Fake Content

Sajjad Dadkhah [1], Xichen Zhang [2], Alexander Gerald Weismann [2], Amir Firouzi [2], and Ali A. Ghorbani [2]

[1]university of new brunswick
[2]Affiliation not available

October 31, 2023

Posted on 12 May 2020 — CC-BY 4.0 — https://doi.org/10.36227/techrxiv.22795130.v1 — This is a preprint. Version of Record available at https://doi.org/10.1109/TCSS.2023.3322303

Abstract

Automatic detection of fake content in social media such as Twitter is an enduring challenge. Technically, determining fake news on social media platforms is a straightforward binary classification problem. However, manually fact-checking even a small fraction of daily tweets would only be possible due to the sheer volume of daily tweets. To address this challenge, we crawled and crowdsourced one of the most extensive ground-truth datasets containing more than 180.000 labels from 2009 to 2022 for tweets with a 5-label and 3-label classification using Amazon Mechanical Turk. We utilized multiple levels of validation to ensure an accurate ground-truth benchmark dataset. We then created and implemented numerous machine learning and deep learning algorithms, such as different variations of BERT-based models, on the data to test the accuracy of real/fake tweet detection with both categories and determine which versions gave us the highest result metrics. Further analysis is performed on the dataset by explicitly utilizing the DBSCAN text clustering algorithm combined with the YAKE keyword creation algorithm to determine topics' clustering and relationships. Finally, we analyzed each user in the dataset, determining their Bot Score, Credibility Score, and Influence Score for a better understanding of what type of Twitter user posts and their influence with each of their tweets, and if there were any underlying patterns to be drawn from each score concerning the truthfulness of the tweet. The experimental results illustrated profound improvement for models dealing with short-length text in solving a real-life problem, such as automatically detecting fake content in social media.

TruthSeeker: The Largest Social Media Ground-Truth Dataset for Real/Fake Content

Sajjad Dadkhah *Member, IEEE*,, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, Ali A. Ghorbani *Senior Member, IEEE,*

Abstract—Automatic detection of fake content in social media such as Twitter is an enduring challenge. Technically, determining fake news on social media platforms is a straightforward binary classification problem. However, manually fact-checking even a small fraction of daily tweets would only be possible due to the sheer volume of daily tweets. To address this challenge, we crawled and crowdsourced one of the most extensive ground-truth datasets containing more than 180.000 labels from 2009 to 2022 for tweets with a 5-label and 3-label classification using Amazon Mechanical Turk. We utilized multiple levels of validation to ensure an accurate ground-truth benchmark dataset. We then created and implemented numerous machine learning and deep learning algorithms, such as different variations of BERT-based models, on the data to test the accuracy of real/fake tweet detection with both categories and determine which versions gave us the highest result metrics. Further analysis is performed on the dataset by explicitly utilizing the DBSCAN text clustering algorithm combined with the YAKE keyword creation algorithm to determine topics' clustering and relationships. Finally, we analyzed each user in the dataset, determining their Bot Score, Credibility Score, and Influence Score for a better understanding of what type of Twitter user posts and their influence with each of their tweets, and if there were any underlying patterns to be drawn from each score concerning the truthfulness of the tweet. The experimental results illustrated profound improvement for models dealing with short-length text in solving a real-life problem, such as automatically detecting fake content in social media.

Index Terms—Fake News Detection, automatic detection, Twitter dataset, fake and real ground truth

1 Introduction

In the contemporary era, social media has become an integral component of human existence. The exponential growth in the usage and popularity of social media has resulted in multifaceted advantages for individuals and enterprises alike. Besides providing a source of leisure and entertainment, social media platforms allow users to disseminate their original content and access a broad audience base to consume diverse information, including local and international news. The prevalence of social media has transformed the communication landscape, creating a ubiquitous platform that facilitates a diverse range of user interactions and behaviors.Sharing fake news has become easier with social media, allowing misleading or incorrect information to reach a large audience quickly. During the 2016 US presidential election, research showed that approximately 14% of Americans relied on social media as their primary news source, surpassing print and radio. One major challenge for analyzing social media platforms is collecting and labeling a large enough training dataset to be used as ground truth [1].

Although social media platforms have numerous advantages, they are also significant sources of false or inaccurate information. A vast volume of incorrect information is disseminated on social media daily, potentially resulting in adverse consequences for individuals and society. The implications of misinformation spread through social media be far-reaching and can significantly impact public perception, decision-making, and political outcomes. For instance, during the 2016 U.S. presidential election, research indicates that social media emerged as a major news source for approximately 14 percent of Americans, surpassing traditional print and radio sources. Therefore, exploring effective methods for identifying and mitigating the spread of misinformation on social media platforms is essential.

The same research [2] found that false news about the two presidential candidates, Donald Trump and Hillary Clinton, was shared millions of times on social media. Likewise, in the 2021 US presidential election campaign, recent research discovered more extensive misinformation campaigns around COVID-19. Moreover, in the aftermath of the 2021 election, specific security associations caught fake news campaigns claiming election fraud detected. These examples show that methods for identifying fake news are a relevant research topic and a pressing societal need. While different issues regarding tweet classification, such as topic or sentiment detection, are considerably researched, automatic fake news detection requires more engagement [3].

A dataset is the most critical component for the credibility and trustability of an ML/DL model. However, the limitations of the existing fake news datasets are undeniable. Most of the existing datasets need to be updated to reflect the advanced generation patterns of the new fake news creators. In addition, many online social media users and posts are unavailable after they have been detected

• *S. Dadkhah,X. Zhang,A. Weismann, M. Firouzi, and A. Ghorbani were with the Canadian Institute for Cybersecurity, the Department of Computer Science, University of New Brunswick, Fredericton, Canada.*

Manuscript received April 19, 2023; revised August 26, 2023.

as malicious or suspicious. High performance on such a dataset cannot guarantee the applicability of any model on new data input. In this paper, we designed and generated a novel Twitter dataset called TruthSeeker. As Figure 1 illustrates, we utilized the Amazon Mechanical Turk crowdsourcing platform to collect an extensive ground truth dataset of tweets for binary and multi-class classification. Furthermore, we conducted comprehensive analyses and evaluations on TruthSeeker, including DL-based detection model establishment, clustering-based event detection, and the relationship between the tweets label and the nature of the online creators/spreaders. The main contributions of this paper can be summarized as follows:

- Obtaining one of the most extensive benchmark datasets with more than 180.000 labels from 2009 to 2022.
- All the collected Tweets are verified with a three-factor active learning verification method which involves utilizing 456 unique Amazon Mechanical Turk highly skilled individuals labeling each Tweet three times by three different Turkers and verifying each decision for 2 and 5 labels by author's institution employee plus train and test models automatically to measure the effectiveness.
- To understand patterns and characteristics of Twitter users for fake/true Tweets and the impact and influence of their content, we introduced and utilized three auxiliary social media scores: Bot, credibility, and influence score.
- To evaluate the TruthSeeker dataset, we utilized different machine learning, deep learning-based detection models, and clustering-based event detection. Additionally, we explored the correlation between tweet labels and online creators/spreaders' characteristics. Our analysis provided valuable insights that enabled us to develop a more precise method for detecting fake content in social media, despite their limited length.

In the spirit of collaborative research, we are making our dataset and all related documents available for download on the Canadian Institute for Cybersecurity (CIC) dataset pages https://www.unb.ca/cic/datasets/truthseeker-2023.html.

2 Existing framework and Data sets

This section involves a detailed literature review and examination of various characteristics of multiple existing data sets for detecting fake content in social media [4], as Table 1 shows. Accurately identifying fake news is essential, and a reliable dataset is a critical component of achieving this. However, without a relevant and complete dataset, it becomes challenging to train models that can accurately identify fake news.The authors of [4] discuss the growing interest in detecting and verifying the authenticity of information related to fake news. They conducted a comprehensive survey of 118 publicly available datasets from the web. The datasets were categorized based on their focus on detecting fake news, verifying facts, analyzing fake news, and detecting satire. The researchers also examined the characteristics and uses of each dataset, highlighting challenges and opportunities for future research.

The construction of truth-based datasets has been an endeavor undertaken for many years. One of the earliest examples of combining truth scores from multiple sources is the original Politifact dataset [5] created by A. Vlachos and S. Riedel. This dataset merged the truth scores from two websites, Channel 4's fact-checking blog and the Truth-O-Meter from Politifact, into a single scale that included five labels: True, Mostly True, HalfTrue, Mostly False, and False. The dataset also includes the URLs and scores of the news. Our dataset creation process relied on this 5-label structure and a combination of expert and crowdsourced data crawling to balance qualitative and quantitative data, which is crucial for creating datasets for models to train on efficiently.

A different way to create a dataset was introduced during the creation of the PHEME dataset [6]. This dataset concentrated on five breaking news incidents and the corresponding discussions on Twitter. The objective was to distinguish between the amount of discussion about the news that was considered rumors or non-rumors. To achieve this, journalists annotated each piece of data, resulting in a relatively small dataset of about 5800 unique annotated tweets for five events.

A similarly small sample size of 2,900 tweets is used in the RumorEval-2017 dataset [7]. Attempting to train a large-scale model on such limited data would result in poor model performance and potential overfitting. Therefore, for our pipeline, we need to find a middle ground. To achieve this, we adopt the idea of expert annotations from the PHEME dataset and apply it to TruthSeeker. We use qualitative labeling by native English speakers for fact-checking each statement and ensuring accurate labeling of source statements.

Other forms of significant dataset creation, including Twitter15 and 16 datasets [8], rely on labeling JUST the source statement and leaving the information propagation up to interpretation, creating a large volume of tweets with potentially correct labels. But more likely needs more granularity and will inevitably produce poor model performance.

Despite 10+ years of work, even the most modern implementations of Politifact's data, such as the LIAR dataset [9], only have around 13000 manually labeled pieces of data. While this is impressive, the dataset could still be much larger and cover more modern forms of news propagation, such as Twitter and Facebook. To address these limitations, TruthSeeker utilizes news articles and social media (specifically Twitter) for a much larger scale of data. These early datasets served as the foundation for TruthSeeker's creation.

Evolutions of older datasets such as PHEME-update [10] and FakeNewsNet [11] seeks to remedy this issue of smaller sample sizes with increased training data. The increase in sample size is a significant improvement. In the PHEME-update dataset, this number has been more than 20x at over 6000 threads rather than the original 300. FakeNewsNet combines the rated and fact-checked news from Politifact and GossipCop to generate a dataset with almost 24,000 unique labeled pieces of information. However, the fundamental approach for generating data will always result in relatively small data size.

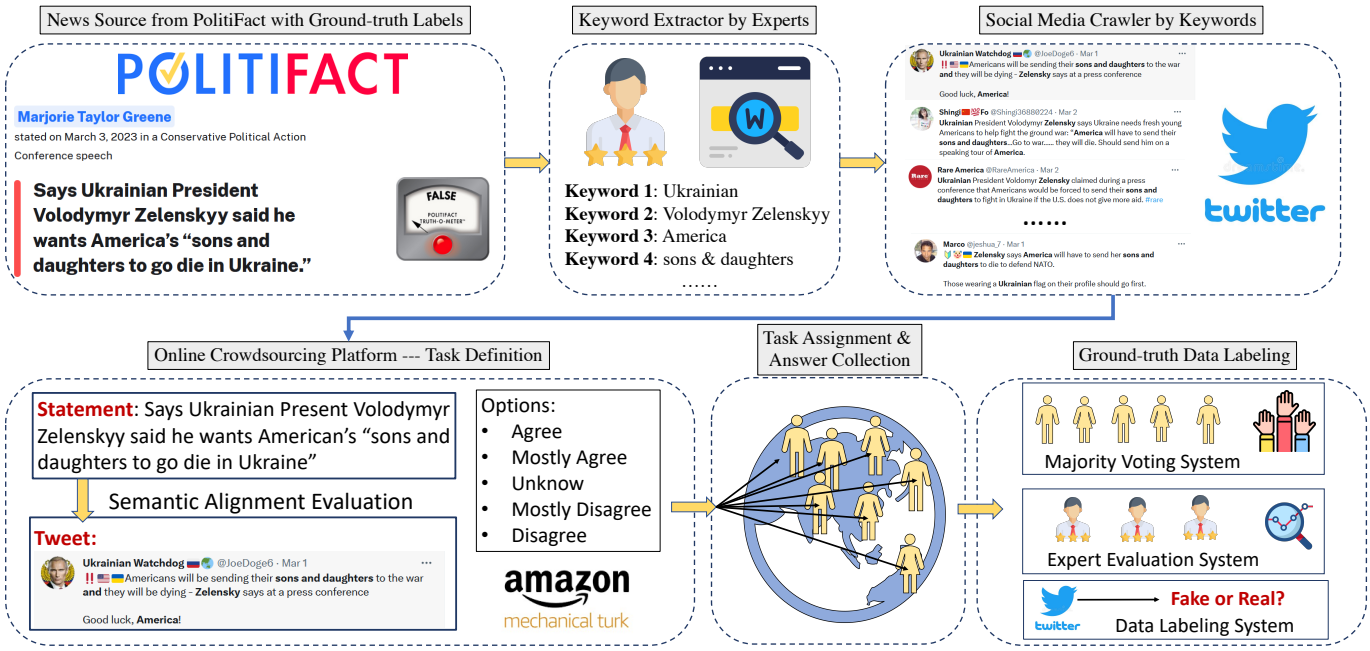


Fig. 1: The overall pipeline of the dataset generation method in this paper.

The Rumor-anomaly dataset [12], among others, produces a vast amount of tweets (4 million across 1000 rumors), but they need to be labeled individually. This is why we use a hybrid data collection and verification approach in TruthSeeker, which allows us to have similar amounts of expertly documented source statements as the original PolitiFact and PHEME datasets while generating over 140,000 actual data points from a smaller sample size. Each data point is labeled individually.

The fast detection of fake content automatically is crucial as it can prevent the spread of such content. There may be better solutions than relying on fact-checking agencies, particularly on social media. In a study by Vo et al. [13], the authors highlight the problem of spreading fake news despite fact-checking systems. They point out that these systems tend to focus on fact-checking and overlook the role of online users in disseminating false information.

In more recent times, a large focus on fake news detection and content analysis of news and tweets has centered around health-related, and specifically, COVID-19 misinformation. HealthStory [14] and HealthRelease [15] attempt to find patterns in data relating to real and fake health news and its spread throughout social media. Examining user information to determine the credibility of users who spread said information. TruthSeeker contains similar features to these two datasets (as will be discussed in a future segment) to provide as much context on the tweet and the person who posted it.

Datasets such as COVID-HeRA [16] attempt to define a more granular classification of tweets. Using categories such as (Real News / Claims, Possibly severe, Highly severe, Other, Refutes/Rebuts Misinformation). From a surface-level view, these categories are extensive. Unsurprisingly, a small data size (just over 61,000 unique tweets) with 5+ categories leads to middling f1 scores. Binary classification performs much better than expected. Similar results were noticed in our research. However, the size of the TruthSeeker dataset seemed to help improve the five-label classification results substantially.

Other covid related datasets, such as MM-COVID [17], indic-covid [18], attempt to generate multilingual datasets for fake news related to COVID-19. Creating a corpus of information large enough to train an accurate model is difficult enough in one language. Thus attempting to cover multiple ones is a herculean effort. The initial goal of the TruthSeeker dataset only included fake news detection through the English language. As English is the lingua franca of the world, it was viewed as the most critical language for generating fake news detection models. In [19] the authors examine misinformation related to COVID-19 on social networks and how it has become a problem, leading the World Health Organization to call it an "infodemic."

Various research studies [11, 20, 21, 22] have tackled the issue of identifying fake news. In a study by Helmstetter et al. [1], the automatic detection of fake news in social media was discussed as a binary classification problem. The authors acknowledged the challenge of obtaining a sizable training corpus, which led them to propose an alternative method using weak supervision to gather a large-scale but noisy training dataset. The dataset was labeled based on the source's trustworthiness, and a classifier was trained. However, the study still needed improvement in working with shorter sentences like Tweets. Despite the efforts to address the issue of fake news through research on fake news detection, more comprehensive, community-driven, and updated fake news datasets still need to be addressed. It is evident that the existing methods in this field have several issues that emphasize the necessity of a comprehensive and extensive dataset for social media, such as TruthSeeker.

ID	Dataset	Instances	Labels	Topic Domain(s)	Evaluation Type	Platform	Language(s)	Year
1	Politifact [5]	221 headlines	5	Politics, Society	Politifact, Channel 4	news articles	English	2014
2	PHEME [6]	330 threads	3	Society, Politics	Crowdsourcing	Twitter	English	2016
3	Twitter-ma [23]	992 threads	2	-	Fact-checking site (Snopes)	Twitter	English	2016
4	RumorEval2017 [7]	297 threads	3	-	PHEME [24]	Twitter	English	2016
5	Twitter15 [8]	1,478 threads	4	-	Fact-checking sites (Snopes, emergent)	Twitter	English	2017
6	LIAR [9]	12,836 claims	6	-	Fact-checking site (Politifact)	short statements	English	2017
5	Twitter16 [8]	818 threads	4	-	Fact-checking sites (Snopes, emergent)	Twitter	English	2017
7	PHEME-update [10]	6,425 threads	3	Society, Politics	PHEME [24]	Twitter	English	2018
8	FakeNewsNet [11]	23,921 news	2	Politics, Celebrity	Fact-checking sites (Politifact, GossipCop)	Twitter	English	2018
9	RumorEval2019 [25]	446 threads	3	Natural disaster	Fact-checking sites (Politcact, Snopes) (Politcact, Snopes)	Twitter, Reddit	English	2019
10	Rumor-anomaly [12]	1,022 threads	6	Politics, Fraud & Scam, Crime, Science, etc.	Fact-checking site (Snopes)	Twitter	English	2019
11	Fang [26]	1,054 threads	2	-	PHEME [24] , Twitter-ma [8], FakeNewsNet [11]	Twitter	English	2020
12	HealthStory [14]	1,690 threads	2	Health	HealthNewsReview	Twitter	English	2020
13	HealthRelease [14]	606 threads	2	Health	HealthNewsReview	Twitter	English	2020
14	CoAID [15]	4,251 threads	2	Covid-19	Fact-checking sites (Politifact, FactCheck.org, etc.)	Twitter	English	2020
15	COVID-HeRA [16]	61,286 posts	5	Covid-19	CoAID, Expert annotators	Twitter	English	2020
16	ArCOV19-Rumors [27]	162 threads	2	Covid-19	Fact-checking sites (Fatabyyano, Misbar)	Twitter	Arabic	2020
17	MM-COVID [17]	11,173 threads	2	Covid-19	Fact-checking sites (Snopes, Poynter)	Twitter	English, Spanish, Portuguese, Hindi, French, Italian	2020
18	Constraint [28]	10,700 posts	2	Covid-19	Fact-checking sites (Politifact, Snopes)	Twitter	English	2020
19	Indic-covid [18]	1,438 posts	2	Covid-19	Expert annotators	Twitter	Bengali, Hindi	2020
20	COVID-Alam [29]	722 tweets	5	Covid-19	Expert annotators	Twitter	English, Arabic	2021
21	COVID-RUMOR [30]	2,705 posts	2	Covid-19	Fact-checking sites (Snopes, Politifact, Boomlive)	Twitter, Websites	English	2021
22	**TruthSeeker**	**186,000 tweets**	**5,2**	**Politics, General Events, Health, Crime, Science, etc.**	**Crowdsourcing (Amazon Mechanical Turkers), & Expert Evaluation**	**Twitter**	**English**	**2022**

TABLE 1: The current detests for fake and true content in social media

3 Dataset Creation

The creation of the TruthSeeker dataset begins with a combination of Real and Fake news crawled from the PolitiFact website. From this data, keywords relating to each piece of text are generated. This is done by painstakingly, manually generating keywords for 700 Real and 700 Fake pieces of news. Many automated keyword generation algorithms were attempted to speed up this manual process (using python packages such as attention approach, **PKE** (Python Keyphrase Extraction) , **RAKE** (Rapid Automatic Keyword Extraction), **RaKUn** (Rank-based Unsupervised Keyword), and **YAKE** (Yet Another Keyword Extractor)). However, in preliminary testing they resulted in poor keyword generation. Providing either:

1. Too few keywords to get meaningfully related tweets when calling the Twitter's Full-archive search API
2. So many keywords that when used in the Full-archive search API, the combination would be too hyper-specific to return any results at all.

This problem occurred with every keyword extraction algorithm attempted. Leading to the conclusion that the sensitive nature of the twitter keyword searching and and low reliability of proper keyword generation algorithms was something that would not provide meaningful or useful results. Making the choice to go the route of manual keyword generation over automated obvious.

Manual keyword generation was the most effective as each set of keywords could be constructed to best summarize the article titles and return the most results. Taking a qualitative approach to assure the most accurate data possible. A general rule followed was to create at a minimum 2 keywords and a maximum of 5 keywords for any of the associated pieces of text. It was observed through both the automated keyword generation and manual creation that any less or more would result in tweets that were unrelated to the topic or so hyper specific that nothing would exist for the results. Thus, the limit of 2-5 keywords was created. The final amount of tweets crawled for 700 Real and 700 Fake pieces of ground-truth news was slightly under 186,000 tweets. Giving on average 133 tweets per piece of news. Exceeding our initial predictions.

Results: Using the getStats() API call from our custom Twitter integrated endpoint we can observe that this piece of news (with the unique ID of 19) returns 88 tweets utilising the manual keywords listed above.

The getTweets() endpoint returns all associated tweets with their full meta data information (created_at, id, text, etc...) in the JSON (JavaScript Object Notation) format. Below is an example of the information returned of one tweet from the 88 of the ground-truth news title with the ID 19:

For the creation of this dataset, the main pieces of information extracted from the returned JSON data were a cleaned version of the "text" called "cleaned_text", the twitter id of the user called "id", and the time of the tweets creation called "created_at". Table 2 illustrates some a sample of the data sent to amazon tukers. This information was then processed and saved in a .csv (Comma Separated Value) file with the appropriate formatting to later be fed into the Amazon Mechanical Turk system. The format is as follows:

Algorithm 1: Example JSON output

Result: Sample values returned from calling the custom getTweets() endpoint

"created_at": "Sun Oct 17 14:19:37 +0000 2021",
"id_str": "1449741894037213184",
"text": "It's not even about jobs. The coal industry has been in decline a 11 by itself for years. Many of the old coal jobs... https://t.co/pMUX6FKZwD"
"user":

"id": 801080368,
"id_str": "801080368",
"name": "Patrick S. Tomlinson",
"screen_name": "stealthygeek",
"location": "Milwaukee",

Statement: End of eviction moratorium means millions of Americans could lose their housing in the middle of a pandemic. **Keywords:** Americans, eviction moratorium

ID	tweet	timestamp
0	Tonight, millions of Americans will sleep without the specter of eviction over their head, but we must move to permanently help them. I want to thank @RepCori for her activism to get this eviction moratorium in place.	Thu Aug 05 03:37:00 +0000 2021
0	On Mon. at 11am EST, US Sens. @ReverendWarnock, @CoryBooker, @ewarren, and a group of Black farmers will hold a briefing on the historic action taken when the American Rescue Plan Act of 2021 was signed into law, a bill that will repeal major injuries to farmers of color. 1/6	Fri Mar 19 21:14:38 +0000 2021
0	Super early this morning we passed the American Rescue Plan. I voted YES for working families, small businesses, farmers, teachers, healthcare workers, and everyone in need of more resources and relief because of this pandemic.	Sat Feb 27 20:40:34 +0000 2021

TABLE 2: Example subset of data sent to the Amazon Mechanical Turkers

Each tweet occupies its own row and includes the meta data that was discussed earlier. A copy of the ground-truth "statement" which is the original article title. The "manual_keywords", and that article title's unique "id" or "query_id". This duplication is done as it is required for creating individual tasks to be completed using the Amazon Mechanical Turk system. One last check is done to make sure that there contains no non UTF-8 encoded characters or symbols. After this, the .csv file is uploaded to Amazon Mechanical Turk for processing.

4 Crowd-soursing and labeling utilizing Amazon Mechanical Turk

The Amazon Mechnical Turk service was a key part of the creation of the TruthSeeker dataset. Allowing for the construction of a much larger dataset with the help of "Turkers" (individuals performing an Amazon Mechanical Turk Task) rather than manually assessing each tweet. Each row of our dataset was translated to and treated as a HIT (Human Intelligence Task). A micro job for a Turker to complete. Below a visualization of the HIT itself is shown.

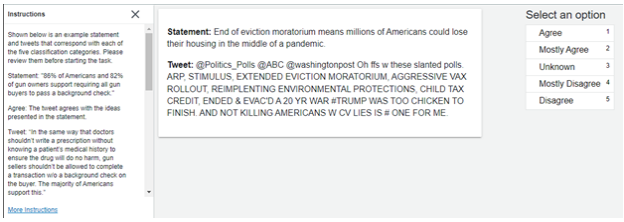


Fig. 2: Example view the Master Turker sees when completing the HIT.

- Our HIT was limited to only Master Turkers. Meaning that only Turkers assessed by Amazon to be of the highest quality were allowed to participate in our HIT. This allowed us to be sure of having the highest skilled Turkers to make the judgments in the tasks we assigned them, rather than rush through to receive payment as fast as possible. Giving us a baseline skill and competency level that using non Master Turkers would not have been guaranteed to afford us.
- The HIT that we published for the Amazon Turkers to complete was a variation of a basic semantic similarity task. We asked the Master Turkers to examine the source statement (i.e. "statement" from the above photo) and an accompanied tweet. They would then need to decide to what degree the tweet agrees with the statement.

A set of the instructions on the side bar was also included for the Turker to read before beginning their task. The instructions provided definitions for each of the 5 options (Agree, Mostly Agree, Unknown, Mostly Disagree, Disagree) and an example tweet that would match each of the the categories.

> Statement: "86 percent of Americans and 82 percent of gun owners support requiring all gun buyers to pass a background check."
> Agree: The tweet agrees with the ideas presented in the statement.
> Tweet: "In the same way that doctors shouldn't write a prescription without knowing a patient's medical history to ensure the drug will do no harm, gun sellers shouldn't be allowed to complete a transaction w/o a background check on the buyer. The majority of Americans support this."
> Mostly Agree: The tweet agrees with the majority of the ideas presented in the statement.
> Tweet: "more than 50 percent of Americans are in favor of some form of gun control, whether it be background checks or something else entirely. . . "
> Unknown: The tweet neither aligns nor differs with the presented statement.
> Tweet: "America is a country that loves guns."
> Mostly Disagree: The tweet disagrees with the majority of the ideas presented in the statement.
> Tweet: "I understand that some people are in favor of background checks, but most REAL Americans are not."
> Disagree: The tweet disagrees with ideas presented in the statement.
> Tweet: "Democrats are busy clutching their pearls over gun control, they claim the founders wouldn't support current Americans right to bear arms. I'd like to remind Democrats, our Founders had just finished a war against their former countrymen. Shall no be infringed is pretty clear."

A final measure was taken was to ensure higher accuracy of HIT responses. This consisted of having each HIT be completed by 3 separate Master Turkers. This allowed us to further verify the final label that would be applied to each tweet after all HIT's were completed.

5 Results

The results we received from the Master Turkers were classified in two separate ways. A 5-way label which included all the original categories (Unknown, Mostly True, True, False, Mostly False) and a 3-way label (Uknown, True, False). The creation of the 5-way labeled dataset was much more restrictive in terms of allowing data to be used. The specific protocol was as follows:

Algorithm 2: 5-Way Label Creation Protocol

```
Result: Dataset of 5 - Way labelled values
while not at end of returned Master Turker dataset do
    read current data;
    if 2/3 majority of same labels then
        add data to dataset;
    else
        /* results too chaotic to use with
           any degree of confidence     */
        remove from dataset;
    end
end
```

If a two-thirds majority exists (i.e. 2/3 Turkers agreed upon a label) then said majority value becomes the final classification label of that tweet. However, if no agreement occurs then the data is expunged. The 3-way labelling is much more forgiving, however.

If 2/3 of the results are positive (i.e. Mostly True, True), negative (i.e. Mostly False, False), or marked Unknown then the final result is labelled as True, False, or Unknown. This method allows for the retention of much more data while still maintain a high level of confidence in the results being accurate. As there was at least some shared sentiment towards the validity of the news and therefore truthfulness

Algorithm 3: 3-Way Label Creation Protocol

```
Result: Dataset of 3 - Way labelled values
while not at end of returned Master Turker dataset do
    read current data;
    if 2/3 majority similar labels then
        /* (True/Mostly True, False/Mostly
           False, or Unknown/Unknown)   */
        add data to dataset;
    else
        /* results too chaotic to use with
           any degree of confidence     */
        remove from dataset;
    end
end
```

of the tweet.Table 3 illustrates breakdown of the 5-way label and 3-way label results using both Master and Standard Amazon Mechanical Turkers. This comparison was done to gauge the quality of Master Turkers over standard ones, as well as show the spread of results from an initial test batch of 1000 tweets.

	Master Turker	Standard Turker	Master Turker	Standard Turker
	5-Way Label		3-Way Label	
Unknown	36.40%	26.99%	36.40%	27.14%
Mostly True	13.80%	20.84%	NA	NA
True	26.20%	20.84%	40.06%	41.52%
False	14.40%	15.94%	23.49%	31.33%
Mostly False	9.08%	15.37%	NA	NA

% of same results from Master and Standard Turkers (5-Label):	36.32%
% of same results from Master and Standard Turkers (3-Label):	55.39%

TABLE 3: 5-way label and 3-way label

Below is a random news statement pulled from our data and an associated tweet of each category (Agree, Mostly Agree, Unknown, Mostly Disagree, Disagree) related to it:

Statement: "Ivermectin sterilizes the majority (85%) of the men who take it."

Tweet(Unknown): ...Now their 'treatment alternative' is not just killing them, but rendering the men functionally or fully sterile.They claimed the free vaccine harms women's fertility & genetics, so instead they pay big bucks for Ivermectin, which mutates sperm & sterilizes the men!

Tweet(False): @90mifromneedles @Blackamazon I think the no schadenfreude train left without me. I saw Ivermectin apparently sterilizes the majority (85 percent) of men that take it & followed the link to the study. My 1st thought was well, at least those pushing its use for COVID will no longer contrib to the gene pool.

Tweet(Mostly False): @Acyn Ivermectin will make them shit out their stomach linings and sterilizes men LOL

Tweet(True): @redsteeze @JerseyWalcott That's absurd. Pretty soon, they're going to start claiming that (life giving) Ivermectin sterilizes men & shrinks their sexual organs.

Tweet(Mostly True): @jeek The study you linked does not say that it sterilizes 85 percent of men that take it. It says that "a recent report showed that 85 percent of all male patients treated in a particular centre with ivermectin in the recent past who went to the laboratory for routine tests were discovered to...

Unknown Tweets Examples: The "Unknown" category contained issue in it being a catch-all category that the Master Turkers used when they were unsure of what response to give, rather than when a tweet had an unknown relation to source statement. However, many of the below examples are not unknown in reality. It may be advantageous to either remove the classification category fully or to split it up into more granular categories to get more accurate results:

Statement: "Ivermectin sterilizes the majority (85 percent) of the men who take it."

6 TruthSeeker Model Analysis

Below showcases the results of training two model types on the TruthSeeker dataset. The first being a standard Binary Classification Model (with categories True and False). The second a 4 - Label Classification Model (False, Mostly False, Mostly True, True). Both of these models attempt to predict the truthfulness of a tweet using various classification categories.

The final TruthSeeker dataset exists in one CSV file that is prepossessed and later used for training our model. Its raw structure is illustrated below:

Feature	Description
statement	Headline of a news article
target	The ground-truth value of the statement
BinaryNumTarget	Binary representation of the target value (1 = True / 0 = False)
manual_keywords	Manually created keywords used to crawl the twitter API
query_id	ID associated with the manual_keywords. Used to reference the associated JSON file with more information.
tweet	Twitter posts related to the associated manual keywords
tweet_id	Unique ID of the twitter post
timestamp	Time the tweet was generated
5_label_majority_answer	Majority answer using 5 labels (Agree, Mostly Agree, Disagree, Mostly Disagree, Unrelated)
3_label_majority_answer	Majority answer using 3 labels (Agree, Disagree, Unrelated)

TABLE 4: List of features in the Truth Seeker dataset with their associated descriptions.

The features that are utilized for creation of the models are:

- **statement**
- **BinaryNumTarget**
- **tweet**
- **5_label_majority_answer**
- **3_label_majority_answer**

6.1 Dataset Preprocessing

After importing the CSV file there are a few reprocessing steps done to the data before model creation. Firstly any

Statement (T/F)	Majority Answer	Truthfulness
T	Agree	True
T	Disagree	False
T	Mostly Agree	Mostly True
T	Mostly Disagree	Mostly False
F	Agree	False
F	Disagree	True
F	Mostly Agree	Mostly False
F	Mostly Disagree	Mostly True

TABLE 5: 4-Label Conversion Truth Table.

rows with a majority answer column value of ”NO MAJORITY” or ”Unrelated” are removed. This is due to the ”NO MAJORITY” label indicating that the analysis of the tweet by 3 separate Amazon Turkers was inconclusive as to the label it should receive. The Unrelated label was used to weed out tweets not directly related to the statement being made, and thus unusable for determining the truth of the tweet in relation to the original statement.

these rows are dropped (using basic dataframe comprehension) and the new dataset is split into two separate dataframes. One containing all data except for the 5_label_majority_answer and one containing all but the 3_label_majority_answer column.

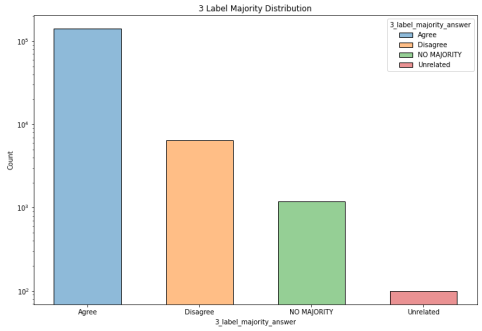
For each of the two newly created dataframes we generate a ”ground_truth_value” and ”categorical_label” columns. the ”ground_truth_value” column takes the BinaryNumTarget of the statement and the majority answer of the tweet as inputs and generates a truthfulness value. Below are the logic tables for the 4 - Label conversion and the 2 - Label conversion.Table 5 illustrates 4-Label Conversion Truth Table. Takes into account the original statements’ validity and the majority answer of the tweet. Then a final truthfulness value is assigned to it. Table 6 shows 2 - Label Conversion Truth Table. Similar to the previous truth table in nature, only two truthfulness values are possible.

After this conversion, the labels are encoded and placed in the ”categorical_label” column for easier use. This dataset contains 150,000 unique tweets coinciding with 1400 unique statements and their manually generated keywords. The balancing of this dataset is exactly 50/50 for True and False statements.

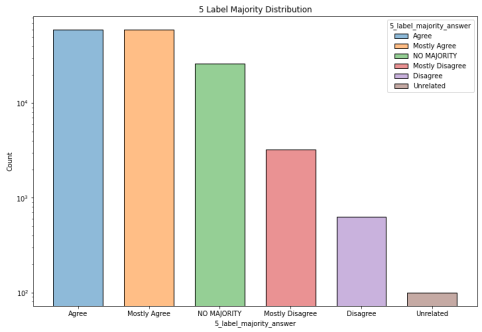
3b showcases a clear majority of Turkers found tweets related to a source statement tend to either agree or mostly agree with source statement. A large percentage of the data was inconclusive and thus marked as NO MAJORITY. Adjusting for this in 3a using a two-thirds majority rule, we see that the majority of the Turker results that we once no majority can be grouped into either the agree column of disagree column, after performing the two-thirds majority conversion. We can see that the turkeys determined the majority of tweets in relation to the source statement are in agreement, with a small subset of disagreeable response

Statement (T/F)	Majority Answer	Truthfulness
T	Agree	True
T	Disagree	False
F	Agree	False
F	Disagree	True

TABLE 6: 2 - Label Conversion Truth Table.



(a)



(b)

Fig. 3: Histograms showcasing the distribution of crowdsourced results from the Amazon Mechanical Turkers.

or answer too difficult to place into either of the major categories easily.

7 Model Testing

It can be difficult to extract important information from short texts like Tweets, even with accurate labels. Our study used 50 unique features and six different machine-learning models (which you can see in Table 8). These features included the number of uncommon or complex words, adjectives and metadata like how many replies the user has. As Table 7 showed, we achieved impressive results in detecting fake social media content, especially considering the limited amount of reliable data available for short texts.

In the next section, we’ll show you how we can improve these results even further by using different versions of Bert model.

	TP	FP	Pr.	Re.	F-1	Acc.
J-48 Decision Tree	0.623	0.378	0.623	0.623	0.623	0.623
Random Forest	0.701	0.300	0.701	0.701	0.701	0.701
IBK KNN	0.626	0.375	0.626	0.626	0.626	0.626
Bays Network	0.618	0.382	0.618	0.618	0.618	0.618
Ada Boost	0.595	0.406	0.595	0.595	0.595	0.595
Log Reg	0.631	0.371	0.631	0.631	0.631	0.631

TABLE 7: Models using 50 texts, lexical, and meta-data features

With the TruthSeeker dataset fully developed and realized, the next goal of our research was to implement multiple BERT (Bidirectional Encoder Representations from Transformers) based models to see if it would be possible to accurately asses the truthfullnes of a tweet. Below we

Tweet Feature Categories	
Text Features	
Name	Description
unique_count	number of unique, complex words
total_count	total number of words
ORG_percent	Percent of text including spaCy ORG tags
NORP_percent	Percent of text including spaCy NORP tags
GPE_percent	Percent of text including spaCy GPE tags
PERSON_percent	Percent of text including spaCy PERSON tags
MONEY_percent	Percent of text including spaCy MONEY tags
DATA_percent	Percent of text including spaCy DATA tags
CARDINAL_percent	Percent of text including spaCy CARDINAL tags
PERCENT_percent	Percent of text including spaCy PERCENT tags
ORDINAL_percent	Percent of text including spaCy ORDINAL tags
FAC_percent	Percent of text including spaCy FAC tags
LAW_percent	Percent of text including spaCy LAW tags
PRODUCT_percent	Percent of text including spaCy PRODUCT tags
EVENT_percent	Percent of text including spaCy EVENT tags
TIME_percent	Percent of text including spaCy TIME tags
LOC_percent	Percent of text including spaCy LOC tags
ORG_percent	Percent of text including spaCy ORG tags
WORK_OF_ART_percent	Percent of text including spaCy WOA tags
QUANTITY_percent	Percent of text including spaCy QUANTITY tags
LANGUAGE_percent	Percent of text including spaCy LANGUAGE tags
Max Word	length of the longest word in the sentence
Min Word	length of the shortest word in the sentence
Avg Word Length	average length of words in the sentence

Lexical Features	
Name	Description
present_verb	number of present tense verbs
past_verb	number of past tense verbs
adjectives	number of adjectives
pronouns	number of pronouns
TO's	number of to usages
determiners	number of determiners
conjunctions	number of conjunctions
dots	number of (.) used
exclamations	number of (!) used
question	number of (?) used
ampersand	number of (&) used
capitals	Number of capitalized letters
quotes	number of quotation makes used
digits	number of digits (0-9) used
long_word_freq	number of long words
short_word_freq	number of short words

Meta-Data Features	
Name	Description
followers_count	number of followers
friends_count	number of friends
favourites_count	number of favorites across all tweets
statuses_count	number of tweets
listed_count	number of tweets the user has in lists
mentions	number of times the user was mentioned
quotes	number of times the user has been quote tweeted
replies	number of replies the user has
retweets	number of retweets the user has
favourites	number of favourites the user has
hashtags	number of hashtags the user has used
URLs	number of URLs the user has posted
BotScoreBinary	Binary score whether the user is considered a bot or not
cred	credibility score
normalized_influence	influence score the user has, normalized.

TABLE 8: Tweet features utilized for ML models

implement 4 variations of as well as the original BERT model. Those being specifically:

- **ROBERTA**
- **BERT**
- **DISTILBERT**
- **BERTWEET**
- **ALBERT**

The results of each model's performance and metrics are as follows. Figure 4a-4d illustrates the results of running the ROBERTA model on the TruthSeeker dataset for 4 Epochs. Extremely promising accuracy and f1 scores are achieved as seen in 4b and 4c with accuracy and f1 score of almost 96% achieved and a relatively low amount of training time. This model appears to converge around 4 Epochs, thus making it doubtful that any meaningful improvements could be made with additional training time concerning iterations.

Figure 4e-5d illustrates the results of running the ROBERTA model on the TruthSeeker dataset for 10 Epochs. The results are not as great but still quite promising. With ten epochs, we see the accuracy hit almost 69% 4e with no apparent convergence. More tests with higher epochs could have us achieve an accuracy of 70% or higher. Other hyperparameters could also be tweaked to see if any meaningful improvement is noticed.

Figure 5k-5l illustrates the results of running the classical BERT model on the TruthSeeker dataset for 5 Epochs. We are able to achieve an accuracy 5j slightly higher than using ROBERTA, DISTILBERT, and ALBERT with our Binary Label. Though they are still fairly close matches. This marginal difference is also potentially attributed to the 1 Epoch difference in training and increased model size of BERT compared to the others mentioned.

Figure 4m-4p illustrates the results of running the classical BERT model on the TruthSeeker dataset. While the results are pretty underwhelming, they are consistent with the accuracy of other pre-trained models. As can be seen in 4m, the model seems to converge with a relatively low accuracy 4n and high evaluation loss 4p. More training time/iterations seem unlikely to generate better results and are more than likely to overfit the model to our dataset.

Figure 5a-5d illustrates the results of running the DISTILBERT model on the TruthSeeker dataset. Results for accuracy 5b and f1 5a are pretty high on this model also.

model	acc	cohen	eval_loss	f1	mcc	prec	recall
ROBERTA	0.9569	0.9137	0.1720	0.9569	0.9137	0.9569	0.9569
BERT	0.9596	0.9192	0.1611	0.9596	0.9192	0.9596	0.9596
DISTILBERT	0.9587	0.9174	0.1616	0.9587	0.9174	0.9587	0.9587
BERTWEET	0.9610	0.9220	0.1653	0.9610	0.9221	0.9610	0.9610
ALBERT	0.9496	0.8993	0.2067	0.9496	0.8993	0.9497	0.9496

TABLE 9: Results of all 2 - Label Classification models.

model	acc	cohen	eval_loss	f1	mcc	prec	recall
ROBERTA	0.6893	0.5859	0.6755	0.6883	0.5867	0.6903	0.6893
BERT	0.5102	0.3471	0.9012	0.5088	0.3476	0.5099	0.5102
DISTILBERT	0.4964	0.3284	0.8951	0.4957	0.3287	0.4963	0.4964
BERTWEET	0.4922	0.3232	0.8596	0.4890	0.3245	0.4912	0.4922
ALBERT	0.4857	0.3141	0.9066	0.4857	0.3141	0.4861	0.4857

TABLE 10: Results of all 4 - Label Classification models.

Giving us a marginally lower accuracy than the base BERT model, yet still maintaining around 95%. Being around 40% smaller than the original BERT model and losing a marginal amount of performance because of this may be the cause of the slightly reduced statistical values for this model. Figure 5e-5h illustrates the results of running the BERTWEET model on the TruthSeeker dataset.

Boasting the highest accuracy 5f and f1 scores 5e of all pre-trained models attempted. BERTWEET seems to provide the best results with our dataset. Being the first public large-scale pre-trained language model for English Tweets, this is not surprising. Figure 5i-5l illustrates running the ALBERT model on the TruthSeeker dataset. Table 9 illustrates the Results of all 2 - Label Classification models. BERTWEET showcases a clear improvement over all other model types with the highest accuracy and f1 scores. Table 10 illustrates the Results of all 4 - Label Classification models. ROBERTA appears to have the highest overall performance. With the lowest accuracy, the ALBERT framework lightweight BERT approach results in poorer performance. However, the performance is still impressive, with a score of over 94% in the two previously mentioned metrics.

8 Social post Text Clustering:

This section focuses on the results from running the DBSCAN text clustering algorithm on our TruthSeeker Dataset with different hyperparameters. We embed our tweets using the Sentence Transformer (allmpnetbasev2) and then apply the DBSCAN algorithm with varying epsilon values after we use the YAKE keyword extractor on each cluster of data

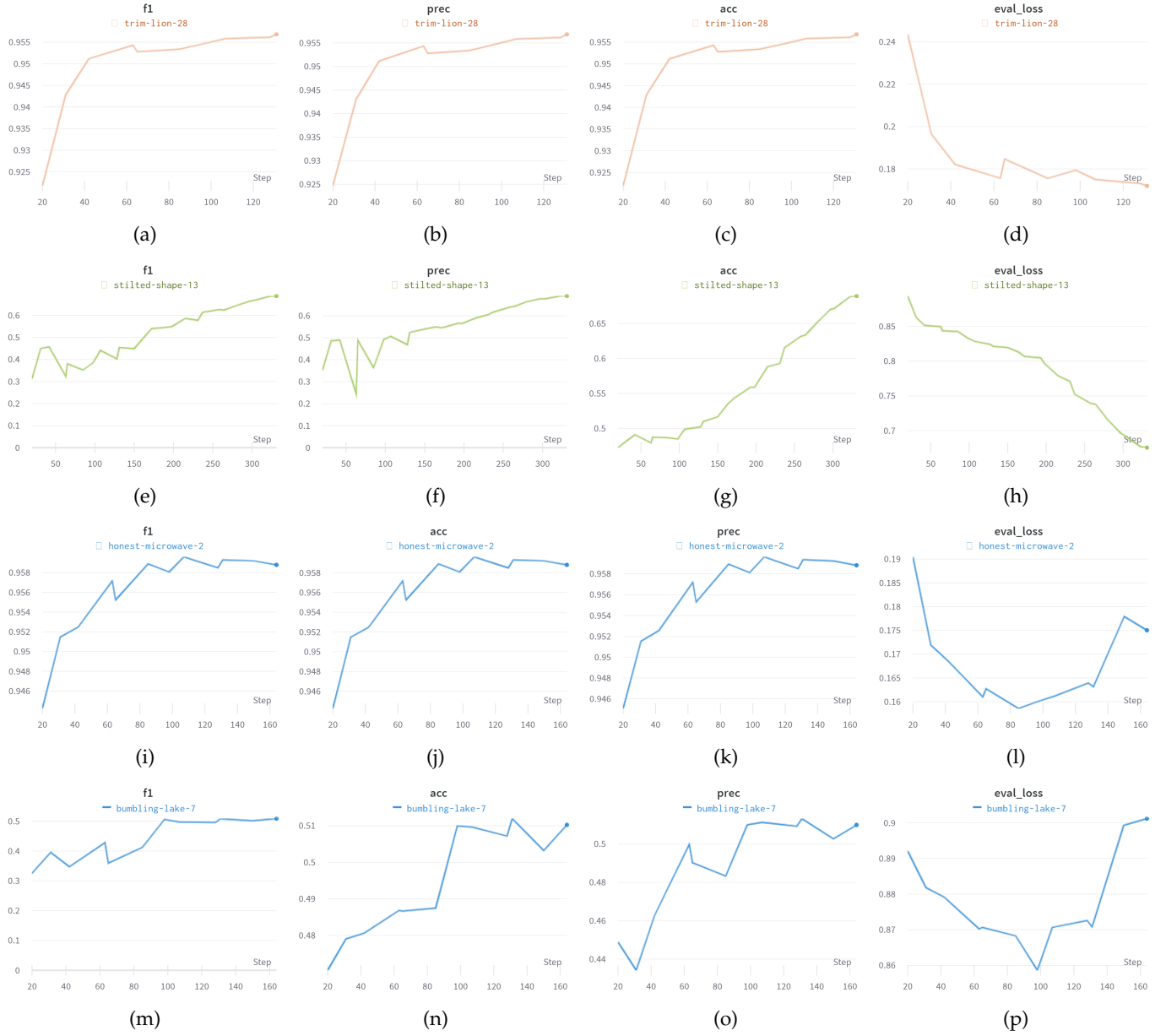


Fig. 4: 4 Epoc-ROBERTA V1.0: 4a Precision maximized at 0.9569,4b Accuracy maximized at 0.9569, 4c F1 score maximized at 0.9569, and 4d Evaluation loss minimized to 0.1720 .Experiment 1-10 Epoc-ROBERTA V1.0: 4e Precision maximized at 0.6903,5b Accuracy maximized at 0.6893, 5c F1 score maximized at 0.6883, and 5d Evaluation loss minimized to 0.6755 .Experiment-BERT 2 label: 5i F1 score maximized at 0.9596,5j Accuracy maximized at 0.9596,5k Precision maximized at 0.9596 and 5l Evaluation loss minimized to 0.1611 . Experiment-BERT 4 label: 4m F1 score maximized at 0.5088,4n Accuracy maximized at 0.5102,4o Precision maximized at 0.5099 and 4p Evaluation loss minimized to 0.9012

created to get an idea of what each cluster is referencing with our tweets/news.

After, we take the list of keywords and remove duplicates / sub-strings while also considering the case sensitivity of words. We then display the top 10 clusters and their associated cleaned keywords. Below are the results of these tests and their outputs. As Table 11 illustrates, the development of DBSCAN algorithm (After applying the DBSCAN clustering to the Fake and Real Tweet data with different epsilon values, The top 10 clusters ranked by size are shown in Table 11)clustering resulted in more than 100 clusters with precise keywords detected for each cluster, which shows how versatile the data in TruthSeeker are, makes a perfect dataset for training automatic detection algorithms in fake news domain. Having access to the Twitter API V.2 Full Archive search enabled us to view tweets as far back as 2007 (The founding of Twitter) and, in our case, from 2009 to 2022.

9 User Analysis

This section focuses on analysing the individual users for each tweet crawled in our data set creation. Focusing on 3 metrics. Bot Score, Credibility, and Influence.

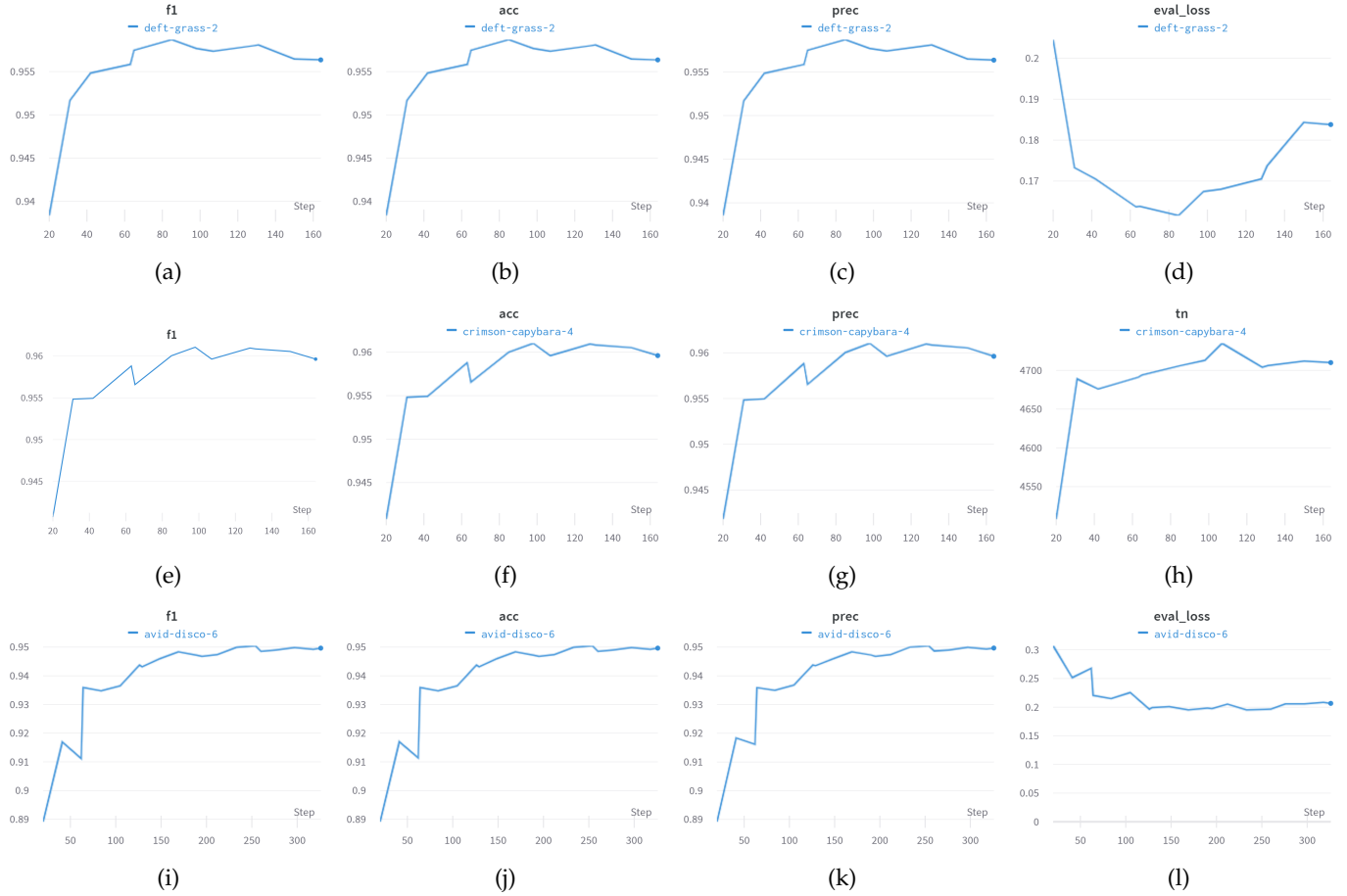


Fig. 5: DISTILBERT: 5c Precision maximized at 0.9587,5b Accuracy maximized at 0.9587, 5a F1 score maximized at 0.9587, and 5d Evaluation loss minimized to 0.1720 .BERTWEET: 5g Precision maximized at 0.9610,5f Accuracy maximized at 0.9610, 5e F1 score maximized at 0.9610. ALBERT: 5i F1 score maximized at 0.9496,5j Accuracy maximized at 0.9496,5k Precision maximized at 0.9497 and 5l Evaluation loss minimized to 0.2067.

9.1 Bot Score

A users Bot Score is a value between 0.0 and 1.0 that is determined by a model trained on 17 features using the users specific follower count (number of followers), friend count (number of friends), favourite count (number of favourites), status count (number of tweets), account age (age of the account), list count (lists created), and url (number of url's posted). With 1.0 being the highest likelihood of being a bot and 0.0 being the lowest.

Figure 6 represents the results of running this bot score test on all data. Any results given a score of less than or equal to 0.5 are considered "Not Bot," whereas anything greater than 0.5 is considered "Bot."As can be seen, bots make up a minority of the overall data but are still sizeable enough for there to be a potential for false information to be disseminated. The split of bots in both fake and real topics is very similar, showcasing that bots are included in all issues on Twitter.

The results above indicate that a user that interacts with True rather than Fake topics appears to have a higher credibility. Which on a common sense level appears to be accurate. More credible people spend time engaging with real topics rather than fake ones.

9.2 User Influence

We can classify an influential user in a few separate ways. Firstly we define an influential user as a user whose actions in a network are capable of affecting the actions or thoughts of many other users in the network.

Below are a few formulas proposed for calculating the influence score of each individual user in a network:

$$GA(i) = \frac{(OT1 + RP1 + RT1 + FT1)}{TotalTweets} \quad (1)$$

$$SignalStrength(i) = \frac{OT1}{(OT1 + RT1)} \quad (2)$$

Equation 1 represents the general activity, and equation 2 is signal strength. Where $OT1$ is the number of original tweets (OTs) the author posts. $RP1$ is the number of replies posted by the author. $RT1$ is the number of retweets accomplished by the author.$FT1$ is the number of tweets of other users marked as favorite (liked) by the author [31].

$$NetworkScore(i) = \log(F2 + 1) - \log(F4 + 1) \quad (3)$$

$\varepsilon = 0.20$			$\varepsilon = 0.25$			$\varepsilon = 0.30$		
Cluster #	**Keywords**	**# Tweets**	**Cluster #**	**Keywords**	**# Tweets**	**Cluster #**	**Keywords**	**# Tweets**
54	Joe Biden Trumps tax cuts Raise taxes Americans	940	9	Pfizer covid vaccine Vaccine Injury Act Covid survival rate Nuremberg Code Fda approved Survival Rates Age	6432	0	President Joe Biden African American children Wisconsin Michigan Arizona Trumps tax cuts	47694
29	million defenseless people Americans gun control United States Century Guns wars died	626	2	Fulton County Georgia Maricopa county Arizona	1595	5	Black Lives Matter Trump labeled antifa terrorist organization Antifa	1103
7	Obama played rounds of golf Trump	472	19	Joe Biden Trumps tax cuts raise taxes Americans	1344	11	NANCY PELOSI Speaker Nancy Pelosi pelosi made millions capitol police insider trading Capitol Police Chief	1047
46	IRS code Ted Cruz Bible words GOPDebate	467	14	million defenseless people United States Americans gun control PEOPLE KILLED FBI killed Guns	1020	6	Obama played rounds of golf Trump	806
41	Survival Rates Age Latest survival rate age CDC Covid SURVIVAL Ages 20-49	438	4	Michigan Judge Kenny Detroit City Election full forensic audit orders HAND RECOUNT	814	50	IRS tax code Ted Cruz Bible words Americans spend	632
8	veto Dream Act dream act DREAM ACT Obama Romney	404	42	Hank Aaron Covid Vaccine heart attack DMX Aarons death	769	22	George floyd died Judge judy george Floyd drug overdose	586
16	CEO compensation Economic Policy Institute Typical worker CEOs	389	21	minimum wage buying power inflation years	766	62	espionage act Obama campaign spied Trump	463
25	heart attack Covid Vaccine DMX dmx died	386	11	Obama played rounds of golf Trump	701	72	Declaration of Independence James Madison sign	457
24	Keystone pipeline permanent jobs State Department	356	5	African American children african american babies NYC born black	5	19	Rick Scott stole VOTE RICK SCOTT Floridian Medicaid millions from Medicare	433
65	Declaration of Independence James Madison sign	349	41	Wind Generation States Clean Green Energy Keystone pipeline green energy jobs	689	66	Black voter turnout voters	421

TABLE 11: Number of Tweets and clusters with respect to keywords

Equation 3 shows our interpretation of social network score and its potential where $F2$ is the number of topically active followers, and $F4$ is the number of topically active followers.

$$InteractorRatio(i) = \frac{RT3 + M4}{F1} \tag{4}$$

$$InteractorRatio(i) = \frac{ReTweet_i + ReplTweet_i}{Tweet_i} \tag{5}$$

Where $RT3$ is the number of users who have retweeted the author's tweets.$M4$ is the number of users mentioning the author. $F1$ is the number of followers.

9.2.1 Proposed Influence Score

The Final influence score that was decided upon measuring the influence of the users within the Truthseeker dataset is described in 6.

$$IF(i) = FC \cdot log((\alpha \cdot LC(i)) + (\beta \cdot SC(i))) \tag{6}$$

Where IF is the influence score, FC is the followers count, SC is the statues count is the number of Tweets author have posted, LC is the listed count which is is the number of times a tweet from that user has been added to a list by another person Normalized Influence. we have normeliized the final score utelizing the followong. norScore(i) = tanh(log(IF(i) + 1))/100, where tanh is tangent function.

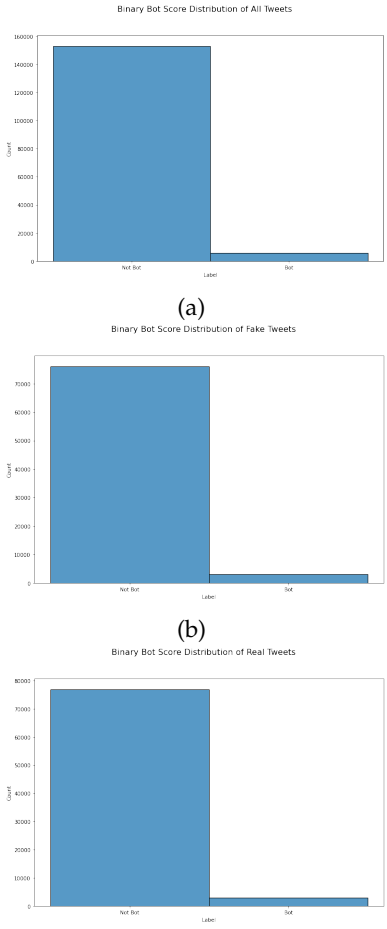


Fig. 6: 6a Bots vs. Non Bots in the whole TruthSeeker dataset, 6bBots vs. Non Bots in the Fake Tweets of the TruthSeeker dataset, 6cBots vs. Non Bots in the Real Tweets of the TruthSeeker dataset

The alpha used for our tests was 0.7 and the beta was 0.3. Weighing the fact that a user was added to a list a a sign that they are viewed a a trustworthy source of information. The graphs below showcase the results of the normalization of the data. As can be seen, the average influence of a user in our system if relatively low. Outliers with massive followings can be easily seen as well as inactive or bot accounts.

9.3 User Credibility

A user's Credibility Score is calculated using a simple (followers/(friends + followers)) equation. 7 shows the results of applying the equation to the full data set of both Real and Fake tweets statements. Showcasing the reality of the Twitter ecosystem as a whole. Figure 7aand 7b showcases that most users in our system have a middling level of influence. Most users have a fairly low impact on their environment and others around them. However, some outliers with a large amount of influence are able to disseminate information easily and widely.

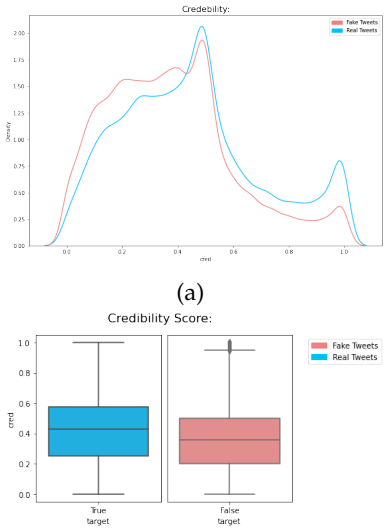


Fig. 7: 7aDensity Distribution of Credibility Score 7bCredibility Distribution Box Plot

10 Conclusion

The expansion of social media platforms, such as Twitter, created several unreliable and unverified platforms, which can create fake news and spread it through several users. Therefore, automatic detection of this misleading information on social media can become an endless challenge for researchers. Addressing this challenge is critical to prevent the spread of misinformation, which can cause significant harm, especially in times of crisis. One of the primary obstacles in detecting fake content on social media platforms is the vast volume of content to be evaluated manually. The massive volume of data demands utilizing different machine learning and deep learning algorithms for automating the progress. However, the success of such algorithms depends heavily on the quality of the dataset used for training.

The existing fake news datasets often need to be updated and expanded in scope. The TruthSeeker dataset can significantly contribute to fake news detection in social media. This dataset, which contains more than 180,000 labels from 2009 to 2022, was collected using Amazon Mechanical Turk, a crowdsourcing platform. The dataset was verified using a three-factor active learning verification method, ensuring its credibility and trust. The authors' institution employees further verified two and five-label classifications and 456 unique Amazon Mechanical Turk highly skilled individuals labeled each tweet three times. Moreover, the dataset contains binary and multi-class classifications, allowing for a more precise and nuanced analysis of tweet content.

To evaluate the accuracy of the detection models, the authors implemented various machine learning and deep learning algorithms, including multiple BERT-based models. The results demonstrated significant improvements in the ability to automatically detect fake content, even with the limited length of tweets. Additionally, the authors introduced three auxiliary social media scores: Bot, credibility, and influence score, to better understand the patterns and characteristics of Twitter users for fake/true tweets and their

impact on the content they post.Furthermore, the authors utilized clustering-based event detection to analyze the relationships between topics and Tweets, and the correlation between tweet labels and online creators/spreaders' characteristics. This analysis provided valuable insights that can help improve the precision and effectiveness of fake content detection models.

In conclusion, the TruthSeeker dataset significantly contributes to the field of fake news detection, specifically regarding Twitter. The TruthSeeker Dataset was a project undertaken by the Canadian Institute for Cybersecurity to determine the validity of tweets posted on Twitter in an automated way. All the data will be available on the dataset page of CIC https://www.unb.ca/cic/datasets/truthseeker-2023.html. The extensive collection of labels, rigorous verification methods, and focus on Twitter content make this dataset valuable for researchers in this area. Additionally, applying multiple BERT-based models and auxiliary social media scores, combined with clustering-based event detection, has provided valuable insights that can help address the long-standing challenge of automatically detecting fake content on social media platforms. While there are still challenges to be addressed, the TruthSeeker dataset has shown promise in advancing the field of fake news detection and is a vital step toward addressing the issue of automatically detecting misinformation on social media platforms.

REFERENCES

[1] S. Helmstetter and H. Paulheim, "Collecting a large scale dataset for classifying fake news tweets using weak supervision," *Future Internet*, vol. 13, no. 5, p. 114, 2021.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 274–277.

[4] T. Murayama, "Dataset of fake news detection and fact verification: A survey," *arXiv preprint arXiv:2111.03299*, 2021.

[5] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 2014, pp. 18–22.

[6] A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter, "Pheme dataset of rumours and non-rumours," 2016.

[7] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours," *arXiv preprint arXiv:1704.05972*, 2017.

[8] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning." Association for Computational Linguistics, 2017.

[9] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv preprint arXiv:1712.07709*, vol. 8, 2017.

[10] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," *arXiv preprint arXiv:1806.03713*, 2018.

[11] 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.

[12] N. T. Tam, M. Weidlich, B. Zheng, H. Yin, N. Q. V. Hung, and B. Stantic, "From anomaly detection to rumour detection using data streams of social platforms," *Proceedings of the VLDB Endowment*, vol. 12, no. 9, pp. 1016–1029, 2019.

[13] N. Vo and K. Lee, "Where are the facts? searching for fact-checked information to alleviate the spread of fake news," *arXiv preprint arXiv:2010.03159*, 2020.

[14] E. Dai, Y. Sun, and S. Wang, "Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 853–862.

[15] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[16] A. Dharawat, I. Lourentzou, A. Morales, and C. Zhai, "Drink bleach or do what now? covid-hera: A study of risk-informed health decision making in the presence of covid-19 misinformation," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 1218–1227.

[17] Y. Li, B. Jiang, K. Shu, and H. Liu, "Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation," *arXiv preprint arXiv:2011.04088*, 2020.

[18] D. Kar, M. Bhardwaj, S. Samanta, and A. P. Azad, "No rumours please! a multi-indic-lingual approach for covid fake-tweet detection," in *2021 Grace Hopper Celebration India (GHCI)*. IEEE, 2021, pp. 1–5.

[19] C. Santana, D. B. Claro, and M. Souza, "Fake news detection in tweets: Challenges and adaptations imposed by the covid-19," *iSys-Brazilian Journal of Information Systems*, vol. 15, no. 1, pp. 11–1, 2022.

[20] S. Dadkhah, F. Shoeleh, M. M. Yadollahi, X. Zhang, and A. A. Ghorbani, "A real-time hostile activities analyses and detection system," *Applied Soft Computing*, vol. 104, p. 107175, 2021.

[21] H. Kirn, M. Anwar, A. Sadiq, H. M. Zeeshan, I. Mehmood, and R. A. Butt, "Deepfake tweets detection using deep learning algorithms," *Engineering Proceedings*, vol. 20, no. 1, p. 2, 2022.

[22] A. Zhang, A. Brookhouse, D. Hammer, F. Spezzano, and L. Babinkostova, "Predicting the influence of fake and real news spreaders (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 13 107–13 108.

[23] 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," 2016.

[24] 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, no. 3, p. e0150989, 2016.

[25] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski, "Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*. Association for Computational Linguistics, 2019, pp. 845–854.

[26] V.-H. Nguyen, K. Sugiyama, P. Nakov, and M.-Y. Kan, "Fang: Leveraging social context for fake news detection using graph representation," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 1165–1174.

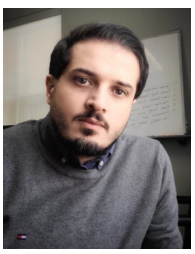
[27] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection," *arXiv preprint arXiv:2010.08768*, 2020.

[28] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer, 2021, pp. 21–29.

[29] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish *et al.*, "Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 913–922.

[30] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, "A covid-19 rumor dataset," *Frontiers in Psychology*, vol. 12, p. 644801, 2021.

[31] F. Riquelme, P. Gonzalez-Cantergiani, D. Hans, R. Villarroel, and R. Munoz, "Identifying opinion leaders on social networks through milestones definition," *IEEE Access*, vol. 7, pp. 75 670–75 677, 2019.



Sajjad Dadkhah is a Faculty member, cybersecurity team leader and Research Associate, at the Canadian Institute of Cybersecurity (CIC), University of New Brunswick (UNB). He has over ten years of experience in digital multimedia security, computer security, and machine learning-based detection systems. He has been involved in several security projects as a team leader, researcher, and security consultant in different organizations such as Kyushu University (Japan), Universiti Malaya (UM), IRIS Smart Technology Complex, and Kyushu Institute of Technology (Japan). In Sept 2016, he was awarded a fellowship by the Kyushu Institute of Technology (Japan) to continue his research for two years. He has been the Managing editor and Board member of Applied Soft Computing (ASOC) Elsevier journal since 2016.

Xichen Zhang received the B.E. degree from Changsha University of Science and Technology in 2010. He received his M.S. degree in Computer Science at the Canadian Institute for Cybersecurity (CIC), Faculty of Computer Science (FCS), University of New Brunswick (UNB) in 2018. After that, he worked as a research assistant in CIC. He is currently working toward the Ph.D. degree with FCS, UNB. His research interests are data mining in cybersecurity, privacy enhancing technologies, and IoT-Big Data security and privacy.

Alexander Gerald Weismann is currently a 4th year undergraduate student in Computer Science and Media Arts and Culture (concurrent) at the University of New Brunswick. He worked as a security software developer at the Canadian Institute for Cybersecurity. His research areas focus on neural network development, Twitter dataset creation, and fake news detection with multi-modal analyses.

Amir Firouzi is currently a Ph.D student at the University of New Brunswick, Canada, and a researcher at the Canadian Institute for Cybersecurity. He received his Masters degree in Software Engineering from the Department of Computer Engineering at Ferdowsi University of Mashhad, Iran. His research interests include Machine Learning, NLP, IoT, IoT Security, Big Data, Data Engineering.

Ali A. Ghorbani has held a variety of academic positions for the past 39 years and is currently a Professor of Computer Science, Tier 1 Canada Research Chair in Cybersecurity, and Director of the Canadian Institute for Cybersecurity, which he established in 2016. He served as the Dean of the Faculty of Computer Science at the University of New Brunswick from 2008 to 2017. He is also the founding Director of the laboratory for intelligence and adaptive systems research. He has spent over 29 years of his 39-year academic career, carrying out fundamental and applied research in machine learning, cybersecurity, and Critical Infrastructure Protection. He is the co-inventor on three awarded and one filed patent in the fields of Cybersecurity and Web Intelligence and has published over 280 peer-reviewed articles during his career. He has supervised over 190 research associates, postdoctoral fellows and students during his career. His book, Intrusion Detection and Prevention Systems: Concepts and Techniques, was published by Springer in October 2010. Dr. Ghorbani developed several technologies adopted by high-tech companies and co-founded three startups, Sentrant Security, EyesOver Technologies, and Cydarien Security in 2013, 2015, and 2019. He is the co-founder of the Privacy, Security, Trust (PST) Network in Canada and its annual international conference and served as the co-Editor-In-Chief of Computational Intelligence: An International Journal from 2007 to 2017. Dr. Ghorbani is the recipient of the 2017 Startup Canada Senior Entrepreneur Award, and Canadian Immigrant Magazine's RBC top 25 Canadian immigrants of 2019.