Domain Bias in Fake News Datasets Consisting of Fake and Real News Pairs

Kato, Shingo
Graduate School of Information Science and Electrical Engineering, Kyushu University

Yang, Linshuo
Graduate School of Information Science and Electrical Engineering, Kyushu University

Ikeda, Daisuke
Faculty of Information Science and Electrical Engineering, Kyushu University

出版情報：2022 12th International Congress on Advanced Applied Informatics IIAI-AAI 2022 Proceedings, pp.101-106, 2022-07. IEEE
バージョン：
権利関係：

Domain Bias in Fake News Datasets Consisting of Fake and Real News Pairs

Shingo Kato
Graduate School of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
kato.shingo.963@s.kyushu-u.ac.jp

Linshuo Yang
Graduate School of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
yang.linshuo.096@s.kyushu-u.ac.jp

Daisuke Ikeda
Faculty of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
ikeda.daisuke.899@m.kyushu-u.ac.jp

***Abstract*—News intentionally containing false information–known as "fake news"–is common on the Internet and often causes social disruption. In order to solve it, research on automatic detection of fake news using supervised learning has been active. Although the accuracy is improving, a major challenge for practical application remains: models can not work well for news in unknown fields (domains) due to domain biases. The goal of this study is to mitigate these domain biases and improve the accuracy of cross-domain fake news detection, which tests news from unknown domains. We firstly try to mitigate the bias by masking noun phrases which are considered a major source of domain bias. However, masking has not improved accuracy. Therefore, we point out that the dataset in this study has the property that it always contains pairs of fake and real news on the exact same topic. In this paper, we focus on this property of dataset and examine how it may affect domain bias and accuracy. Comparative experiments show that accuracy is higher when trained on a dataset with the property shown in this study. We suggest that a fake news dataset consisting of paired news could be effective for cross-domain detection.**

***Index Terms*—fake news detection, cross-domain, BERT**

I. Introduction

Thanks to the widespread use of the Internet, we can easily gather information. However, some information on the Internet is incorrect. Fake news is one of such information. There are various definitions for fake news, but all of them can be described as news disseminating misinformation for some purpose. For example, Zhou et al. defined fake news as "intentionally false news published by a news outlet" [1].

Fake news has been a serious issue around the world. For example, during the 2016 U.S. presidential election, many fake news was spread and it is even said that fake news changed the result of the election [2]. The need for fake news detection has been recognized. Determining the veracity of such news generally requires prior knowledge of the news and a lot of cost to verify the information. The need for automatic detection of fake news is increasing because of the large amount of fake news being disseminated on social networking sites.

There are two major approaches to the fake news detection task: knowledge-based approach and feature-based one. In the former case, a technique called fact-checking is often used. In the latter one, detection is based on capturing unique characteristics of fake news. Supervised learning is often used in this approach.

Feature-based detection have improved accuracy due to large-scale pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) [3]. Although it has been detected with high accuracy in experiments, there is a significant challenge for its practical application, that is, fake news detection heavily depends on the genre (domain) of news in the training data and can not work well for news in unknown domains [4]. In other words, standard models for fake news detection are overfited to given domains. This is mainly due to *domain bias* caused by differences in vocabulary among domains. For instance, a detection model that judges news with many "Donald Trump" words as fake is not effective for news in the sports domain. The simplest way to mitigate overfitting is to create a dataset containing news from as many domains as possible. However given the high cost of creating fake news datasets, *cross-domain fake news detection*–which can detect even unknown domains–is important.

The goal of this study is to mitigate these domain biases and improve the accuracy of cross-domain fake news detection. At first, we try to mitigate the bias by masking noun phrases which are considered a major source of domain bias. However, masking has not improved accuracy, and it is likely not an effective bias mitigation method for the dataset used in this study.

Therefore, we point out that the dataset in this study has the property that it always contains pairs of fake and real news on the exact same topic. The authors consider the possibility that this property may have some effect on domain bias. In this paper, we focus on this property of dataset and examine how it may affect domain bias and accuracy. To the best of the authors' knowledge, no previous studies have examined the effect of such properties of the dataset on the accuracy of cross-domain detection.

We use the dataset FakeNewsAMT [5] and employ BERT as the classification model. As a prior experiment, we conduct a cross-domain fake news detection experiment by masking noun phrases that are considered to be a source of domain bias. Next, we conduct a comparison experiment to determine whether the property of FakeNewsAMT affects its accuracy,

and show that models trained on a dataset that satisfies this property have better accuracy in cross-domain detection.

II. Related Work

In this section, we briefly show three previous studies relevant for cross-domain detection using stylistic features of fake news. Firstly, we present a study in which stylistic features were manually extracted and were classified using SVM. We show that there is some stylistic feature difference between fake news and real news, and that these can be captured and classified. Secondly, we introduce a study that used deep learning to conduct a fake news cross-domain detection experiment. Unlike the first study, this study uses deep learning to capture more latent features and achieve higher accuracy in cross-domain detection. Finally, as a clue to the methods of domain bias mitigation, we present a study that proposed a bias mitigation method between datasets.

A. Stylistic Characteristics of Fake News

Given a text of news, what are characteristics that we intuitively find suspicious? The study by Benjamin et al. [6] statistically tested for differences between authentic and fake news by extracting three categories of features: stylistic features, complexity features, and psychological features. The results show that characteristics differed significantly. Fake news have characteristics such as less jargon, more lexical redundancy, and more self-referential (e.g. "I", "We" are used more often). Classification using SVM with these features result in over 70% accuracy, well above the baseline of 50%. From this results, we can conclude that fake news can be detected by capturing features.

B. Cross-Domain Fake News Detection Using Deep Learning

While the research presented in Section II-A extracted features manually, it is expected that deep learning can be used to obtain more latent features automatically. Saikh et al. attempted to improve the accuracy of fake news detection using deep learning [7]. They also used a dataset called FakeNewsAMT, which contains six domains, to test the cross-domain detection accuracy. In the experiment, five domains were used for training, and data from the remaining one domain were classified as fake or non-fake using a neural network. The results show a relatively high accuracy of 73-91% for classification.

ELMo (Embeddings from Language Models) [8] was used in this study to vectorize words. Although ELMo is a context-aware embedded representation compared to traditional Word2Vec and others, it is a shallow contextual consideration compared to pre-trained models on large datasets such as BERT. In addition, factors that enabled highly accurate detection of cross-domain fake news–which is considered to be difficult–had not been verified. We consider that the method used to create FakeNewsAMT may be a contributing factor.

FakeNewsAMT is a dataset of news consisting of a title, contents, and a label. The dataset contains 40 news in each of six domains (business, education, politics, entertainment, sports, and technology) verified as factual, and then crowdsourced to create fake news based on each factual news, giving instructions to write the news in a journalistic style and avoid unrealistic content. Due to the instructions, the dataset can resemble actual fake news. We focus on the nature of FakeNewsAMT due to the method of its creation and examine its impact on cross-domain detection in Section III-D.

C. Bias in Fake News Detection

The distribution of words used in each domain is different, which prevents generalization to unknown domains. That is, the domain bias in the training dataset makes cross-domain fake news detection difficult. Also, fake news is a type of news and is therefore influenced by trends and interests. Therefore, the data collected varies greatly depending on when the dataset was created. Murayama et al. named this "diachronic bias" [9], and noted bias among the fake news datasets.

Assuming that this bias is mainly caused by named entities such as person names, Murayama et al. attempted to improve the classification accuracy for datasets created at different times by masking these named entities. The results show improved accuracy, suggesting that masking named entities mitigated the bias between datasets and let the model more generalizable to unknown datasets.

Since there are biases among the six domains included in the FakeNewsAMT, used in this study, due to vocabulary and other factors, we test whether the method of Murayama et al. can be applied to reduce the domain bias.

III. Experiment

In this section, we conduct cross-domain fake news detection experiments on the six domains of FakeNewsAMT by using BERT. We use the similar method as Murayama et al. to try to mitigate the bias between domains, and verify whether there is a change in accuracy compared to training with normal data. Also, we quantitatively examine properties of the dataset, and test the impact of these properties on the classification model.

A. Data and Preprocessing

We use FakeNewsAMT in this paper, and it is composed of news data consisting of a title and body. Here is an exmple of a news in the dataset:

```
Robots Taking Over the World

Robots are slowly taking over the
workforce of the world. Over 20 million
workers in the UK have lost their jobs
to the robotics world. The consultancy
Firm PwC has found...
```

The first sentence is the title of the news, and the next block is the body. Each domain of FakeNewsAMT is composed 40 fake news and the same number of real news. The basic statistics are shown in Table I.

As preprocessing, the publication date and time of the news and the URL were removed. In addition, there are several news

articles without their titles. To put these data into BERT, "No title" was added to the title of the data.

B. Bias Mitigation Experiments with Masking

In this section, we employ BERT as a pre-trained classification model and conduct the cross-domain detection using FakeNewsAMT, according to Saikh et al. The authors also consider that noun phrases vary widely in distribution across domains, and they can be one of the factors contributing to domain bias. We try to mitigate the bias by masking noun phrases and compare the cross-domain detection accuracy when using each normal and masked data.

POS (Part-of-speech) tagging was used before masking noun phrases in this experiment. It is the task of estimating the part-of-speech of words in a sentence. Tokens estimated as proper nouns were replaced with [NNP] and those estimated as nouns with [NN] labels. An example of the masking results for the actual data is shown below.

```
Original data:
Trump's next legislative target:tax reform
Masked data:
[NNP]'s next legislative [NN]:[NN] [NN]
```

We used flair[1], a Python framework, to do the POS tagging.

In this study, we employ BERT as a fake news detection model. BERT can be used for variety of tasks such as classification problems and sentence generation. Also, fine-tuning BERT–which has been trained on a large dataset–can perform well on a small dataset.

BERT is given two sentences or one sentence as input, where the input format is "[CLS] 1st-sentence [SEP] 2nd-sentence [SEP]", where [CLS] is a special token indicating the beginning of a sentence, and [SEP] is a special token indicating the end of a sentence. The embedding of [CLS] tokens is sometimes used in classification problems. The input to BERT is the sum of the embedded representation of the word, the representation indicating whether it is the first or second sentence, and the representation with positional information.

We use a pre-trained model published on HuggingFace[2], and denote it by $\mathrm{BERT}_{\mathrm{BASE}}$. The title and body of a news are given as two input sentences. An overview diagram of the model is shown in Fig. 1. The embedding of [CLS] tokens in the final layer of BERT ($T_{[CLS]}$) is given as input to the fully connected layer (FFN). In the output layer of the FFN, the Softmax function is used for binary classification as fake or non-fake. The number of neurons in each FFN layer is 768, 10, and 2. AdamW is used as Optimizer and the learning rate is set to 1e-05. To prevent overfitting, we drop out 20% of the output of the input layer. The special tokens for masking, [NN] and [NNP], are added as tokens for BERT and we train the FFN layer and fine-tuning BERT.

Four of the six domains in the FakeNewsAMT are used for training, validation is performed in another domain, and the remaining one is tested with the least lossy parameter in the validation domain. There would be five models for one testing domain. Fig. 2 shows an example of how to split the data for training when the business is the testing domain.

Python 3.8.10 and AllenNLP 2.8.0[3], a framework for natural language processing, are used in this experiment. The OS is Ubuntu 20.04 LTS, the CPU is AMD Ryzen 7 3700x (3.6GHz), and the GPU is NVIDIA GeForce RTX 3080 (10GB).

These tables show the accuracy when training with normal data (Table II) and with noun-masked data (Table III). The rows represent the test domain and the columns represent the validation domain, and the value of each cell is the accuracy when tested with the classification model trained in the four training domains. The rightmost column is the average accuracy for each test domain.

Firstly, looking at the results for each domain, the average accuracy increases in the politics and entertainment domains due to masking. On the other hand, the average accuracy decreases for the business, education, and technology domains, and remain the same for the sports domain. Looking at the overall results, the average accuracy for the six domains is 0.815 for the normal data and 0.801 for the masked data. Masking result in a 1.4% decrease in accuracy.

As a result, masking noun phrases does not improve accuracy. For FakeNewsAMT, it is likely that masking noun

TABLE I: FakeNewsAMT statistics: average number of words and sentences per news

label	No. of news	Avg. words	Avg. sentences
Fake	240	132	5
Legit	240	139	5

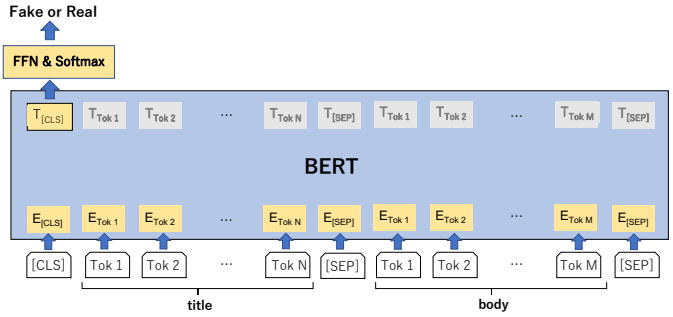

Fig. 1: Figure of classification model in this study

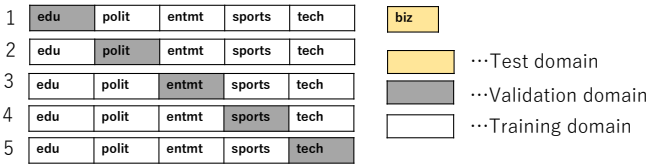

Fig. 2: How to split the training data. In line 1, edu is used for validation and others are used for training.

[1]https://huggingface.co/flair

[2]https://huggingface.co/models

[3]https://allenai.org/allennlp

TABLE II: Accuracy when training with normal data

Test \Validation	biz	edu	polit	entmt	sports	tech	**Average**
biz		0.89	0.93	0.85	0.86	0.90	**0.886**
edu	0.84		0.84	0.80	0.85	0.80	**0.826**
polit	0.82	0.82		0.79	0.80	0.84	**0.814**
entmt	0.76	0.78	0.70		0.68	0.79	**0.742**
sports	0.86	0.85	0.74	0.79		0.86	**0.820**
tech	0.88	0.70	0.84	0.78	0.82		**0.804**

TABLE III: Accuracy when training with masked data

Test \Validation	biz	edu	polit	entmt	sports	tech	**Average**
biz		0.89	0.89	0.85	0.84	0.84	**0.862**
edu	0.82		0.79	0.79	0.86	0.72	**0.796**
polit	0.88	0.79		0.84	0.84	0.80	**0.830**
entmt	0.76	0.74	0.78		0.70	0.75	**0.746**
sports	0.85	0.81	0.79	0.81		0.84	**0.820**
tech	0.78	0.72	0.76	0.76	0.75		**0.754**

phrases has little effect on domain bias. However, it is noteworthy that while cross-domain fake news detection is considered difficult, the accuracy when training on regular data is very high for some domains, as in the experiment by Saikh et al.

This result suggests that the classification model may have been unaffected by bias for some reason, i.e., the noun phrases, a major source of bias, may not have affected the detection. The fact that no significant difference in accuracy is observed despite the masking of noun phrases also suggests this possibility. We examine this possibility in detail in the next section.

C. Lexical Overlap between Paired Data

In this section, we quantitatively analyze properties of FakeNewsAMT to determine if there are any factors that may contribute to the results in the Section III-B.

FakeNewsAMT consists of correct news and crowd-sourced fake news based on the correct news. In other words, this dataset always contains pairs of fake and real news on the same topic. At this time, it can be assumed that news on the same topic have similar noun phrases, and in fact, FakeNewsAMT shows overlapping noun phrases between the paired news data. An example is shown in Fig. 3. It can be seen that the noun phrases are somewhat similar between the fake and real news pair. We test whether overlap of noun phrases between paired data is found across the entire dataset.

Before calculating the noun phrase overlap rate between the paired data, we preprocessed the data using the method of Juan et al. [10]. Firstly, all letters were converted to lowercase. Secondly, frequently used words such as "I", "a", and "of", called stop words, were removed. Finally, lemmatization was performed to convert words into headwords. For example, "dogs" is converted to "dog", and "met" to "meet".

The overlapping noun phrases shown in blue in Fig. 3 are often unique to the news, such as proper nouns. We suspect that noun phrases unique to that news may have more overlap between the paired data, so in addition to calculating the overlap rate for all noun phrases, we also calculate the overlap rate for characteristic noun phrases. We use the TF-IDF method to check whether the noun phrases are news-specific or not. The definition and calculation method of TF-IDF are described below.

The TF value of a word t in a document d is defined as

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}, \tag{1}$$

where $n_{t,d}$ is the frequency of a word t in document d and $\sum_{s \in d} n_{s,d}$ is the sum of the frequencies of all words in document d. In this experiment, a document d refers to a news article.

The IDF value for the word t is defined as

$$idf(t) = \log \frac{N}{df(t)} + 1, \tag{2}$$

where N is the number of all documents and $df(t)$ is the number of documents in which the word t appears. In this experiment, all documents refers to 480 news data including fake and real news.

Finally, the product of TF and IDF is TF-IDF. Document-specific words are assigned a higher TF-IDF value.

Lexical overlap rates are calculated between all paired data of fake and real news. The results are shown in Table IV. In all domains, the overlap rate for noun phrases only is higher than that for the whole vocabulary. We also calculate the overlap rate for noun phrases with the top 20 TF-IDF values, which is higher in most domains. The top TF-IDF words include many words that could be a major sources of domain bias, such as named entity and nouns specific to that domain. The results show that many of these words overlapped between pairs of data.

There is a lot of overlap of noun phrases between pairs of FakeNewsAMT data. The similarity of noun phrases between the fake and non-fake data suggests that the model may have learned to make judgments without noun phrases. It is very important for cross-domain fake news detection that the model is not influenced by noun phrases, which can be a source of domain bias.

D. The Effect of Training Data Properties on Accuracy

In this section, we examine whether training on paired data with overlapping noun phrases affects domain bias and improves the accuracy of cross-domain fake news detection. We created training data consisting of paired data only and, conversely, training data consisting of unpaired data. We

TABLE IV: Average percentage of lexical overlap between paired data

	Whole vocabulary	**Noun only**	**TF-IDF Top 20 Noun**
biz	0.300	0.367	0.397
edu	0.247	0.308	0.391
polit	0.343	0.409	0.402
entmt	0.238	0.319	0.405
sports	0.278	0.340	0.343
tech	0.210	0.279	0.408

"Alex Jones , **purveyor** of the independent investigative news website Infowars and **host** of The Alex Jones Show , has been vindicated in his **claims** regarding the so-called "Pizzagate" **controversy** . Jones and **others** uncovered **evidence** last **year** that top Democratic **Party officials** were involved in a bizarre , satanic **child sex cult** and **pornography ring** using the Washington D .C . pizza **parlor** Comet Ping Pong Pizza as a **front** . The **allegations** rocked the Democratic **Party** and may have caused serious **damage** to the **Hillary Clinton** presidential **campaign .** Top **U.S.** federal **investigators** have now confirmed that they have verified many of these **claims** after executing raids on the **offices** of several of the key **players** . **Charges** are expected to be filed in the coming **days** .

(a) Fake news

Alex Jones a prominent **conspiracy theorist** and the **host** of a popular right-wing **radio show** has apologized for helping to spread and promote the **hoax** known as Pizzagate . The **admission** on **Friday** by Mr . Jones the **host** of **"The Alex Jones Show"** and the **operator** of the website Infowars was striking . The Pizzagate **theory** which posited with no **evidence** that top Democratic **officials** were involved with a satanic **child pornography ring** centered around Comet Ping Pong a pizza **restaurant** in Washington D .C . grew in online **forums** before making its **way** to more visible **venues** including Mr . Jones's **show.**

(b) Real news

Fig. 3: Paired news data
(Bold : Noun phrase, Blue : Overlapping noun phrase)

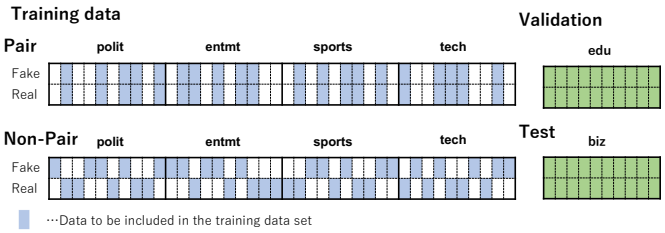

Fig. 4: An example of how to create a data set

evaluate the accuracy on the test domain of models trained on these datasets.

The paired dataset consists of 80 randomly selected fake and non-fake pairs of data from each of the four training domains. Conversely, the unpaired dataset consists of 160 randomly selected (paired data not included) data from each domain (Fig. 4). The amount of data for both datasets is 160, and 10 training datasets were created for each dataset.

The same classification model is used as in the Section III-B. Table V shows the results when training on the dataset constructed with only paired data or without paired data. The results show that the average accuracy is about 5% to 7% higher when training on paired data. However, although the average accuracy differs in the sports domain, there is only a littel difference in accuracy, except in some cases where accuracy rates are extremely low.

Therefore, to check whether there is a significant difference in the accuracy of the two groups, we use Mann-Whitney's U test.

The results show differences at the 5% significance level in the business, education, and entertainment domains, and at the 1% significance level in the technology domain. No significant differences are found in the politics and sports domains in this experimental data.

There are significant differences in four of the six domains, suggesting the possibility of improving accuracy for unknown domains by composing the dataset with paired data. Although we were not able to conduct a quantitative analysis of the causes, the sports domain contains several fake news that seem to reverse the wins and losses of games, suggesting that it is quite difficult to judge if such news is fake or not, and some unique nature of fake news in the sports domain may have influenced the results.

IV. Conclusion

In this study, we have validated the property of the dataset for cross-domain detection. First, we have attempted to mitigate the bias by training on data with masked noun phrases. As a result, we have not seen any improvement in accuracy due to masking, but have noted that the accuracy is very high when learning with normal data. At this time, we have focused on the property of FakeNewsAMT that it consists of pairs of fake and real news about the exact same topic, and that noun phrases overlap between the paired data. We have hypothesized that due to this property, the classification model would learn to ignore noun phrases to determine whether they are fake or not, and have been less susceptible to domain bias. To test this hypothesis, we created a dataset consisting of only paired data and only unpaired data, and have compared the accuracy of the classification models when trained on each dataset. The results have shown that training on the paired dataset is more accurate, with four of the six domains showing higher accuracy at the 5% level of significance rather than differences due to randomness.

We have shown that it may be important to collect pairs of real and fake news with similar noun phrases on the same topic in order to improve the accuracy of cross-domain detection.

There are two future issues in this study. The first is how to create the actual dataset. The FakeNewsAMT uses crowd-

TABLE V: Average accuracy for each domain when training on a dataset built with corresponding fake and real news paired data or without paired data

Training Data	Test	1	2	3	4	5	6	7	8	9	10	Average
Paired Data	**biz**	0.78	0.85	0.81	0.79	0.82	0.81	0.80	0.86	0.81	0.84	**0.817**
	edu	0.70	0.69	0.71	0.71	0.69	0.68	0.68	0.62	0.68	0.69	**0.685**
	polit	0.74	0.68	0.80	0.78	0.66	0.71	0.75	0.79	0.70	0.78	**0.739**
	entmt	0.64	0.69	0.65	0.66	0.74	0.65	0.70	0.68	0.66	0.66	**0.673**
	sports	0.78	0.89	0.82	0.76	0.89	0.80	0.84	0.84	0.79	0.81	**0.822**
	tech	0.79	0.75	0.75	0.78	0.75	0.79	0.79	0.81	0.75	0.70	**0.766**
Non-Paired Data	**biz**	0.81	0.82	0.79	0.61	0.80	0.72	0.68	0.85	0.71	0.66	**0.745**
	edu	0.56	0.72	0.65	0.66	0.69	0.68	0.60	0.57	0.59	0.64	**0.636**
	polit	0.72	0.66	0.56	0.69	0.75	0.76	0.68	0.75	0.59	0.68	**0.684**
	entmt	0.53	0.69	0.62	0.64	0.62	0.57	0.60	0.65	0.69	0.62	**0.623**
	sports	0.80	0.84	0.78	0.80	0.61	0.86	0.85	0.78	0.61	0.80	**0.773**
	tech	0.70	0.78	0.76	0.71	0.70	0.70	0.71	0.57	0.72	0.68	**0.703**

TABLE VI: The result of Mann-Whitney U test

Test domain	p-value
biz	**0.045**
edu	**0.028**
polit	**0.076**
entmt	**0.014**
sports	**0.307**
tech	**0.008**

sourcing to create fake news. When constructing a dataset from news and actual fake news, it is necessary to devise a method to collect paired news data. Next, it is necessary to quantitatively analyze why few differences are found only in the sports domain. If the nature of fake news specific to the sports domain is clarified, it may provide useful information for conducting research on fake news detection.

REFERENCES

[1] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5), 2020.

[2] Washington Post. A new study suggests fake news might have won donald trump the 2016 election. https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, NAACL-HLT, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[4] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 557–565. AAAI Press, 2021.

[5] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, 2018.

[6] Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):759–766, 2017.

[7] Tanik Saikh, Arkadipta De, Asif Ekbal, and Pushpak Bhattacharyya. A deep learning approach for automatic detection of fake news. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 230–238, 2019.

[8] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[9] Murayama Taichi, Wakayama Shoko, and Aramaki Eiji. Diachronic bias in fake news detection datasets (in Japanese). *Proceedings of the Twenty-seventh Annual Meeting of the Association for Natural Language Processing*, pages 1011–1016, 2021.

[10] Juan Pablo Posadas-Durán, Helena Gomez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent and Fuzzy Systems*, 36:4868–4876, 2019.