Detecting Fake News: Exploring Key Features in Multilingual Arabic Dialect Corpus

Hocini Abdelouahab, Kamel Smaïli

Detecting Fake News: Exploring Key Features in Multilingual Arabic Dialect Corpus

Abdelouahab Hocini and Kamel Smaili

LORIA, University of Lorraine, F-54600, France
abdelouahab.hocini@univ-lorraine.fr, smaili@loria.fr

Abstract. As misinformation continues to spread rapidly on social media platforms identifying and stopping the dissemination of fake news has become an urgent need. In this article, we propose a deep learning approach leveraging keywords for feature extraction and classification of Arabic dialect fake news. Our method achieves an accuracy of 82.3% on a corpus comprising 3000 news articles in Algerian and Tunisian dialects, Modern Standard Arabic (MSA), French, and English, featuring instances of code-switching between these languages; as well as an accuracy of 93.7% on an English fake news corpus. Our experimentation shows that the shortcut learning problem that can arise when using keyword based features can be solved using regularization techniques. Our findings also show that our approach will achieve better performance on larger Arabic dialect corpora.

Keywords: Fake News · Classification · Arabic Fake News · Deep Learning.

1 Introduction

The surge in fake news across social media platforms has become a major obstacle for researchers since the internet's widespread adoption, as it provides global access to these platforms for a diverse range of users.This problem has gained greater visibility, especially during the 2016 US presidential elections, when the spread of false information reached unprecedented heights in an attempt to sway public opinion and influence the election results. The potential for exploiting this phenomenon becomes even more alarming when contemplating its implications within the context of fourth-generation warfare, characterized by a fusion of war and politics, blurring traditional distinctions between the two.Such occurrences are commonplace in the Arab world; each day sees the circulation of numerous pieces of fake news across social media platforms, spanning topics from the demise of prominent figures to matters concerning a nation's economy or diplomacy. Swift and accurate detection of these reports is crucial to mitigate their enduring effects. In this work, we aim to approach the problem of fake news detection from a different perspective than traditional text classification. Our main goal is to construct an identity map of fake news containing its various features and to exploit these features to recognize it. Our study focuses

on the possibility of building a set of keywords that will allow us to determine, based on their presence or not, whether a given piece of information is real or not. Our investigation will focus on a multilingual Arabic dialect fake news corpus, featuring news articles in Algerian and Tunisian dialects, Modern Standard Arabic (MSA), French, English, Arabizi, and code-switching—comprising any combination of these languages. Following that, we will validate our findings by implementing our technique on English fake news corpora. This will help us identify any disparities, if present, in the efficacy of our method when applied to a single-language corpus compared to a multilingual one. The remainder of this paper is structured as follows: in Section 2 we discuss research work related to our study and give aguments for the choice of our representation. Then we present our work and results in section 3. In Section 4 we discuss and compare the behaviour of our proposed approach on an english corpus. "Conclusion" presents the conclusion and future work on the problem.

2 Related Works

When addressing the challenge of detecting fake news, there are two key considerations. Firstly, it is crucial to determine the approach, namely, which aspects of fake news and/or social media platforms to leverage in making the decision. Once this direction is established, a corpus is essential for testing and evaluating the proposed technique. In this section, drawing from our review of existing literature, we will provide an overview of various fake news detection techniques, with an emphazis on those centered on analyzing the content of fake news, along with a brief examination of available fake news corpora.

2.1 Fake News Detection Approaches

The task of detecting fake news on social media involves determining the authenticity of a given piece of information. However, this task isn't limited to binary classification alone. Existing literature presents various techniques to tackle this challenge. Depending on the approach taken, these techniques can be categorized into different groups as shown in Figure 1. The first major categories are relating to which information is leveraged when trying to decide on the truth value of a news. **Network based approaches:** Previous studies showed that fake news behave differently from real news. They are characterized by a high velocity when spreading between social media platform users [16]. In the same perspective, network based fake news detection techniques focus on studying the social network and the dissemination of a news to identify fake ones. In [20] the authors proposed a network-based pattern-driven fake news detection approach, their technique uses graph representation to highlight the following patterns: Fake news (i) spread further and (ii) attract more spreaders who are often (iii) more engaged and (iv) more densely connected in the network. In [12], de Souza et al. proposed a network-based approach based on Positive and Unlabeled Learning by Label Propagation (PU-LP).

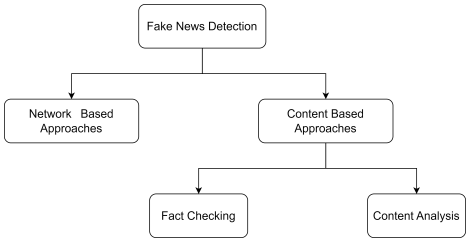

Fig. 1. Fake news detection approaches

Content based approaches: These approaches focus on studying the content of the news article (text or image or both).

Fact checking: it's the process of comparing the news article with a previous ground truth knowledge base. This process can be done manually by experts or specialized agencies, crowd sourced or done automatically. Some known fact-checking websites are: politifact[1], Fake news DZ[2] and Falso[3].Two surveys have been recently conducted for news automatic fact-checking. In [5], a three step fact-checking pipeline has been presented as shown in Figure 2. On the other side [19] presented a review of relevant research on claim detection and claim validation components in the fact-checking process. ***Content analysis***: Mainly

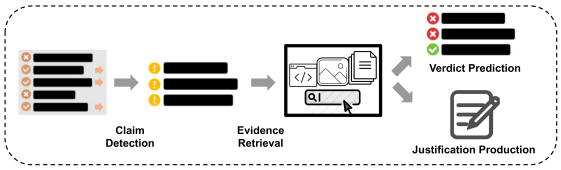

Fig. 2. A natural language processing framework for automated fact-checking [5].

supervised machine and deep learning algorithms are used to assign predefined class labels to news articles. Ahmed et al. [3] utilized a linear support vector machine (LSVM) trained on TF-IDF vectors on a dataset comprising 2000 news articles, resulting in an accuracy rate of 92%. In [18] the authors conducted extensive experimentation using Random Forest and XGBoost with tf, tf-idf and character/word n-gram features for fake news classification achieving an accuracy of up to 95% on Kaggle's real_or_fake dataset. Furthermore, beyond the previously cited techniques, various deep learning architectures have been used for the classification of fake news. Deep Average Network [6] is an architecture that views the text as a Bag of Words of which the representation is the aver-

[1] www.politifact.com
[2] www.facebook.com/FakenewsDZ
[3] www.facebook.com/falso.tn

age of the embeddings, a Multi-Layer Perceptrons is then used as a classifier. CNN-based models aim at capturing patterns in texts while RNN-based models recognize word dependencies and text structures. In [2] the authors proposed a deep architecture of which the backbone was composed of a CNN followed by an LSTM for feature extraction, the resulting feature maps were used for the classification achieving an accuracy of 97.2%.

2.2 Fake News Corpora

Fake news datasets typically consist of collections of news articles, social media posts, or other textual content, annotated with labels indicating their veracity. In recent research work, multiple fake news datasets are available: LIAR [17] contains 12.8K manually labeled statements into 6 classes (true, mostly-true, half-true, barely-true, false, pants-fire). PolitiFact [15] is a multimodal (text and image) fake news dataset composed of a collection of tweets and fact checked articles from politifact.com labeled as real or fake news. FakeNewsNet [10] is a fake news dataset built with articles collected from politifact and gossipcop labeled into two classes (real and fake).

In this work, we conducted our experimentation on BOUTEF[11], a comprehensive Arabic dialect fake news corpus, the nature of the dialect implies that the corpus contains articles in various languages, namely: Algerian and Tunisian dialects, Modern Standard Arabic (MSA), French, and English, featuring instances of code-switching between these languages. The distribution of the languages accross news articles is shown in Table 1. This dataset contains news post collected from Facebook, Twitter, Youtube, TikTok and other websites reflecting the diverse sources of misinformation. BOUTEF classifies news articles into three classes: Fake News, Fake News Comment and Fake News Denial, further more it offers a labeling scheme consisting of 40 categories.

Table 1. Quantitative information regarding the corpus and language distribution

	Fake	Fake Comments	Denial	Total
MSA	313	788	206	1307
ALGDIA	24	786	1	811
TUNDIA	0	382	0	382
CODE SWITCHING	7	308	6	321
ARABIZI	0	110	0	110
TARABIZI	0	61	0	61
ENG	18	40	15	73
FRE	90	419	91	600
MARABIZI	0	1	0	1
TOTAL	**452**	**2895**	**319**	**3666**

3 Arabic Dialect Fake News Classification

In this work we propose a novel approach for Multilingual Arabic dialect fake news classification, the approach we propose is composed of three main steps

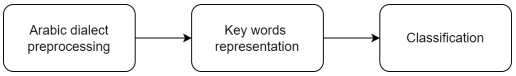

Fig. 3. the workflow of our approach for Arabic dialect fake news classification.

as shown in Figure 3: Arabic dialect text preprocessing, the data representation using the key words and finally the classification task.

3.1 Data Preprocessing

BOUTEF corpus is complex due to the nature of the news elements in it, all the posts were collected from social media platforms which make them cover a huge spectrum of lexical entities; In order to address this issue, initially, we partitioned the corpus based on language into distinct sub-datasets to facilitate customized preprocessing for each language. This resulted in the creation of four subsets: English, French, Arabic, and code-switching. For the Arabic subset, we first applied a cleaning step for dialect news articles, where we removed repeated letters within the same word, and the we used Farasa segmentation tool [1]. We applied lemmatization for French and English written news articles. Finally, we reassemble the sub-datasets into one dataset containing our three classes: Fake, FakeComment and NoFake. The last step of the preprocessing is to build a list of the most frequent words for each one of the classes.

3.2 Data Representation

For the use of ML and DL models for fake news classification, we need to give a numerical representation of the news articles to be able to extract features, usually in literature tf-idf[9, 7] or word embedding[14] are the most common representations.

TF-IDF: Term Frequency-Inverse Document Frequency is a statistical measure utilized to assess the significance of terms within a corpus of Y documents.

$$tf\text{-}idf(x, y) = tf_{x,y} \times \left(1 + \log\left(\frac{Y}{df_x}\right)\right) \tag{1}$$

The tf-idf formula we used (Equation 1) is different from the classic one, we added 1 to the idf of a term to avoid null values for words that are common to all three classes in BOUTEF.

The Figure 4 shows the process of building our feature map for a given news article, N being the desired vector size, we consider the $N/3$ most frequent words for each of the three classes of BOUTEF (Fake, FakeComment and Denial). We iterate through the list of keywords and verify whether each word is present within the news article. If a word is found, we encode it with its corresponding tf-idf value in the respective position of the output vector. Otherwise, we assign a value of 0.

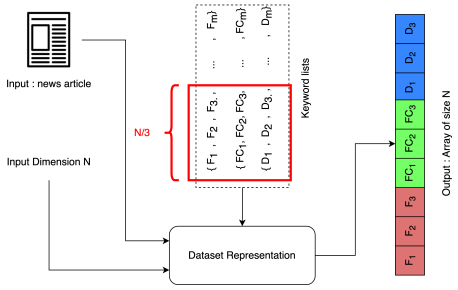

Fig. 4. News article representation using tf-idf.

3.3 Classification

According to previous works machine learning models and shallow deep learning models are more efficient for text classification tasks [8], this encourages that we first explore using machine learning models for our classification task. The figure 5 illustrates the performance of some ML models in classifying news articles in BOUTEF namely: naive bayes, decision tree, random forests and Support Vector Machine (SVM). The models' accuracy is plotted against the input vector dimension.

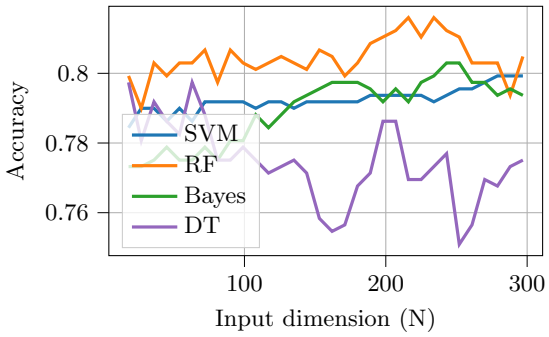

Fig. 5. Supervised machine learning models' accuracy on BOUTEF corpus.

Table 2 shows the best achieved accuracy with each classifier after fine tuning model parameters. The best performing classifier is Random Forest with an accuracy of 81.6%.

We propose to use a Multi-Layer Perceptron as a deep learning model to do the classification task, the choice of this model is based on the following criteria: Bag of Word news representation (see section 3.2) which excludes CNN's and RNN's and relatively small corpus which excludes the usage of large data-hungry models. The input layer of our model is the $tf\text{-}idf$ vector of size N, followed by

Table 2. Supervised machine learning models' best accuracy on BOUTEF.

Classifier	Parameters	Accuracy	Input Dimension
Naïve Bayes	$\alpha = 1$	80.3%	252
Decision Tree	Criterion = gini	79.74%	63
Random Forest	Random states = 100	**81.6%**	216
SVM	Kernel = RBF	79.92%	279

two hidden layers of size $N \times 2$ and $N/2$ and an output layer of size 3 with a softmax activation function. We achieved a validation accuracy of **82.3%**.

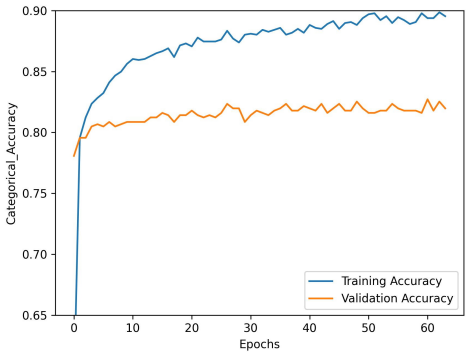

Fig. 6. Learning curve of the MLP with an input vector of size 279.

4 Investigating the Results

Classifying fake news contained in BOUTEF yielded satisfying results, but we can notice an issue, the model doesn't generalise very well (see Figure 6). We propose in this section to investigate this phenomenon and explain it as it can be caused by the relatively small size of the dataset, the class imbalance or by the nature of the Arabic dialect that adds to the complexity of the problem. In order to answer this question, we proposed to observe the behaviour of our model on two English fake news datasets. As shown in table 3, in contrast with BOUTEF, the datasets we chose were split into two classes only (Real or Fake news) and both classes are equally represented and finally they are larger than the corpus we trained on previously.

Our choice of there two datasets was made with an emphasis on the following criteria:

– Class balance: We choose corpora with balanced classes to avoid the phenomenon of overfitting. If one class dominates the corpus, the model will over-fit that class.

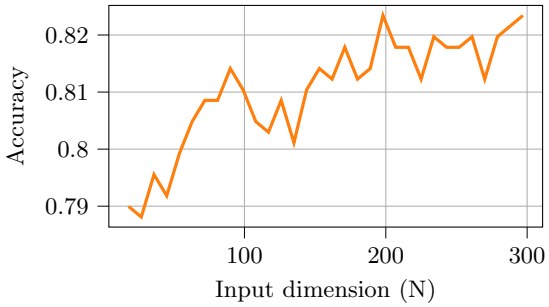

Fig. 7. MLP accuracy on BOUTEF.

Table 3. Kaggle fake news datasets used for the investigation.

Dataset	Classes	Real news count	Fake news count
Dataset 1	2 (Real, Fake)	3171	3164
Dataset 2		21418	23538

– Dataset's size: We have chosen the first dataset with a size similar to that of BOUTEF, to have comparable results. The second one is larger to determine if having more training data could lead the model to achieve better accuracy.

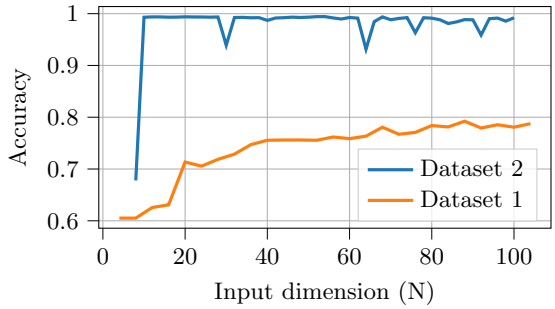

Fig. 8. Validation accuracy of the model based on the dimension of the input vector.

Figure 8 shows the achieved validation accuracy on both datasets. We can notice that comparing between BOUTEF and the first dataset the model performs similarly achieving a maximum accuracy of 79.23% on Dataset1 vs 82.3% ,so we can conclude that given the same amount of training data the model behaves in the exact same way despite the complexity of the language and the class imbalance in BOUTEF (see section 2.2). But when looking at the curve of the model's validation accuracy on dataset2, a strange phenomenon becomes apparent; With an input array of size 8 the accuracy is of 67.8% and the next

Table 4. Keyword lists for input size 8 and 10.

Input vector size (N)	Keyword list
10	Real: (”say”,”trump”,”state”,”would”,”reuters”) Fake: (”trump”,”say”,”president”,”people”,”go”)

step (i,e with adding one key word to each class thus getting an input size of 10) the model's accuracy spikes to 99.3% and stabilizes. This indicates a strong correlation between either word and one of the classes, implying that the model relies on this shortcut rather than truly learning the underlying relationship between the features and the classes; This phenomenon is known as **shortcut learning**[4]. In Table 4, we provide a list of keywords, among which **reuters** and **go** stand out as potential sources of shortcut learning. To pinpoint the specific culprit, we conducted additional model training experiments: initially omitting one keyword and then the other, we observe the model's behaviour. We can

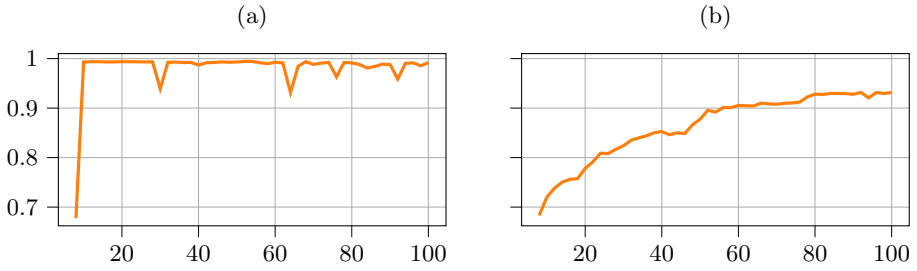

Fig. 9. Model's validation accuracy against input vector size (N) on dataset2: **(a)** The word ”go” was removed from the keyword list. **(b)** The word ”reuters” was removed from the keyword list.

notice that in Figure 9.a, after removing the word **go**, the shortcut learning problem persists meaning that it probably wasn't at the origin of this problem. On the other hand, on Figure 9.b after removing the word **reuters**, the problem disapears. We found out that the word **reuters** is strongly correlated to one of the classes in dataset2, leading the MLP to exploit this shortcut instead of trying to learn the relation between the features and the class lables. We then conduct more experimentation in order to understand if this problem arises anytime we consider the word **reuters** in the list of keywords. The detailed methodology of our experimentation can be summarized in two simple steps: Firstly, we exclude **reuters** from the list of keywords to prevent it from automatically influencing the construction of our feature map (i.e., input vector). Secondly, we reintroduce it at various positions (10, 20, 54, 76) and repeat the training process. The results of this experiment are depicted in Figure 10. The findings validate that

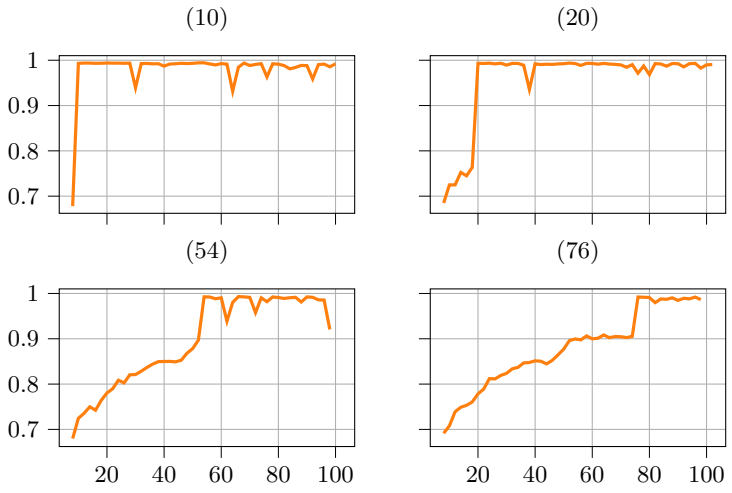

Fig. 10. MLP accuracy on Dataset2 against input size after inserting the word "reuters" in the keywords at different positions.

reuters indeed induces the phenomenon of shortcut learning. Whenever this problematic word is included in the list of keywords, the model's accuracy spikes to 99.3% and remains stable at this level. To resolve this issue we could use regularization techniques to smooth the model's weights and avoid large weights, the most popular techniques used in literature are: Dropout regularization [13], L1 and L2 regularization.

Dropout: it's a regularization technique that consists in randomly removing some of the neurons in the network during training with a probability of $p = DropoutRate$ with $0 \leq DroupoutRate \leq 1$. These neurons are put back in the network during validation to perform inference.

L1 and L2 regularization: these two techniques modify the loss function by adding a term in function of the model's weights. The modified loss functions are show in Equation 2 and Equation 3 for L1 and L2 regularization respectively:

$$\mathcal{L}' = \mathcal{L} + \alpha \cdot \sum w_i \tag{2}$$

$$\mathcal{L}' = \mathcal{L} + \alpha \cdot \sum (w_i)^2 \tag{3}$$

With $0 \leq \alpha \leq 1$, we determine at which extent we punish the model for having higher weights with 0 meaning no punishment and 1 meaning the highest possible punishment. We conducted extensive experimentation to choose the best regularization algorithm to use for mitigating the shortcut learning problem, the best results were obtained with L2 regularization, Figure 11 shows the validation accuracy of the MLP with two different values of α, we can infer that L2 regularization effectively addresses the issue of shortcut learning by appro-

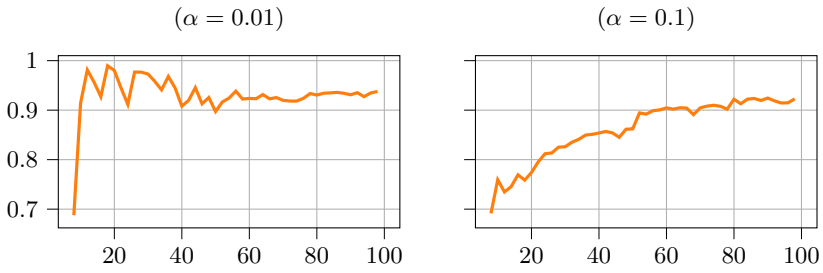

Fig. 11. Effect of L2 regularization on preventing shortcut learning phenomenon. The X axis shows the input vector size, the Y axis shows the validation accuracy.

priately penalizing large weights and enabling the model to reach an accuracy of 93.7%.

5 Conclusion

In this paper, we introduce a deep learning model for fake news detection on a multilingual Arabic dialect corpus, we leverage the Bag of Words representation combined with keywords for features extraction. Our approach showed great resiliency and efficiency achieving a 82.3% accuracy on an unbalanced and relatively small corpus of approximately 3,000 articles comprising news content in Arabic dialect, Modern Standard Arabic (MSA), French, and English; and an accuracy of 93.7% on a larger English fake news corpus. Our studies also showed that L2 regularization can solve the shortcut learning problem that could occur when leveraging keywords for the classification task. The primary focus of this work is to detect fake news early, facilitating early and efficient response from social media platforms and to protect their users.

References

1. Abdelali, A., Darwish, K., Durrani, N., Mubarak, H.: Farasa: A fast and furious segmenter for Arabic. In: DeNero, J., Finlayson, M., Reddy, S. (eds.) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 11–16. Association for Computational Linguistics, San Diego, California (Jun 2016)
2. Agarwal, A., Mittal, M., Pathak, A., Goyal, L.M.: Fake news detection using a blend of neural networks: An application of deep learning. SN Computer Science **1**, 1–9 (2020)
3. Ahmed, H., Traore, I., Saad, S.: Detection of online fake news using n-gram analysis and machine learning techniques. In: Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1. pp. 127–138. Springer (2017)

4. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
5. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Transactions of the Association for Computational Linguistics **10**, 178–206 (2022)
6. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers). pp. 1681–1691 (2015)
7. Liu, C.z., Sheng, Y.x., Wei, Z.q., Yang, Y.Q.: Research of text classification based on improved tf-idf algorithm. In: 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE). pp. 218–222. IEEE (2018)
8. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: A comprehensive review. ACM Comput. Surv. **54**(3) (apr 2021)
9. Qaiser, S., Ali, R.: Text mining: use of tf-idf to examine the relevance of words to documents. International Journal of Computer Applications **181**(1), 25–29 (2018)
10. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
11. Smaili, K., Hamza, A., Langlois, D., Amazouz, D.: Boutef: Bolstering our understanding through an elaborated fake news corpus. In: International conference on Arabic language processing. Springer (2024)
12. de Souza, M.C., Nogueira, B.M., Rossi, R.G., Marcacini, R.M., Dos Santos, B.N., Rezende, S.O.: A network-based positive and unlabeled learning approach for fake news detection. Machine learning **111**(10), 3549–3592 (2022)
13. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
14. Verma, P.K., Agrawal, P., Amorim, I., Prodan, R.: Welfake: Word embedding over linguistic features for fake news detection. IEEE Transactions on Computational Social Systems **8**(4), 881–893 (2021)
15. Vo, N., Lee, K.: Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020) (2020)
16. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
17. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)
18. Wynne, H.E., Wint, Z.Z.: Content based fake news detection using n-gram models. In: Proceedings of the 21st international conference on information integration and web-based applications & services. pp. 669–673 (2019)
19. Zeng, X., Abumansour, A.S., Zubiaga, A.: Automated fact-checking: A survey. Language and Linguistics Compass **15**(10), e12438 (2021)
20. Zhou, X., Zafarani, R.: Network-based fake news detection: A pattern-driven approach. SIGKDD Explor. Newsl. **21**(2), 48–60 (nov 2019). https://doi.org/10.1145/3373464.3373473