduce that changes in the number of individuals do not have a strong impact on the performance of the MiDe22 dataset. In class-based analysis, we observe the same patterns in the change of true and false claims with the increasing number of individuals. We can conclude that the varying numbers of individuals have a

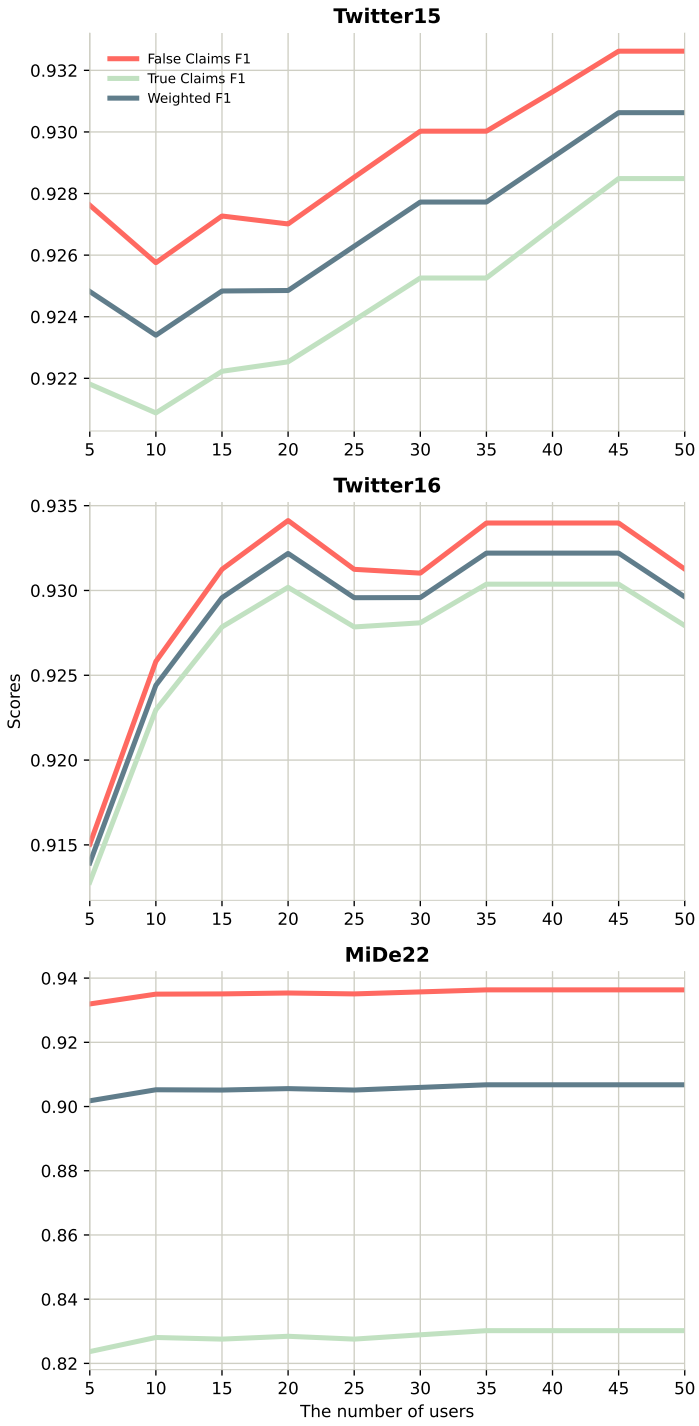


Figure 7.3: The model performance on the varying number of individuals within communities. Each metric is indicated with different lines of color.

Table 7.1: The main characteristics of the top five users in terms of betweenness centrality scores in each social network. We indicate root and engager user types with and , respectively. Users who share/retweet true tweets are colored green, otherwise red under the truthfulness column. Account usage shows the main use purpose of a user on social media. The follower and following columns are colored in terms of values (darker is higher).

Network	User Type	Truthfulness	Account Usage	Follower	Following
Twitter15		Shared false tweets	Politics	>100k	<100k
		Shared false tweets	Personal	>1m	<100k
		Shared true tweets	News	>100k	<1k
		Mostly true retweets	Personal	>1k	<10k
		Mostly false retweets	Personal	>1k	<10k
Twitter16		Shared false tweets	Personal	>100k	<100k
		Shared false tweets	Parody account	>1m	<1k
		Shared true tweets	News	>1m	<1k
		Shared true tweets	News	>100k	<100k
		Shared true tweets	Politics	>100k	<1k
MiDe22		Shared true tweets	Journalist	>1m	<10k
		Shared true tweets	News	>100k	<1k
		Shared false tweets	Personal	>100k	<10k
		Mostly true retweets	News	>100k	<10k
		Mostly false retweets	Personal	>1k	<10k

similar impact on the performance of each class. To support this outcome and investigate the effectiveness of our model, we report the top five users in Table 7.1 and argue the characteristics of the top five users.

From Table 7.1, we observe the top five users for each social network in terms of betweenness centrality scores. Solely using these users in the network to obtain similarity vectors to the identification of misinformation, we observe challenging scores (see Figure 7.3). Note that, the users are called root or engager if they share false or true tweets or retweet them, respectively. The lower performance in Twitter16 when using the top five users compared to other datasets may be related to a lack of engagers as observed in Table 7.1. This is because the engagers have many connections to root users, which provide diverse tweet content to feed our text similarity-based approach. Furthermore, we can see that the distribution of the top five users is balanced in terms of the truthfulness of their social media posts. This can highlight the promising model performance in both detecting true and false tweets when only using five users in Figure 7.3. We also observe that accounts sharing news or doing journalism mostly share/retweet

true-labeled tweets while personal accounts may tend to share misinformation detection. Moreover, we observe that the top five users have mostly greater than 100k followers, and lower than 100k followings. This can show that they have many connections with other social media users. Therefore, these users' social media posts can expose a lot of visibility and impact on their connections.

Answer to RQ5:

Overall, we observe challenging performance employing only five users with the highest betweenness centrality. Our model can effectively identify misinformation content by achieving a consistently high weighted F1 score of over 90% across all datasets, even with a small number of users. Considering that, our social networks are snapshots of very large networks, the model can achieve challenging performance with a small network. We argue that there may be no need to increase the social network size, which is costly in terms of time and data crawling. These findings can highlight its potential for early and accurate misinformation detection in low-resource settings. Furthermore, this can show the applicability of our approach to combating misinformation in real-world scenarios with extensive and dynamic social data.

Chapter 8

Limitations

We acknowledge a set of limitations in this study. First, the type of social media posts is limited to tweets and English texts. One can further construct a dataset including posts shared on other platforms (e.g., Reddit and Facebook) and other languages such as Turkish [88] to generalize the results. Our study relies solely on tweet texts and user engagements from real-world datasets. Furthermore, we conduct our experiments on relatively small datasets, which could potentially limit the generalizability of our approach. We randomly select the users from the social networks to construct user communities. Similarly, we randomly pair tweets for the fine-tuning of the SBERT model. These random selection processes can affect the robustness of our proposed approach.

Our proposed model does not incorporate propagation networks [29, 30, 46]. Future enhancements to our model could explore the integration of propagation networks to potentially improve its performance. The process of crawling user engagements may face challenges, particularly with the recent rebranding of Twitter to X. The procedure of collecting data may need to be modified as a result of the shift to a new branding.

The preparation of social networks poses inherent challenges, given its complexity and time-consuming nature. For instance, constructing accurate social

network representations involves meticulous data collection, cleaning, and integration processes. Furthermore, investigation of social networks visually requires specialized tools such as Gephi [74].

Social media users may unintentionally contribute to misinformation spread. This may lead to a potential risk of social bias against a specific user if their identities are revealed. Note that, our study investigates the supportive effect of communities to classify tweet texts instead of labeling misinformation spreaders.

Chapter 9

Conclusion

In this thesis, we curate a multi-event tweet dataset for misinformation detection that has novelty in terms of the variety of languages, i.e., English and Turkish, covering a wide range of topics. For each language, the dataset features information on 40 different events, along with how users interacted with the tweets (likes, replies, retweets, and quotes).

We propose an end-to-end text similarity-based misinformation detection framework that utilizes a community of users who interact with each other. The model consists of three stages; namely, social network construction, contrastive learning, and similarity-based classification. We observe that the top three communities, in terms of size, are prone to form a specific veracity label group. This shows community-centric dialogue about misinformation or true information spread. Inspired by the community information, we include knowledge from social networks by using similarities between users. Next, we fine-tune an SBERT model with contrastive learning. We do oversampling the datasets by making pairs and converting the human-annotated labels to pseudo-labels for the pairs. After the fine-tuning of the text encoder, we obtain similarities between the input tweet and the social network users' tweets to identify the truthfulness of the given tweet. Despite its simplicity, it achieves promising results compared to various strong baselines.

We compare our model with state-of-the-art baseline models on three real-world datasets. We show each component of our model has important contributions to the full model. We also conduct ablation studies about the choice of oversampling amount and the number of users in the communities. The choice of oversampling amount is not straightforward since it depends on the datasets. In other words, implementing a greater amount of oversampling does not mean higher performance for all cases. We obtain influential users in our social networks using the betweenness centrality metric. We then employ only a few users instead of the whole network to observe the impact of users. Even with a small number of individuals in a social network, the proposed model achieves challenging results compared to the whole network.

In future work, we plan to improve the feature quality for user representations in the communities with specific domain features, such as user timelines and profiles. Furthermore, we plan to explore the explainable artificial machine learning techniques [89] to enhance our model interpretability. Integrating methods such as retrieval augmented generation [90] is one of the possible choices since with this technique we can inject world knowledge from the web to identify misinformation. Investigation of the interpretability of our model's outcome will not only encourage its trustworthiness but also contribute to a deeper understanding of misinformation dynamics within social networks.

Bibliography

[1] O. Ozcelik, C. Toraman, and F. Can, "SiMiD: Similarity-based misinformation detection via communities on social media posts," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 1–8, 2023.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, oct 2008.

[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[4] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[5] U. K. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen, "The psychological drivers of misinformation belief and its resistance to correction," *Nature Reviews Psychology*, vol. 1, no. 1, pp. 13–29, 2022.

[6] B. Person, F. Sy, K. Holton, B. Govert, A. Liang, S. C. O. Team, B. Garza, D. Gould, M. Hickson, M. McDonald, *et al.*, "Fear and stigma: the epidemic within the sars outbreak," *Emerging infectious diseases*, vol. 10, no. 2, p. 358, 2004.

[7] M. S. Islam, T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar, *et al.*, "Covid-19–related infodemic and its impact on public health: A global social media analysis," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 4, pp. 1621 – 1629, 2020.

[8] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[9] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, sep 2020.

[10] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, pp. 211–36, May 2017.

[11] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Social Network Analysis and Mining*, vol. 13, pp. 1–36, Feb 2023.

[12] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, pp. 1–20, 2020.

[13] P. Törnberg, "Echo chambers and viral misinformation: Modeling fake news as complex contagion," *PLOS ONE*, vol. 13, pp. 1–21, 09 2018.

[14] C. R. Sunstein, "The law of group polarization," *Journal of Political Philosophy*, vol. 10, no. 2, pp. 175–195, 2002.

[15] A. R. Edwards, "(How) do participants in online discussion forums create echo chambers?: The inclusion and exclusion of dissenting voices in an online forum about climate change," *Journal of Argumentation in Context*, vol. 2, no. 1, pp. 127–150, 2013.

[16] M. Grömping, "'echo chambers': Partisan facebook groups during the 2014 thai election," *Asia Pacific Media Educator*, vol. 24, no. 1, pp. 39–59, 2014.

[17] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau, "Tweeting from left to right: Is online political communication more than an echo chamber?," *Psychological Science*, vol. 26, no. 10, pp. 1531–1542, 2015. PMID: 26297377.

[18] M. D. Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

[19] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group, 2011.

[20] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004.

[21] W. L. Bennett and S. Livingston, "The disinformation order: Disruptive communication and the decline of democratic institutions," *European Journal of Communication*, vol. 33, no. 2, pp. 122–139, 2018.

[22] E. Noelle-Neumann, "The spiral of silence a theory of public opinion," *Journal of Communication*, vol. 24, no. 2, pp. 43–51, 1974.

[23] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, (New York, NY, USA), p. 675–684, Association for Computing Machinery, 2011.

[24] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (New York, NY, USA), p. 1751–1754, Association for Computing Machinery, 2015.

[25] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PLOS ONE*, vol. 12, pp. 1–19, 01 2017.

[26] S. Gupta, R. Thirukovalluru, M. Sinha, and S. Mannarswamy, "Cimtdetect: A community infused matrix-tensor coupled factorization based method for fake news detection," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 278–281, 2018.

[27] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, (New York, NY, USA), p. 797–806, Association for Computing Machinery, 2017.

[28] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018.

[29] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), p. 395–405, Association for Computing Machinery, 2019.

[30] Y.-J. Lu and C.-T. Li, "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 505–514, Association for Computational Linguistics, July 2020.

[31] L. Tian, X. Zhang, and J. H. Lau, "DUCK: Rumour detection on social media by modelling user and comment propagation networks," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Seattle, United States), pp. 4939–4949, Association for Computational Linguistics, July 2022.

[32] X. Chen, F. Zhou, G. Trajcevski, and M. Bonsangue, "Multi-view learning with distinguishable feature fusion for rumor detection," *Knowledge-Based Systems*, vol. 240, p. 108085, 2022.

[33] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Advances in Knowledge Discovery and Data Mining*, (Cham), pp. 354–367, Springer International Publishing, 2020.

[34] J. Alghamdi, Y. Lin, and S. Luo, "Towards covid-19 fake news detection using transformer-based models," *Knowledge-Based Systems*, vol. 274, p. 110642, 2023.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[36] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1724–1734, Association for Computational Linguistics, Oct. 2014.

[37] X. Dong, U. Victor, and L. Qian, "Two-path deep semisupervised learning for timely fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 6, pp. 1386–1398, 2020.

[38] X. Li, P. Lu, L. Hu, X. Wang, and L. Lu, "A novel self-learning semi-supervised deep learning network to detect fake news on social media," *Multimedia tools and applications*, vol. 81, no. 14, pp. 19341–19349, 2022.

[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, 2006.

[40] V. Morini, L. Pollacci, and G. Rossetti, "Toward a standard approach for echo chamber detection: Reddit case study," *Applied Sciences*, vol. 11, no. 12, 2021.

[41] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, Nov. 2019.

[42] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021.

[43] Z. Yue, H. Zeng, Y. Zhang, L. Shang, and D. Wang, "MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Toronto, Canada), pp. 5223–5239, Association for Computational Linguistics, July 2023.

[44] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, (New York, NY, USA), Association for Computing Machinery, 2012.

[45] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on Twitter with tree-structured recursive neural networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 1980–1989, Association for Computational Linguistics, July 2018.

[46] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 549–556, Apr. 2020.

[47] K. A. Qureshi, R. A. S. Malick, M. Sabih, and H. Cherifi, "Deception detection on social media: A source-based perspective," *Knowledge-Based Systems*, vol. 256, p. 109649, 2022.

[48] Y. Yuan, N. Pang, Y. Zhang, and K. Liu, "Which cascade is more decisive in rumor detection on social media: Based on comparison between repost and reply sequences," *Knowledge-Based Systems*, vol. 278, p. 110857, 2023.

[49] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, pp. 44–53, 08 2017.

[50] C. Toraman, O. Ozcelik, F. Şahinuç, and F. Can, "Not good times for lies: Misinformation detection on the Russia-Ukraine war, COVID-19, and refugees," *arXiv preprint arXiv:2210.05401*, 2022.

[51] S. G. Taskin, E. U. Kucuksille, and K. Topal, "Detection of turkish fake news in twitter with machine learning algorithms," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2359–2379, 2022.

[52] G. K. Koru and C. Uluyol, "Detection of turkish fake news from tweets with bert models," *IEEE Access*, vol. 12, pp. 14918–14931, 2024.

[53] J. Ma, W. Gao, and K.-F. Wong, "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)*, (Vancouver, Canada), pp. 708–717, Association for Computational Linguistics, July 2017.

[54] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PLOS ONE*, vol. 11, pp. 1–29, 03 2016.

[55] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020. PMID: 32491943.

[56] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. Gurevych, "The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation," in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 5–9, Association for Computational Linguistics, Aug. 2018.

[57] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[58] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[59] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *arXiv preprint arXiv:1911.03854*, 2019.

[60] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104*, 2017.

[61] W. Y. Wang, ""Liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Vancouver, Canada), pp. 422–426, Association for Computational Linguistics, 2017.

[62] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (New York, NY, USA), p. 1867–1870, Association for Computing Machinery, 2015.

[63] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, p. 3818–3824, AAAI Press, 2016.

[64] L. Cui and D. Lee, "CoAID: COVID-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[65] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, "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, Dec. 2020.

[66] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "Recovery: A multimodal repository for covid-19 news credibility research," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, (New York, NY, USA), p. 3205–3212, Association for Computing Machinery, 2020.

[67] M. Weinzierl and S. Harabagiu, "VaccineLies: A natural language resource for learning to recognize misinformation about the COVID-19 and HPV vaccines," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (Marseille, France), pp. 6967–6975, European Language Resources Association, June 2022.

[68] D. Küçük and F. Can, "Stance detection: A survey," *ACM Comput. Surv.*, vol. 53, feb 2020.

[69] Q. Li, Q. Zhang, L. Si, and Y. Liu, "Rumor detection on social media: Datasets, methods and opportunities," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, (Hong Kong, China), pp. 66–75, Association for Computational Linguistics, Nov. 2019.

[70] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, p. 22–36, sep 2017.

[71] J. Zarocostas, "How to fight an infodemic," *The Lancet*, vol. 395, no. 10225, p. 676, 2020.

[72] K. Singh, G. Lima, M. Cha, C. Cha, J. Kulshrestha, Y.-Y. Ahn, and O. Varol, "Misinformation, believability, and vaccine acceptance over 40 countries:

Takeaways from the initial phase of the covid-19 infodemic," *PLOS ONE*, vol. 17, pp. 1–21, 02 2022.

[73] S. A. Memon and K. M. Carley, "Characterizing COVID-19 misinformation communities using a novel Twitter dataset," *arXiv preprint arXiv:2008.00791*, 2020.

[74] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 3, pp. 361–362, Mar. 2009.

[75] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005. IJCNN 2005.

[76] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 9–14, Association for Computational Linguistics, Oct. 2020.

[77] K. Pelrine, J. Danovitch, and R. Rabbany, "The surprising performance of simple baselines for misinformation detection," in *Proceedings of the Web Conference 2021*, WWW '21, (New York, NY, USA), p. 3432–3441, Association for Computing Machinery, 2021.

[78] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (A. Moschitti, B. Pang, and W. Daelemans, eds.), (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.

[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[80] M. G. Kim, M. Kim, J. H. Kim, and K. Kim, "Fine-tuning bert models to classify misinformation on garlic and covid-19 on twitter," *International Journal of Environmental Research and Public Health*, vol. 19, no. 9, 2022.

[81] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[82] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[83] O. Ozcelik and C. Toraman, "Named entity recognition in turkish: A comparative study with detailed error analysis," *Information Processing & Management*, vol. 59, no. 6, p. 103065, 2022.

[84] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, jan 2020.

[85] C. Toraman, O. Ozcelik, F. Şahinuç, and U. Sahin, "Arc-nlp at checkthat! 2022: Contradiction for harmful tweet detection," in *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pp. 722–739, 2022.

[86] A. Martín, J. Huertas-Tato, Álvaro Huertas-García, G. Villar-Rodríguez, and D. Camacho, "Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference," *Knowledge-Based Systems*, vol. 251, p. 109265, 2022.

[87] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[88] O. Ozcelik, A. S. Yenicesu, O. Yildirim, D. S. Haliloglu, E. E. Eroglu, and F. Can, "Cross-lingual transfer learning for misinformation detection: Investigating performance across multiple languages," in *Proceedings of the 4th Conference on Language, Data and Knowledge*, (Vienna, Austria), pp. 549–558, NOVA CLUNL, Portugal, Sept. 2023.

[89] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Explainable machine learning for fake news detection," in *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, (New York, NY, USA), p. 17–26, Association for Computing Machinery, 2019.

[90] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, Curran Associates, Inc., 2020.

Appendix A

Data and Code Availability

The datasets used in the experiments and the source code of the proposed model are publicly available and listed below:

- Twitter15 and Twitter16 datasets
  https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0
- MiDe22 dataset
  https://github.com/ogozcelik/MiDe22
- SiMiD model
  https://github.com/ogozcelik/simid-misinformation-detection

Appendix B

Annotation Tool

The open-source annotation application, INCEpTION [56], is utilized for the data annotation process. INCEpTION offers conveniences such as project creation, monitoring of the labeling process, and providing users with an interface related to the project. During the installation of the application, a database is automatically created, and the annotations made during the project are automatically saved. Since the project focuses on labeling misinformation data shared on Twitter, we improved the user interface to make the labeling process more efficient.

Figure B.1 shows the interface provided to annotators in the INCEpTION program. The tweets shown to the annotators are divided according to the main topics as mentioned in Section 3.3. The red box at the top of the page is for the fact-checked news related to an event from the main topics. We provide the title of the event, the confirmation date of the claim (17.7.2021), and the source link where the annotators can visit the web page for further information. When annotating a tweet, annotators are expected to first investigate the claim in the red box by visiting the web page and then checking the confirmation of the event.

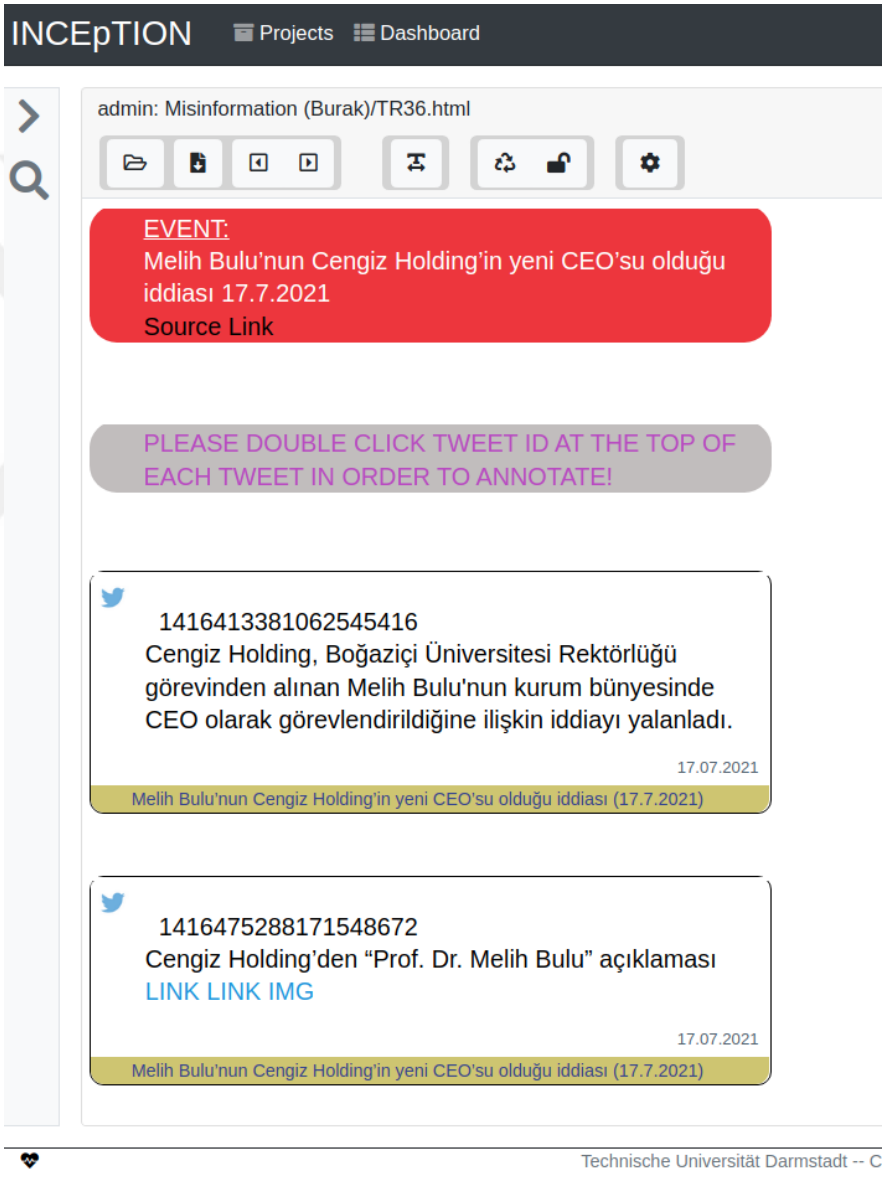


Figure B.1: The interface of the annotation tool.

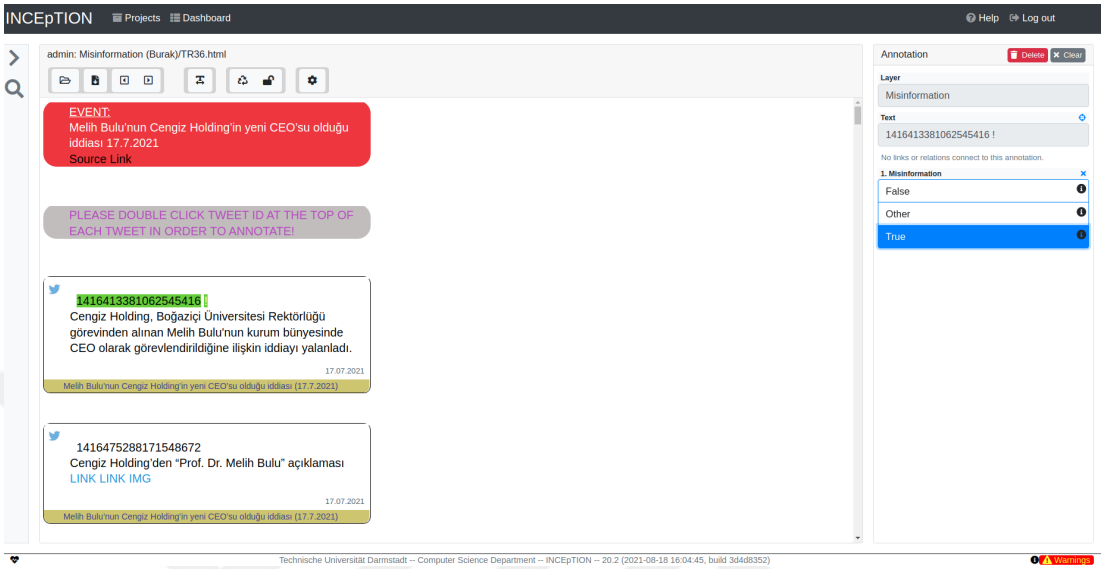


Figure B.2: The interface of the annotation tool and the annotation scheme.

When the annotator double-click a tweet ID on the page, a box appears on the right side of the page for annotation purposes (as shown in Figure B.2). The annotator is presented with three different label options in terms of veracities: False, True, and Other. After selecting the appropriate label, the annotator has the right to make changes later if necessary.