Detected Source-Based fake News via Word2vec Algorithm

Hamid Sharifi[1] , Jafar Sheykhzadeh [2]
[1,2] Department of Computer Science, Ahar Branch, Islamic Azad University, Ahar, Iran
[1] Email: HamidSharifi22@yahoo.com
[2] Email: j-sheykhzadeh@iau-ahar.ac.ir (Corresponding Author)

Abstract

Today and regarding the increase of social media platforms and the people welcoming these networks has led to share different data throughout the world without the confirmation by the platforms. This has increased the incorrect data frequency and has had great effects on political, economic, and social fields. Such incorrect data are called fake news. This has changed into one of the topical issues in today's society. Through the proposal of an appropriate solution and first through analyzing the news resources in the dataset called Buzzfeed News, we have concluded that websites with better fames propagate less fake news. We changed the data into vector using Word2vec and investigated the similarity of the taught data and the tagged data in the dataset and got the least precision amounting to 0.60 and the highest precision amounting to 0.94 out of 1 and the results showed that our algorithm has been very helpful in discovering the qualified news.

Keywords: Word2vec, fake news detection, machine learning, Data Mining

1. Introduction

The previous decade has been full of growth and rapid success of online social networks. These networks have disturbed traditional media regarding fundamental changes in time and location of the propagate of the latest news. Social media prepare proper tools for the users to create and share varied data. Due to the vast content and ease of access, more and more people are searching and receiving news data online. In a global report in 2021, 60% of the world's population is online [1] for example, about 68% of adults in the United States have followed news in social media in year 2018[2] and this has increased tremendously compared with the 49% in year 2012. Along with the advantages of social media, they have created some dangers such as propagating incorrect data[3][4]. The incorrect data are a type of exciting reports or imperial news propagating comprised of intended strategies. Public thought disturbance is one of the biggest threats for business, news agencies, and democracy and it has violated many and has created lots of damages. On the whole, the creators of fake news aiming at political or financial success create and send such news in online social media. Following that, the employ social robots or spam senders to make money[5]. The effect of fake news was globalized in 21st. century. Besides all this, fake news was criticized because of the spread of political polarizing during the political decision making campaigns. The voters can also be affected by improper effects or political fraudulent orations. In the year 2016, the term fake news was proposed and it meant the highest possible misuse of web-based networks. It also was more publicly spread in 2016 after the critical political competitions in the United States [6]. In recent research literature, different approaches have been utilized to recognize and reduce incorrect data. Although these approaches are different considering the selection of algorithm techniques and their adaptation, they have common methodology and settlement techniques and mostly try to discover news using content[7]–[9] and grammar[10]–[12] comprehension. Researchers have dealt with fake news within a research framework but they have not been able to reach a complete approval resolution. The different viewpoints used to study and probe fake news are categorized as follows:

Style-based: it focuses on the writing style of the news to recognize it being correct or incorrect[11].

Propagation-based: as it is apparent through the term it is related to propagation. Thus, it focuses on incorrect data propagation[13], [14].

User-based: this outlook deals with the users. For example, how do people deal with fake news?

Practically, the approaches above are inconsistent because regarding the dynamic nature of news, the coverage of novel events, topics, and speeches permanently change. Therefore, the categorization using the content of the articles published within a certain time interval [15] may probably become inefficient in future as demonstrated in their experiments, which may be too late. According to research [16]

Our proposed method can perform a very in time and rapid recognition through understanding the validity of news propagation domain. We estimated the Alexa rank of fake news using the url of the broadcaster and showed that websites which are more famous propagate less fake news. We taught the documents gained using word2vec[17] algorithm and changed them into a vector and finally investigated about the similarity of unreal tagged news using the teaching set in the presence of tf-idf [18] similarity and gained the set with great care and precision.

On the whole, there would be 5 overall sections here: the data and familiarity with topic will be gained through investigating traditional approaches. Some current frameworks welcomed by the researchers will be used to recognize fake news and then the predicted framework will be explained. The implementation using the proposed algorithm and the analysis of the experimental results will be presented next. After that, the limitations and suggestions for future works will be supplied.

The above challenges raise many research questions and the complexity of the problem requires new and solid solutions.

This work is motivated by the following two important research questions:

First question: Is there a source that shows the credibility of the news publisher?

Second question: Is the source-based feature a fast and reliable method?

In response to the above questions, we use the source-based method to check the validity of the sources and then evaluate them and check the ready-made results and their similarity with the real data set, which is briefly as follows:

2. Research literature

Fake news is information that is false or misleading and is presented as real news[19]. Nikiforos et al [20] also focus on fake news detection on tweets related to the Hong Kong protests, using similar data collection methods, but different tweet filtering criteria, discarding any non-English ones, and feature selection methodologies, as they also include network and account related features. As a consequence, their dataset is smaller and more unevenly distributed, which is balanced out using SMOTE oversampling [21], with their results yielding 99.8% accuracy.

Recently Sonia and et al [22] proposed a topical assortment strategy (TAG) which uses linguistic and web marking to recognize the pages that entail fake news. They showed that their approach can categorize not only political news changing through the pass of time, but news in different areas could be covered. They perform better than content based approaches while they entail very few characteristics and do not require retraining.

Mangel and et al [23] utilized image and text as the input. They send images and texts to find related connections to the search engines and use Web scraper to recognize the first 20 connections and investigated the weight of the resources using tf-idf to recognize whether news is real or unreal.

Vishvakarma and et al determined an algorithm which approves the correctness of image content through the web search. Then it studies the validity of 15 outstanding google search results through calculating reality parameter (Rp). If it exceeds the threshold, it is considered as a category. We then calculate recognition precision in order to test the proposed approach performance and compare the highest precision with similar advanced models to represent the better performance of our approach.

Paschalides and et al [24] presented a new low volume plug in browser called Check-It which recognizes fake news in several stages. It works through the integration of different methods using Flag-list and Fact-Check and investigating the users' behavior and utilizing the linguistic approach. In fact, Check-It has gained an integration of knowledge excavated through different signals with a precision of higher than 90%.

3. Our ProposedApproach

There are lots of approaches to recognize fake news proposed by many researchers and our proposed method has alleviated their major defects. In our proposed method, the recognition of fake news is done through trained news recognition and comparing the news with unreal tagged news to calculate the precision of our training. Through the extraction of the characteristics using site_url we could calculate domain validity, domain age, and those deleted from the domain. Then we added another characteristic to our dataset called Alexa [25]. Alexa is a very popular and valid site to investigate the validity and popularity of a domain. (Ranking in this terminal starts from 0. This means that any site close to 0 has a higher validity). Then the data related to domains of each of the sites are added in Alexa field. Now we need to calculate the ranks and categorize real and fake news based on it. In the first stage, there is a need to add information as an input into the set to check whether the news is real or not. One of the methods previously used to achieve the important words using TF-IDF algorithm and was very common utilized the words gained to enter into their training set. The major problem with this method was that maybe the word we intend to recognize whether it is real or not may not be among TF-IDF algorithm's most frequent words. For this reason, we first enter a word as an input to discover the category and then extract tables or articles through which the text characteristic related to the word could be seen. Therefore, we have an extracted set of data from our dataset including our input data and then compare the extracted set using Alexa isolation characteristic considering real or unreal news. Finally, we use word2vec algorithm to achieve the similarity between them. The overall method has been represented in figure.

3-1: Dataset

In our proposed method, first the datasets utilized to represent news agency website addresses to achieve the propagator website information to train our algorithm how to recognize real and fake news. Also we needed news to be identified whether they are real or fake. On the whole, there were two appropriate datasets for our project [26][27]. They were considered as the most encouraging versions for pre-processing, characteristic extraction, and categorization model. The reason is that other data sets lack url resources where articles/texts of the statement are produced and propagated. The citation of resources for article text to investigate the reliability is considered a very important news. It also helps tagging the data as fake or unreliable. These features are as follows:

Text, Title, Label, Image_URL, Site_URL, Published date, Author

This dataset is comprised of 9 tags and the articles related are categorized within this category. We have used dataset [27] due to the limitations mentioned in last section.

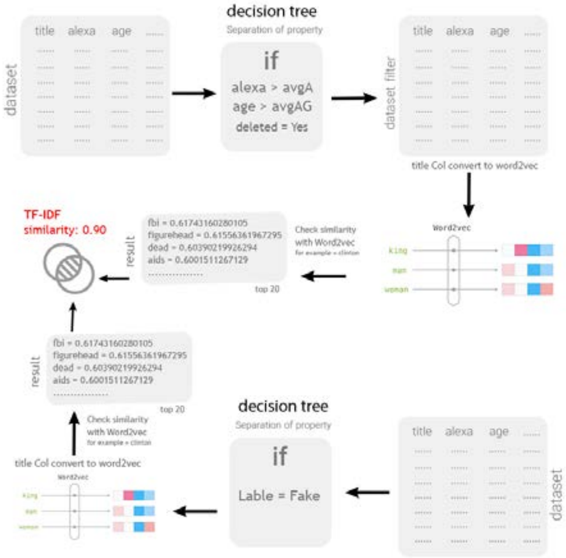


Fig.1. An overall view of our proposed algorithm

3-2: Data analysis and method

In this section we have identified the methods that can be used to teach dataset to recognize correct news.

3-2-1: Calculating domain age

The very first stage to recognize news is the identification of the domain age. We expect that websites with longer history should have entailed almost lower amounts of unreal news than the rest of websites in our database. Therefore, we have shown in figure 1 that the longer history of websites leads to less incorrect news propagation compared to newer websites. We used the Domaintools [28] tool to check the age of the site

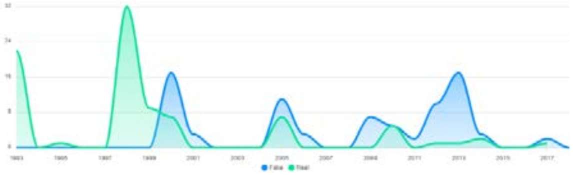


Fig.2. The age of websites propagating unreal news

3-2-2: Unreal news being hidden

A valid news agency tends to maintain its rank to continue their popularity and to keep the propagated news for a longer time. Meanwhile, fake news websites often hide news and make it offline after they achieve their short-time goals to astray their readers. Our analysis regarding the set of unreal and real news sets approves such a common approach.

As it has been represented in table 2, three sets of data from among unreal datasets show the hiding algorithms of fixed news. Meanwhile, the set of real news contains much less offline news. On the whole, there are only 8 real news from among 57 deleted news and 7 of them have completely stopped their activities completely in the website. There could be several different reasons for their lack of activities and only one of them had eliminated the real news after being propagated. Therefore, we believe that news becoming hidden could be considered as a valuable characteristic to recognize news.

3-2-3: Domain popularity using Alexa

To investigate about the website popularity, we have used Alexa. Alexa utilizes a tool called ranking. Alexa ranking is a number attributed by Alexa to rank the websites. Using this number, based on the traffic, your website will be estimated to represent your website in search results and to measure the popularity of your website by Alexa. The lower amounts of this number, represents higher traffic and more acceptability of your website on the part of your users. It is natural that a more popular website has higher numbers of daily visits because the visitors tend to spend more time to browse the website. Naturally, famous websites have greater audiences than those newly established websites which propagate unreal news. As it has been represented in figure 4, valid websites which propagate less real news are not very popular. The news agencies with a rank of lower than 2000 did not propagate any unreal news. Thus, the domain popularity parameter could be considered as a fundamental step in recognizing unreal news.

Figure 4- Domain popularity based on Alexa

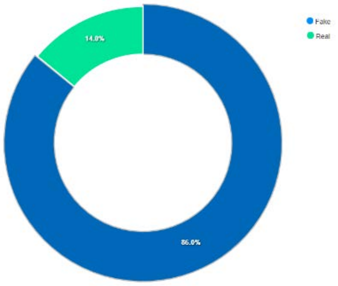


Fig.3. Unreal news becoming hidden

4. Discussion and Conclusion

The proposed method here is known as a very precise method using vector. In fact, words are changed into vectors to have a more consistency capability. In our method, first it was required to utilize datasets with internet addresses of the new agency in order to gain the information of the publisher website in your algorithm to give required instructions in order to recognize real and fake news. Also we need news where the real and fake ones are recognized and thus we utilized a dataset [27] to train and then used word2vec algorithm to investigate about the similarities. The reason to choose the present algorithm refers to the fact that the process of vast datasets is time consuming. This is not only due to the vast volume of the data, but also due to the fact that the type and structure of the data could be different and sophisticated. Currently, many of the data investigation techniques and machine learning are used to fight against great amounts of data. Some of them can make good learning algorithms considering great training examples. Meanwhile, due to the data dimension, if the learning algorithm is able to select useful characteristics or reduce the features, it would be more efficient [29].Word2vec which is presented and supported by google is formed of two learning models called Continuous Bag of Words (CBOW) and Skip-gram, respectively. Word2vec enters text data into one of learning models, the word vectors are given as the outputs which could be represented as a big text piece or even through the total article. In the present research, first we taught the data through word2vec model and assessed the similarity of the word. In our method, there is a need for two sets to compare the similarity. One is the data with fake tags isolated from the cluster using the decision tree algorithm of the data identified as fake in the dataset and then using the word2vec algorithm they will be sent to a set in order to investigate about the similarity. Then, using the parameters gained, the domain age, Alexa rank, and news being hidden were utilized in order to filter and isolate the news using a decision tree algorithm.

Since the domain age is known as a criterion to discover the news but it is not an absolute reason to discover the news and also the reason to become hidden, based on our investigations, could refer to several different conditions such as lack of activity permanence and this is not due to unreal news, this case was utilized in the dataset. Thus, it is known as the only most precise method that could help us in discovering the unreal news and it was due to Alexa rank. We received the Alexa average and divided it into two parts of unreal and real news based on our own data and taught them. In fact, if the Alexa rank of the news agency website was lower than the average it would be real and if it was higher than the average, we considered it to be unreal. Due to Alexa rules, the closer to zero numbers represents higher fame of the news agency. In fact, our goal was to discover unreal news and therefore we were looking for websites with higher than the average Alexa.

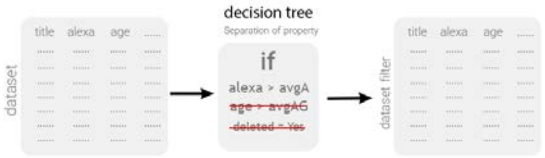


Fig.4. Lack of requirement of domain age and getting hidden

In second phase, a dataset filtered in isolation is gained through which we decide whether Stop Word words are invalid words or not. For example (from, that, he, for ,…) are deleted from title of the articles or the data present in the dataset and then are changed into vectors. This results in our training set volume becoming lower and also more precise. Although word2vec puts the similar words into the same vector, the training time requires lots of resources through which stop word could play a proper role in optimizing the training. After preparing the optimized constituents, we enter data in order to process into word2vec. Now we need an input to study the similarity in word2vec. For example, Clinton chosen by word2vec in the form of an element starts to look for its neighbors based on the algorithm and calculates 20 superior words related with Clinton along with the related percent.

On the other hand, we need data to investigate about the set power gained. Therefore, in our own dataset we use the decision tree algorithm to put an unreal tagged set in another element and change the texts into vectors.

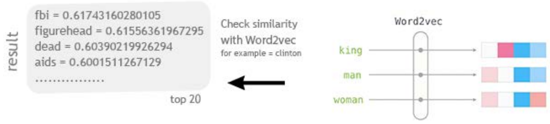


Fig.5. Unreal tagged data

To compare the datasets, we need a tf-idf similarity algorithm which is the best tool to investigate about the vector similarity. TF-IDF means the reverse frequency of the document frequency. The statistical method results in a number to compute the importance of a word against a document within a set or a frame. Based on other investigations, we concluded that the most precise tool to investigate the issue is vector similarities comparison. We have drawn the overall method in third figure in a complete form.

According to the results gained, the precision of our algorithm in some of the words is between 0.60 and 0.94 and the closer to 1 means the greater similarity between the two sets. We believe the word2vec algorithm is the best possible option for vector similarity and it presents very precise calculations for us in a way that even a minor mistake could result in a high similarity percentage which shows the high precision and sensitivity of this algorithm. We share some important words based on investigation carried out and resulted in our findings.

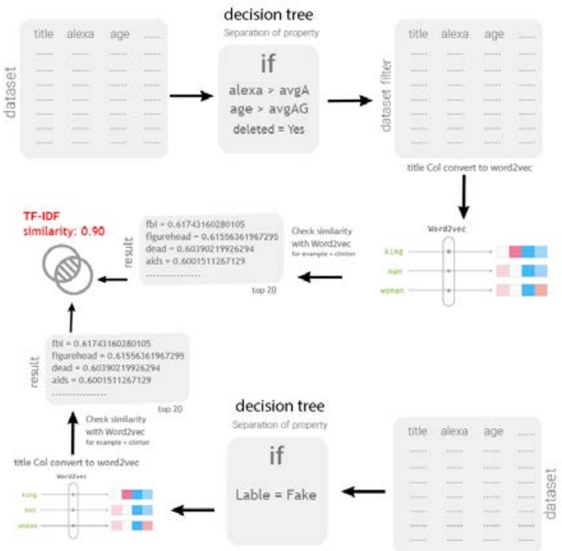
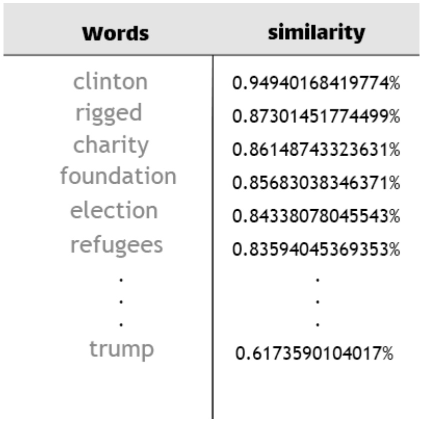


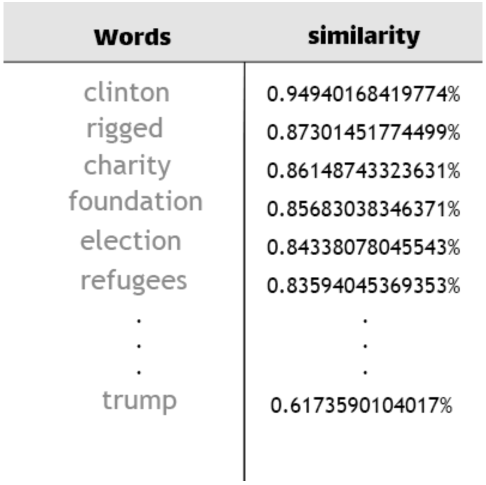
Fig.6. The overall stages of unreal news recognition in a glance



Words	similarity
clinton	0.94940168419774%
rigged	0.87301451774499%
charity	0.86148743323631%
foundation	0.85683038346371%
election	0.84338078045543%
refugees	0.83594045369353%
⋮	⋮
trump	0.6173590104017%

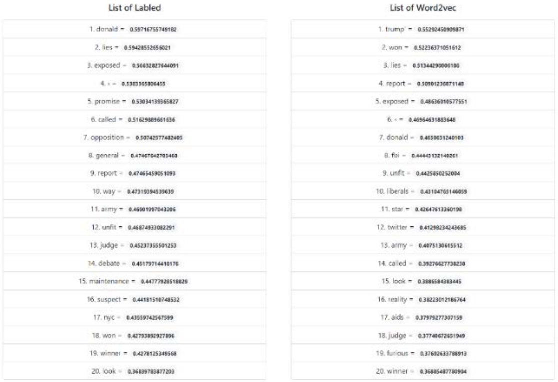
Fig.7. Results gained in different words

Regarding two words of Clinton and trump, the similarity of the two sets using tf-idf similarity algorithm, the results were equal to 61 percent and 94 percent as follows.



Words	similarity
clinton	0.94940168419774%
rigged	0.87301451774499%
charity	0.86148743323631%
foundation	0.85683038346371%
election	0.84338078045543%
refugees	0.83594045369353%
⋮	⋮
trump	0.6173590104017%

Fig.8. The search gained through the word Clinton



List of Labled	List of Word2vec
1. donald = 0.59716755749182	1. trump' = 0.55292450909871
2. lies = 0.59428552656021	2. won = 0.52236371051612
3. exposed = 0.56632827644091	3. lies = 0.51344290006186
4. ‹ = 0.5383165806455	4. report = 0.50901230871148
5. promise = 0.53034139365827	5. exposed = 0.48636010577551
6. called = 0.51629889661636	6. ‹ = 0.46964631883648
7. opposition = 0.50742577482405	7. donald = 0.4650631240103
8. general = 0.47467042705468	8. fbi = 0.44443132140201
9. report = 0.47465459051093	9. unfit = 0.4425850252004
10. way = 0.47319394539639	10. liberals = 0.43104765146059
11. army = 0.46901997043206	11. star = 0.42647613360198
12. unfit = 0.46874933082291	12. twitter = 0.41298234243685
13. judge = 0.45237355501253	13. army = 0.4075130615512
14. debate = 0.45179714418176	14. called = 0.39276627738238
15. maintenance = 0.44777928518829	15. look = 0.3886584383445
16. suspect = 0.44181510748532	16. reality = 0.38223012186764
17. nyc = 0.43559742567599	17. aids = 0.37979277307159
18. won = 0.42793892927896	18. judge = 0.37740672651949
19. winner = 0.4278125349568	19. furious = 0.37692633788913
20. look = 0.36839783877203	20. winner = 0.36885487780904

Fig.9. The search gained through the word Trump

5.Suggestions

We believe that through finding news source and avoiding the propagation of them we could achieve success because it would be difficult to avoid news propagation. Thus, we need an environment and a big platform to discover news in order to recognize and block news immediately. Therefore, it would be necessary to recognize and discover unreal news using meaning understanding and recognizing certain characteristics. It could not work with a high precision because the distributor can permanently affect the news propagation and this method could not be a really effective one.

In an article about Daily Time on Site, Alexa is utilized as one of the overall calculation indexes to rank and isolate the data. To do so, the major rank and the overall Alexa are utilized to recognize the news because the maintenance rate or Daily Time on Site can be computed using different methods such as making the website graphic more attractive and creating the recreations to enhance this rate. This solely cannot represent popularity and we decided to utilize the major Alexa ranking for our own training set.

6. Limitations and Suggestions for Future works

Our proposed method has had some administration constraints. We needed a high power server to train in order to be able to change the data and words into vectors. To do so, we only could use low numbers to train. Surely, if there exists a great deal of news, the recognition power would be greater. Also we added only news titles to our training set meanwhile there could be explanations and detailed news to include more meaningful words. Also, the presentation of an input based on images and videos is known as an activity that could be dealt in future in details. Additionally, learning the data based on search engines could be investigated in future works.

References

[1] Simon Kemp, “Digital 2022: Global overview report,” wearesocial, 2022. https://wearesocial.com/au/blog/2021/04/60 percent-of-the-worlds-population-is-now-online/

[2] E. S. and K. E. Matsa, “News Use Across Social Media Platforms 2018,” Pew Res. Cent., 2018, [Online]. Available: https://www.pewresearch.org/journalism/2018/09/10/news-use-across-social-media-platforms-2018/

[3] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” ACM Comput.Surv., vol. 53, no. 5, Sep. 2020, doi: 10.1145/3395046.

[4] B. D. Horne, J. N\orregaard, and S. Adali, “Robust Fake News Detection Over Time and Attack,” ACM Trans. Intell. Syst. Technol., vol. 11, no. 1, Dec. 2019, doi: 10.1145/3363818.

[5] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” Nat. Commun., vol. 9, no. 1, p. 4787, 2018, doi: 10.1038/s41467-018-06930-7.

[6] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election (No. w23089).” National Bureau of Economic Research Cambridge, MA, 2017.

[7] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT),” Appl. Sci., vol. 9, no. 19, 2019, doi: 10.3390/app9194062.

[8] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, “FNDNet – A deep convolutional neural network for fake news detection,” Cogn. Syst. Res., vol. 61, pp. 32–44, 2020, doi:https://doi.org/10.1016/j.cogsys.2019.12.005.

[9] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, “DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning.” arXiv, 2018. doi: 10.48550/ARXIV.1809.06416.

[10] 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 &amp; Knowledge Management, 2020, pp. 1165–1174. doi: 10.1145/3340531.3412046.

[11] P. Przybyla, “Capturing the Style of Fake News,” Proc. AAAI Conf. Artif. Intell., vol. 34, no. 01, pp. 490–497, 2020, doi: 10.1609/aaai.v34i01.5386.

[12] R. M. Silva, R. L. S. Santos, T. A. Almeida, and T. A. S. Pardo, “Towards automatically filtering fake news in Portuguese,” Expert Syst. Appl.,

vol. 146, p. 113199, 2020, doi: https://doi.org/10.1016/j.eswa.2020.113199.

[13] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: {A} New Multimodal Benchmark Dataset for Fine-grained Fake News Detection," CoRR, vol. abs/1911.0, 2019, [Online]. Available: http://arxiv.org/abs/1911.03854

[14] Y. Liu and Y.-F. B. Wu, "FNED: A Deep Network for Fake News Early Detection on Social Media," ACM Trans. Inf. Syst., vol. 38, no. 3, May 2020, doi: 10.1145/3386253.

[15] K. Shu, S. Wang, and H. Liu, "Beyond News Contents: The Role of Social Context for Fake News Detection," in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 312–320. doi: 10.1145/3289600.3290994.

[16] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science (80-.)., vol. 359, no. 6380, pp. 1146–1151, 2018, doi: 10.1126/science.aap9559.

[17] Wikipedia contributors, "Word2vec {Wikipedia}{,} The Free Encyclopedia." 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Word2vec&oldid=1073402438

[18] Wikipedia contributors, "tf–idf." https://en.wikipedia.org/wiki/Tf–idf

[19] 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, Jun. 2020, doi: 10.1089/big.2020.0062.

[20] M. N. Nikiforos, S. Vergis, A. Stylidou, N. Augoustis, K. L. Kermanidis, and M. Maragoudakis, "Fake News Detection Regarding the Hong Kong Events from Tweets BT - Artificial Intelligence Applications and Innovations. AIAI 2020 IFIP WG 12.5 International Workshops," 2020, pp. 177–186.

[21] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "{SMOTE}: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[22] S. Castelo et al., "A Topic-Agnostic Approach for Identifying Fake News Pages," in Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 975–980. doi: 10.1145/3308560.3316739.

[23] D. Mangal and D. K. Sharma, "A Framework for Detection and Validation of Fake News via Authorize Source Matching BT - Micro-Electronics and Telecommunication Engineering," 2021, pp. 577–586.

[24] D. K. Vishwakarma, D. Varshney, and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," Cogn. Syst. Res., vol. 58, pp. 217–229, 2019, doi: https://doi.org/10.1016/j.cogsys.2019.07.004.

[25] "alexa ranking." https://www.alexa.com/help/privacy

[26] R. Bhatia, "Source based Fake News Classification," 2020. https://www.kaggle.com/ruchi798/source-based-news-classification

[27] Craig Silverman, "BuzzFeed News dataset," buzzfeednews, 2016. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-onfacebook?utm_source=dynamic&utm_campaign=bfsharecopy

[28] Domaintools, "Whois Lookup." https://www.domaintools.com/

[29] L. Ma and Y. Zhang, "Using Word2Vec to process big text data," in 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 2895–2897. doi: 10.1109/BigData.2015.7364114.

.