



Fake news detection: deep semantic representation with enhanced feature engineering

Mohammadreza Samadi¹ · Saeedeh Momtazi¹

Received: 4 September 2021 / Accepted: 10 February 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

Due to the widespread use of social media, people are exposed to fake news and misinformation. Spreading fake news has adverse effects on both the general public and governments. This issue motivated researchers to utilize advanced natural language processing concepts to detect such misinformation in social media. Despite the recent research studies that only focused on semantic features extracted by deep contextualized text representation models, we aim to show that content-based feature engineering can enhance the semantic models in a complex task like fake news detection. These features can provide valuable information from different aspects of input texts and assist our neural classifier in detecting fake and real news more accurately than using semantic features. To substantiate the effectiveness of feature engineering besides semantic features, we proposed a deep neural architecture in which three parallel convolutional neural network (CNN) layers extract semantic features from contextual representation vectors. Then, semantic and content-based features are fed to a fully connected layer. We evaluated our model on an English dataset about the COVID-19 pandemic and a domain-independent Persian fake news dataset (TAJ). Our experiments on the English COVID-19 dataset show 4.16% and 4.02% improvement in accuracy and f1-score, respectively, compared to the baseline model, which does not benefit from the content-based features. We also achieved 2.01% and 0.69% improvement in accuracy and f1-score, respectively, compared to the state-of-the-art results reported by Shifath et al. (A transformer based approach for fighting covid-19 fake news, arXiv preprint [arXiv:2101.12027](https://arxiv.org/abs/2101.12027), 2021). Our model outperformed the baseline on the TAJ dataset by improving accuracy and f1-score metrics by 1.89% and 1.74%, respectively. The model also shows 2.13% and 1.6% improvement in accuracy and f1-score, respectively, compared to the state-of-the-art model proposed by Samadi et al. (ACM Trans Asian Low-Resour Lang Inf Process, <https://doi.org/10.1145/3472620>, 2021).

Keywords Fake news detection · Feature engineering · Contextualized text representation · Deep neural network

Mathematics Subject Classification Computing methodologies · Artificial intelligence · Natural language processing

1 Introduction

The controversial debates about disseminating fake news have been growing rapidly since social media have been an inseparable part of people's lives. In social media, each user can share various information with others; consequently, there is a high potential to produce, share, or redistribute

misinformation based on different intents. These days, most governments are concerned about spreading misinformation on social media among people since it can be used as a powerful tool for manipulating the general public's thoughts about different topics. In addition, spreading fake news can endanger people's physical and mental health when they are in a difficult situation. Such misinformation conceals healthy behaviors and promotes wrong practices, which increases the spread of the virus and leads to poor physical and mental health outcomes [50]. Similar to other events, such as the 2016 US presidential election, the COVID-19 pandemic showed the potential of spreading unauthentic information on social media. Such wrong information led to a wrong understanding of the virus; e.g., a fake video about COVID-19 indicated that wearing a mask will activate the virus [13].

✉ Saeedeh Momtazi
momtazi@aut.ac.ir

Mohammadreza Samadi
mhmd.samadi@aut.ac.ir

¹ Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran

Detecting fake news is a complex and vital task, which has been introduced with different definitions. Lazer et al. [24] defined the term “fake news” as a piece of fabricated information whose content’s structure is the same, but the intentions of dissemination are not. Also, they claimed that 9% to 15% of Twitter accounts are bots. Based on Facebook’s estimation, there are 60 million active bots on Facebook, most of whom were in charge of spreading misinformation during the 2016 US election. Rubin et al. [41] classified fake news into three categories: (1) fabrication, (2) hoaxing, and (3) satire. In this study, similar to most studies on fake news detection, we focus on the first category. Another challenge in the fake news detection task is that collecting a high-quality, comprehensive dataset is tricky; Shu et al. [49] tried to propose a solution to this problem with weak supervision.

In recent years, the great success of deep learning architectures like contextualized text representation language models in different tasks of Natural Language Processing (NLP) such as question answering Wang et al., [55], text chunking [27], named entity recognition [1], sentiment analysis [6,29,58], and semantic classification [25,34,35] motivated researchers to propose different transformer-based models for fake news detection. The models can extract high-level features based on the context of input texts. Before such architectures emerged, researchers tried extracting valuable features by using feature engineering techniques. They mainly extract different features from the input data to represent all aspects well. Although these feature engineering techniques have shown reasonable performance in the field, they received less attention by the advent of semantic representations. The proposed models by Kaliyar et al. [23]; Shishah [47]; Liu et al. [26]; Samadi et al. [43]; Goldani et al. [16]; Jwa et al. [22] are examples on these studies which only utilized semantic representation models for the task. The traditional approaches, however, have their own advantages that should be considered to achieve a high-performance system. One of these advantages is that selecting valuable features using ensemble methods can dominate state-of-the-art models [19].

In this paper, we aim to show that although deep contextualized models can extract high-level features from textual information related to the context of the input texts, for solving a complex problem like fake news detection, we still need to apply various feature engineering methods in order to pay attention to all aspects of news articles. We use four additional text processing modules to extract content-based features: (1) latent Dirichlet allocation (LDA) to extract major topics that each news includes, (2) a deep classification model for extracting the category of news articles, (3) a named entity recognition model for creating a vector that represents all named entities in each news article, and (4) a sentiment classifier for specifying that each news has negative or positive polarity. This approach helps us to analyze the impact of

content-based features besides semantic textual features. We use both types of features within a deep CNN framework. Also, this study focuses on two different languages from rich-resource and low-resource languages, namely English and Persian. For extracting semantic features with deep contextualized models, we utilize transformer-based models from the BERT [8] family, namely RoBERTa [28] and ParsBERT [10].

This paper is organized as follows: In Sect. 2, we review the related works on fake news detection. Section 3 describes the content-based features used in our model. Section 4 presents our approach for representing and processing news articles and the details of our models. In Sect. 5, we reported our experimental results, and finally, in Sect. 6, we conclude this paper and explain our future works.

2 Related works

Due to the complexity of fake news detection, researchers have been trying to extract a different set of features from texts to improve the performance of the automated models. The following subsections discuss previous works on manual feature engineering, semantic text representation, and hybrid models. We also review related studies on Persian fake news detection.

2.1 Feature engineering for fake news detection

Before the emergence of deep semantic contextualized models pretrained on a huge amount of textual data, researchers aimed to extract manual features for representing news articles from different aspects and using these features within traditional machine learning classifiers.

Pérez-Rosas et al. [40] proposed two fake news datasets in English. They analyzed fake news articles using a linear SVM classifier by extracting linguistic features such as words’ n-grams, punctuation, psycholinguistic features (e.g., words per sentence, part-of-speech categories), and readability (e.g., number of characters, paragraphs). They showed the effectiveness of manual feature engineering on the performance of machine learning classifiers by improving 9% in terms of accuracy compared to a model that uses merely syntactic features.

Janze and Risius [21] trained logistic regression, support vector machines, decision tree, random forest, and extreme gradient boosting classifiers and utilized a feature engineering approach for extracting different features in the following groups: (1) cognitive cues (e.g., textual features like having question mark signs), (2) visual cues (e.g., number of faces in image posts), (3) affective cues (e.g., emotional reactions to the posts), (4) behavioral cues (e.g., number of shares and

comments each post received). Finally, they calculated the effectiveness of each feature for detecting fake news.

Shu et al. [48] provided a survey on different approaches for fake news detection in social media, including fake news characterizations focusing on data mining algorithms. They discussed some unique characteristics of spreading fake news on social media. They suggested features for representing each post, including linguistic features (e.g., lexical, syntactic), visual features (e.g., clarity score, similarity distribution histogram), user-based features (e.g., number of followers/followings), topic-based features (e.g., topics extracted by latent Dirichlet allocation), and network-based features (e.g., the friendship network of users).

Dey et al. [9] performed linguistic analysis on a simple dataset including 200 tweets about “Hillary Clinton”. After categorizing and linguistic analysis, such as parts of speech tagging and named entity recognition, they utilized the K-nearest neighbor algorithm for classifying news.

Braşoveanu and Andonie [4] explored fake news detection using a hybrid approach utilizing machine learning techniques to extract valuable features such as named entity, named entity links, and semantic analysis. Besides, they discovered relations between the speaker name and the subject of the news using both parts of speech tagging and the DBpedia knowledge base.

Hakak et al. [19] utilized ensemble models such as decision tree, random forest, and extra tree classifiers as fake news detection models. They mainly focused on selecting valuable features and discerning fake and real news. After pre-processing, they extracted statistical features like word count, character count, sentence count, average word length, average sentence length, and named entity features. This study indicated the importance of appropriate feature selection and hyperparameter tuning.

To summarize, manual feature engineering has been used widely among researchers in the fake news detection task. Although researchers utilized different sources of information for extracting handcrafted features like user profiles or visual features, we extract content-based features, including named entities, sentiments, topics, and categories of news articles, as well as their semantic textual features.

2.2 Semantic text representation for fake news detection

After introducing neural word representation models, embedding vectors such as Word2vec [30] and GloVe [39] have received researchers’ attention in various tasks including fake news detection.

Wang [54] used Word2vec embedding vectors for representing news articles and proposed a hybrid CNN model for detecting fake news. Goldani et al. [15] utilized the novel capsule neural network architecture for detecting fake news

of different lengths. They used different settings of GloVe for representing news articles. In another study, Goldani et al. [16] proposed a CNN architecture with margin loss for fake news detection. They used the pretrained GloVe model as the embedding layer.

By introducing the transformer’s architecture [53], researchers proposed deep contextualized language models based on the architecture of the transformer. Devlin et al. [8] proposed the BERT model that significantly improved the performance of deep models in different NLP tasks. The advanced architecture of these models in capturing semantic features from text motivated researchers to use contextualized text representation for their tasks.

More specifically, in fake news detection, Liu et al. [26] proposed a two-stage approach for detecting fake news. In the first stage, they used BERT outputs for representing both textual information and metadata with a classifier that calculates a coarse-grained label for each news (e.g., fake or true). In the second stage, the BERT model encodes all previous information in addition to the predicted label in the first stage. The second classifier predicts the fine-grained label (e.g., barely true, half-true, mostly true, or true).

Jwa et al. [22] trained the BERT model on a large corpus of CNN and Daily Mail news data. The model is then used to represent news articles. They trained their classification using weighted cross-entropy.

Zhang et al. [60] proposed an end-to-end model called BDANN for multimodal fake news detection. They used textual and visual channels to extract features using BERT and VGG-19, and they evaluated the effectiveness of their model on Twitter and Weibo multimedia datasets.

Giachanou et al. [14] proposed a multimodal fake news detection system that utilized both textual and visual features. They connected VGG16 to an LSTM model for image representation. Also, they used $BERT_{base}$ as a deep contextual text representation. Finally, after concatenating all extracted feature vectors, the vector is fed into a multi-layer perceptron (MLP) for classification.

Samadi et al. [43] provided a comparative study on different contextualized text representation models, including BERT, RoBERTa, GPT2, and funnel transformer within single-layer perceptron (SLP), MLP, and CNN architectures.

After the COVID-19 outbreak, researchers proposed different methods to prevent spreading fake news and misinformation. Wani et al. [56] implemented different classification algorithms with distinct representations such as transformer-based models like BERT, DistilBERT, and COVID-TwitterBERT [33] to detect fake news. Shifath et al. [46] utilized deep contextual models like BERT, GPT-2, RoBERTa, DistilRoBERTa, and fine-tuned them by COVID-19 corpus. Also, they combined different contextualized models to create ensemble models for detecting fake news.

Therefore, previous studies illustrated the effectiveness of deep contextualized language models for the fake news detection task to extract rich semantic representations for tokens or sentences. This study aims to utilize deep contextualized language models for extracting semantic features from the input news.

2.3 Hybrid approaches

In some research studies, the combination of semantic text representation with content features has been considered. Sabeeh et al. [42] proposed a two-step model which uses BERT and LDA topic modeling for fake news detection. Gautam et al. [11] also used XLNet and LDA representation for fake news detection. Gölo et al. [17] proposed using multimodal variational autoencoder for fake news detection in one-class learning. Their proposed model uses text embeddings and topic information to represent news articles.

However, these studies benefit from only one type of feature, topic modeling in most cases, and do not use different kinds of content-based features. Moreover, they limited their research to English, a rich-resource language that does not face the challenges of feature engineering on low-resource languages.

2.4 Persian fake news detection

In contrast to English, researchers have done numerous studies on detecting fake news; in Persian, there are a few studies on fake news detection using feature engineering.

Zamani et al. [59] crawled 783 Persian rumors from two Iranian websites and added equal numbers of randomly selected non-rumor tweets. They created a user graph using metadata based on the relations and network-based factors like page rank and clustering coefficient. Moreover, they added user-specific features and concatenated all metadata to textual features.

Jahanbakhsh-Nagadeh et al. [20] examined linguistic features on the rumor detection task, and they believe that the emotion of news, number of sensitive adverbs, and ambiguous words caused differences between real news and fake news.

Samadi et al. [44] created a Persian fake news dataset crawled from news agencies and proposed two architectures, BERT-CNN and BERT-SLP, for detecting fake news.

As can be seen from the related literature, both content-based features and high-level semantic features have been studied in misinformation analysis. Considering the advantages of the models, however, no work tried to benefit from both approaches. Traditional machine learning models only focus on intensive feature engineering, and recent deep learning models only work based on semantic text representation. This gap between traditional machine learning and deep

learning approaches motivated us to propose a model to benefit from the advantages of both sides within a deep learning framework for fake news detection.

Moreover, to the authors' best knowledge, this is the first study that benefits from state-of-the-art contextualized representation with a deep learning model while using different kinds of content features and applying them to two different languages.

3 Content-based features

In this study, we focus on obtaining high-level features from news articles, all of which, besides the semantic features, can assist the neural classifier in predicting a news label accurately. For extracting text-based features, we utilize deep contextualized text representation models. Furthermore, other valuable features in news articles provide information beyond the contextual characteristics.

The first feature that can be useful for discerning fake and real news is the category of the articles. We propose capturing topic-related features from news texts. We utilize a topic modeling approach and a deep neural classifier for categorizing each news. We hypothesize that some specific topics/categories may deal with more fake news than other topics/categories. To evaluate this hypothesis, we use this information as additional evidence to decide if a news article is fake or real.

The occurrence of named entities in the text is another potential feature that can help us recognize fake and real news better. We hypothesize that it is more likely to have fake news about persons, such as celebrities, than other entities, such as organizations. To this end, we use named entity recognition and extract the entity labels of each text.

Another feature that can provide us with information about a piece of news is its sentiment which is an important parameter revealing the intentions of its producer. We hypothesize that it is more likely to have fake news among negative texts rather than positive texts. To evaluate this hypothesis, we use a deep neural model to obtain the polarity of texts.

This paper shows that this extra information helps detect fake news and misinformation in different concepts and languages. To this aim, we use two datasets in English and Persian in different domains. The following subsections review the details of the modules we use for content-based feature engineering.

3.1 Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA) [3] provides a vector representation for each document based on the distribution of each document over different topics. This model has been used in various information retrieval and NLP tasks, including

sentiment analysis [36], publication analysis [7], and query suggestion [31]. Considering the unsupervised nature of this approach, no training data is required for this module, and we use the same approach for both English and Persian experiments.

3.2 News category classification

To extract the category of each news article, we train a deep neural classifier. We use the RoBERTa-CNN model to detect English categories based on the RoBERTa representation and the CNN classifier. The BBC news dataset [18] is used for training the classifier, which includes news articles in 5 categories: business, entertainment, politics, sport, and technology.

We benefit from a BERT-SLP model for the Persian news that includes an SLP connected after the BERT embedding layer. In this architecture, we use multilingual BERT to represent Persian news and train it on the Hamshahri news corpus [2]. Each news has a label in 82 fine-grained categories, but the coarse-grained labels include five categories: social, cultural, politics&economy, science, and sport used for training our model.

3.3 Sentiment analysis

Similar to news category classification, we benefit from deep neural classifiers for capturing the sentiment of the text. For classifying English news based on their sentiments, we utilized the pretrained DistilBERT [45] fine-tuned on the Stanford Sentiment Treebank (SST2) dataset.

For Persian news, we train a deep context-sensitive neural model called XLM-RoBERTa-CNN by connecting three parallel convolutional layers after the sequenced output of XLM-RoBERTa for extracting the sentiments of the Persian dataset with either positive or negative polarity. To this aim, we use the Ghasemi et al. [12]’s dataset.

3.4 Named entity recognition

For named entity recognition on English news, we utilize a pretrained HuggingFace model for named entity recognition trained on the CoNLL-2003 dataset [51], which includes five entity types, namely person, location, organization, miscellaneous, and other.

For Persian news, following Abdollah Pour and Momtazi [1] we use a conditional random field layer connected to XLM-RoBERTa for annotating and extracting Persian named entities in 16 categories. The list of named entity labels in our Persian data is as follows, which are extracted by training on the MoNa dataset [32]: person individual, person group, location, organization, language, nationality, event, job, book, film, date, religion, field, magazine, and other.

4 Proposed model

Our proposed model includes a contextualized pretrained model for extracting semantic features from news articles, several modules for capturing content-based features, and a neural classifier for classifying fake and real news.

4.1 Text representation

Various methods have been proposed for text representation [52]. In our proposed model, we use contextualized text representation, the state-of-the-art representation in the field.

Devlin et al. [8] introduced the BERT model based on the transformer architecture. They trained it on a large corpus of Wikipedia and BooksCorpus with different approaches of masked language model and next sentence prediction. After introducing the BERT model, Liu et al. [28] claimed it was not completely trained. Therefore, they proposed the RoBERTa model with the same architecture as BERT and trained it on a much larger raw corpus without using the next sentence prediction learning method. We use RoBERTa for our English experiments.

ParsBERT [10] is a deep contextualized model that is specifically trained on Persian raw corpora such as Wikipedia, BigBangPage, Chetor, Eligasht, Digikala, Ted Talks subtitles, several fictional books and novels, and Miras-Text. We utilize ParsBERT for representing Persian news articles.

4.2 Classification with content-based features

Convolutional neural networks have been used widely in different domains such as computer vision and natural language processing. This architecture’s main advantage is extracting high-level features using trainable filters. In our proposed model, the sequenced output of the text representation module is connected to three parallel groups of convolutional and max-pooling layers. Each part includes a convolutional layer followed by a max-pooling layer to extract features based on the collocation of tokens in news articles.

The sequenced-output is a $n \times d$ matrix including n token vectors with d dimension, presented as “Semantic Contextualized Vector Representation” in Figs. 1 and 2. The convolutional layers can extract high-level features with different levels of the n-gram. In the convolutional layer, we benefit from 30 kernels with different sizes of 3, 4, and 5 applied on the sequenced-output matrix. The kernel sizes help to capture tri-gram, 4-gram, and 5-gram models. The obtained features are downsampled using max-pooling layers. After extracting textual features using CNN, we fed them to a flattened layer.

In the case of using no content-based feature, the output of the flattened layer is fed to a fully connected layer, and the

Fig. 1 The architecture of our proposed model with semantic features

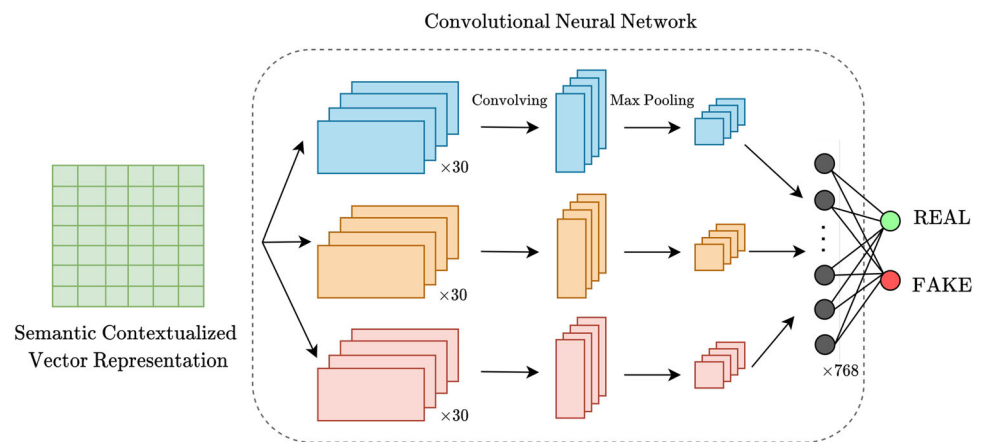
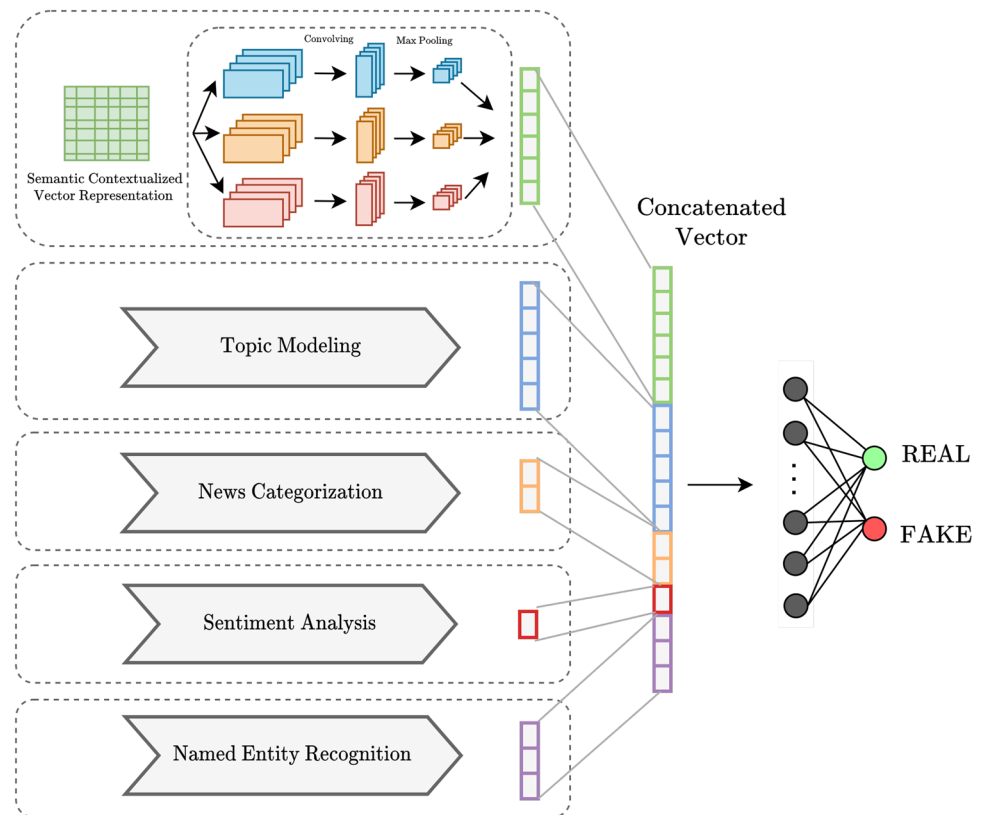


Fig. 2 The architecture of our proposed model with both content-based and semantic features



probability of belonging to each class is calculated. Figure 1 shows the architecture of our model for news classification based on semantic features only.

In the case of using content-based features, the output of the flatten layer is concatenated to the content-based feature vectors that were explained in Section 3. Finally, for each news, a feature vector, consisting of semantic representation from the contextualized text representation model, topics, categories, named entities, and the sentiment, is fed into fully connected layers for calculating the loss and predicting the label. In this architecture, the content-based features

can include the output of one or several modules from Sect. 3. Figure 2 shows the architecture of our model in detail.

5 Experimental results

5.1 Datasets

To evaluate the effectiveness of applying content-based features besides extracting contextualized semantic features from news articles for fake news detection, we use two completely different datasets.

Table 1 The statistics of the English COVID-19 fake news dataset introduced by Patwa et al. [38] and the Persian TAJ dataset

Split	Dataset			
	COVID-19		TAJ	
	Real	Fake	Real	Fake
Training	3360	3060	1294	1311
Validation	1120	1020	367	378
Test	1120	1020	200	171

We evaluate our approach on a Persian fake news dataset, called TAJ [44], which includes 1,860 fake news and 1,860 true news in a diverse range of topics from real-world sources. Moreover, we test our approach on an English dataset to show that the proposed model does not depend on a specific language. Considering the advent of fake news detection on social media about the COVID-19 pandemic, we use the COVID-19 dataset [38] for our English experiments. The data contain 5,600 real and 5,100 fake posts related to the pandemic, which were spread on Twitter. Table 1 shows the statistics of the two datasets.

5.2 Settings

For implementing the proposed model, we use Keras [5] and huggingface/transformers [57] libraries for implementing deep neural models. We trained our fake news detection model with a learning rate of $5e - 5$, four epochs, and 64 tokens as the maximum length of input news articles. In order to prevent models from overfitting the training data, we apply the L2 kernel regularization term to all convolutional layers.

Table 2 The results of evaluating our model with different content-based features on the Persian TAJ dataset

Metadata	Precision	Recall	F1-score	Accuracy
—	89.13	93.71	91.36	91.64
Latent Dirichlet allocation	91.42	91.42	91.42	91.91
News categorization	96.87	88.57	92.53	93.26
Sentiment analysis	91.52	92.57	92.04	92.45
Named entity recognition	91.16	94.28	92.69	92.99
All features	93.64	92.57	93.10	93.53

Table 3 The results of evaluating our model with different content-based features on the English COVID-19 dataset

Metadata	Precision	Recall	F1-score	Accuracy
—	88.85	99.31	93.79	93.73
Latent Dirichlet allocation	97.63	97.15	97.39	97.52
News categorization	96.88	94.61	95.73	95.98
Sentiment analysis	97.78	95.29	96.52	96.73
Named entity recognition	94.94	97.55	96.22	96.35
All features	97.19	98.43	97.81	97.89

The parameters have been selected based on the experiments on the validation set.

For the LDA model, we use gensim's LDA model by training it on the datasets with 2000 iterations and extracting 50 topics for each news article represented by a 50-dimensional vector.

5.3 Results and discussion

In our experiments, we want to show the effect of each content-based feature besides the semantic features extracted by deep contextualized models. First, we evaluated our approach to the Persian dataset as a low-resources language. The results of our experiments on the TAJ dataset are reported in Table 2.

As can be seen from the tabulated results, using each of the mentioned features improved the model results. Category features and named entities achieve the best improvements. Since the Persian dataset includes general news from different sources, the news articles' category feature improved accuracy and f1-score by 1.6% and 1.2%, respectively, which is related to the fact that the number of fabricated news articles is not balanced in different categories. Named entity features also boosted accuracy and f1-score by 1.35% and 1.33%, respectively. This indicates that appearing named entities in real and fake news follow different patterns, and this feature can help to detect fake news better. Moreover, compared to the model, which uses semantic features only, we achieve 0.81% and 0.3% improvement in accuracy by sentiment and topics features, respectively. Ultimately, by concatenating all feature vectors and feeding a single vector to the fully connected layer, the f1-score and accuracy increased 1.74% and 1.89%, respectively.

We believe that the effectiveness of feature engineering is not related to the language or domains of text articles. To this aim, we also evaluated this approach on an English dataset and reported the results in Table 3. Based on the obtained results, the extracted information using the LDA topic modeling approach achieves the best improvement in accuracy and f1-score. It is because the news articles in the COVID-19 dataset are about the pandemic, and LDA provides extra information about different aspects of each news; e.g., topics related to the impact of COVID-19 on economics are more prone to distribute misinformation rather than topics about preventing the disease. The topical content improved the accuracy and f1-score of our model by 3.79% and 3.6%, respectively. Also, named entities, news categories, and sentiment feature improved accuracy by 2.6%, 2.2%, and 3%, respectively.

In the next step of our experiments, we use all four features in addition to the RoBERTa. The results of this experiment show that all features together boosted the model accuracy and f1-score by 4.16% and 4.01%, respectively.

To better understand how content features affect the model's performance, we present the distribution of different content features in our datasets. Figure 3 shows the distribution of all entities in the COVID-19 datasets. As can be seen in this figure, named entities with organization and location labels are more pronounced in real news, while person and misc labels appear more in fake news. This indicates that fake news articles are more about people rather than locations or organizations; i.e., the number of fake news which includes a person's name is much larger than the number of real news with a person's name. Having 5792 news with person's names, 4710 out of 5792 news (81%) are fake. However, considering ORG and LOC labels, the number of real news that includes those named entities is much more than the number of fake news. Having 19,531 news with the organization's names, only 6651 out of 19,531 (34%) are fake. Having 16,895 news with the location's names, only 4584 out of 16,895 (27%) are fake.

Figure 4 shows the distribution of categories in each class of the TAJ dataset, showing how categorical information can help us in the prediction phase.

Moreover, by comparing different topics and their occurrence in real and fake news, we can capture topics that are more pronounced in real news and those that occur more in fake news. Tables 4 and 5 show the set of words representing each topic id in the COVID-19 and TAJ datasets, respectively. The label column shows that the topic is more pronounced in which part of the data, real or fake. The extracted topics using LDA are meaningful in both datasets; e.g., in Table 4, words in Topic 1 carry important information about preventing the pandemic by wearing a mask, which is mainly associated with real news. In contrast, Topic 0 is about the different views of workers, races about the pandemic, and

black people, leading to the spreading of fake news on social media. Likewise, in Table 5, Topic 31 is about celebrities exposed to rumors, while Topic 7 shows valid information reported by Police offices.

To summarize, Fig. 5 presents the improvement of using content-based features extracted by the feature engineering modules compared to utilizing only semantic features obtained by contextualized text representation models, known as the state-of-the-art model in the field.

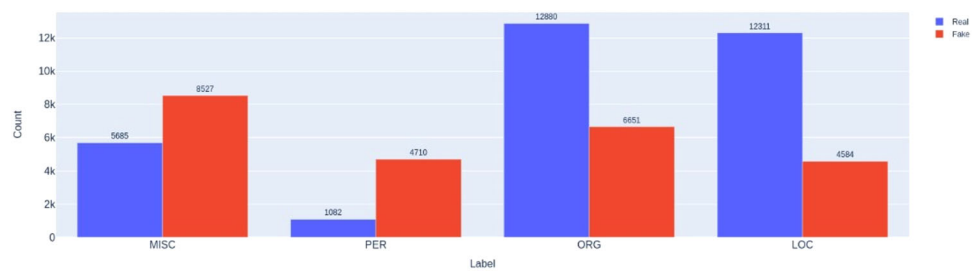
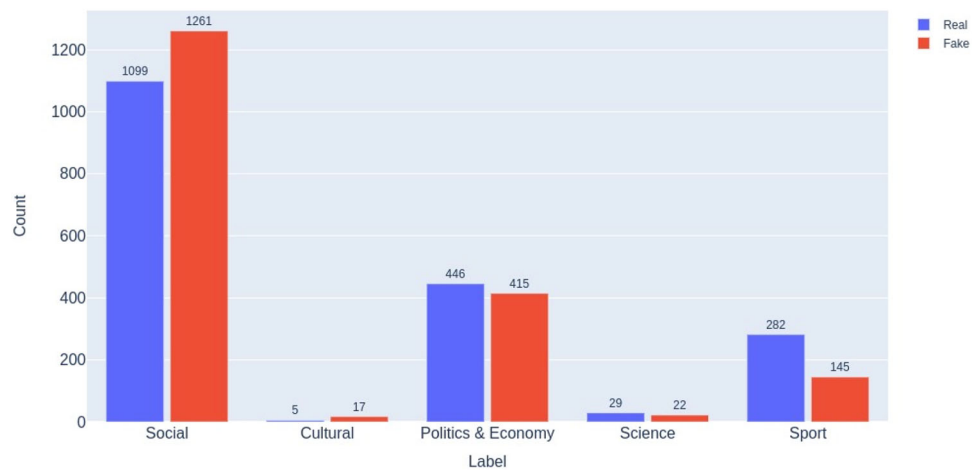
In order to substantiate our claim about the positive effects of feature engineering besides using deep contextualized models, we compared our results with the state-of-the-art models in the field as presented in Table 6. We also presented the results of conventional machine learning models to show how deep learning approaches improved the performance of models and how adding content-based features on top of deep learning can further improve the performance.

As can be seen in the tabulated results, deep models outperformed conventional machine learning models. Moreover, our proposed model achieved a 2.13% improvement compared to the model proposed by Samadi et al. [44]. Also, comparing our approach with the proposed models by Pathwar and Gill [37], Shifath et al. [46], and Wani et al. [56], we achieved 4.02%, 2.01%, and 2.1% improvement, respectively. It should be mentioned that the differences between our model and the proposed model by Samadi et al. [44] on both f1-score and accuracy are statistically significant according to the two-tailed t -test (p -value < 0.05). The difference between the accuracy of our model and all three baseline models on the COVID-19 dataset is also statistically significant according to the two-tailed t -test (p -value < 0.05); the difference on f1-score is statistically significant compared to Pathwar and Gill [37]'s, and Wani et al. [56]'s results.

Although our experiments show that using content-based features improves the performance of the task, in some cases, such features may have a negative impact. The following texts are examples of real news in the dataset that our model detects as fake news. Both news is labeled as *negative* and categorized as *political news*, which both have higher distribution in fake news. These features, together with the textual information of the news articles, caused our model to label them as fake, which is incorrect.

news example 1: "A small fraction of deaths in long-term care facilities are staff members. Nonetheless, researchers estimate that COVID-19 will make working in an LTC facility the most dangerous job in America by year's end in 2020."

news example 2: "Independent SAGE adviser' withdraws lockdown claim - as the UK records highest #coronavirus daily cases since May."

Fig. 3 The distribution of named entities in each class for the COVID-19 dataset**Fig. 4** The distribution of different categories in each class for the TAJ dataset**Table 4** List of words and their label in LDA topics for the COVID-19 dataset

Topic id	Topic words	Label
Topic 0	Pandemic, media, important, workers, american, woman, black, social, communities, leader	Fake
Topic 8	Taking, 30, white, registered, delhi, spreading, followed, left, muslim, sir	Fake
Topic 1	Spread, may, help, mask, masks, protect, public, others, wear, wearing	Real
Topic 3	Symptoms, medical, discharged, announced, deaths, covid19nigeria, cases, new, support, confirmed	Real

Table 5 List of words and their label in LDA topics for the TAJ dataset

Topic id	Topic words	Label
Topic 21	Syria, forces, US, news agency, report, terrorist, Korea, newspaper, AD, Lebanon	Fake
Topic 31	Film, show, program, historical, network, Nowruz, cinema, days, artists, directing	Fake
Topic 7	Police, manual, police, discovery, crescent, red, colonel, smuggling, Nowruz, accused	Real
Topic 48	Work, government, hint, create, continue, economic, conditions, emphasis, problems, attention	Real

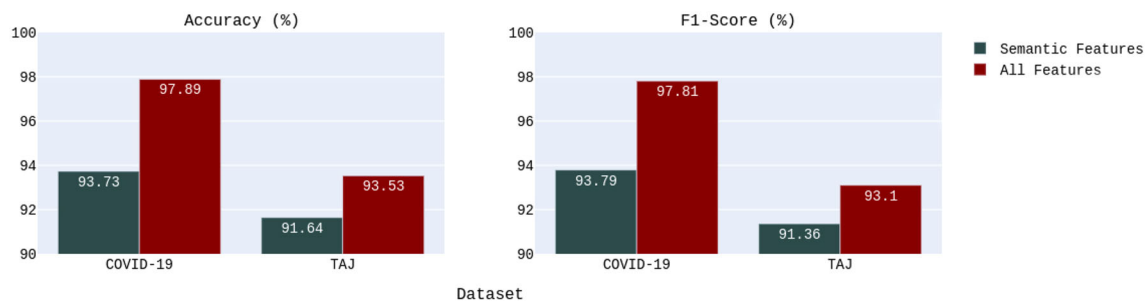
**Fig. 5** Comparison of accuracy and f1-score by using different features

Table 6 Comparison of our feature enhanced model with previous methods on TAJ and COVID-19 datasets

Model	Dataset	Precision	Recall	F1-score	Accuracy
ML (DecisionTree)	TAJ	71.12	77.77	74.30	75.20
ML (Logestic Regression)		75.77	85.96	80.54	80.86
ML (Support Vector Machine)		79.79	90.05	84.61	84.90
ML (Gradient Boosting)		80.95	89.47	85.00	85.44
[44]		90.40	92.63	91.50	91.40
[37]		94.03	94.28	94.15	93.87
Feature Enhanced Model (ours)		93.64	92.57	93.10	93.53
ML (DecisionTree)	COVID-19	85.47	85.37	85.39	85.37
ML (Logestic Regression)		92.01	91.96	91.96	91.96
ML (Support Vector Machine)		93.33	93.32	93.32	93.32
ML (Gradient Boosting)		87.24	86.96	86.96	86.96
[56]		98.94	92.15	95.43	95.79
[46]		96.49	96.68	97.12	95.88
Feature Enhanced Model (ours)		97.19	98.43	97.81	97.89

6 Conclusion

In this paper, we claimed that although deep contextualized text representation models have been a great success story in different tasks of NLP in recent years, we still need to utilize feature engineering methods to capture different features based on different aspects of texts. To substantiate this issue, we used semantic features extracted by contextualized models and four content-based features captured by different text processing tasks, including topic modeling, news categorization, sentiment analysis, and named entity recognition. Also, we showed that manual feature engineering is not limited to the domain of input texts or their language. We evaluated our approach in different experiments on a general Persian fake news dataset and a domain-specific English dataset. Our results showed that content-based feature engineering is still an essential part of fake news detection besides the semantic features and can assist our model in detecting fake and real news more accurately.

Since most fake news is spread on social media, additional sources of information can be used besides textual information to improve the performance of an automated model for future work. One of the informative sources that can be considered in future work is the users' graph on social media, which indicates how users are connected and how they spread information.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Abdollah Pour, M.M., Momtazi, S.: A comparative study on text representation and learning for persian named entity recognition. *ETRI* (2022)
2. AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F.: Hamshahri: a standard persian text collection. *Knowl.-Based Syst.* **22**(5), 382–387 (2009)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res* **3**(1), 993–1022 (2003)
4. Braşoveanu, A.M., Andonie, R.: Integrating machine learning techniques in semantic fake news detection. *Neural Process. Lett.* 1–18 (2020)
5. Chollet, F., et al.: Keras. <https://keras.io> (2015)
6. Dai, A., Hu, X., Nie, J., Chen, J.: Learning from word semantics to sentence syntax by graph convolutional networks for aspect-based sentiment analysis. *Int. J. Data Sci. Anal.* **14**(1), 17–26 (2022)
7. Danesh, F., Dastani, M., Ghorbani, M.: Retrospective and prospective approaches of coronavirus publications in the last half-century: a latent Dirichlet allocation analysis. *Library Hi Tech* **39**(3), 855–872 (2021)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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*, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186, <https://doi.org/10.18653/v1/N19-1423> (2019)
9. Dey, A., Rafi, R.Z., Parash, S.H., Arko, S.K., Chakrabarty, A.: Fake news pattern recognition using linguistic analysis. In: *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 305–309. *IEEE* (2018)
10. Farahani, M., Gharachorloo, M., Farahani, M., Manthouri, M.: Parsbert: transformer-based model for persian language understanding. *arXiv preprint arXiv:2005.12515* (2020)
11. Gautam, A., Venkatesh, V., Masud, S.: Fake news detection system using xlnet model with topic distributions: Constraint@aaai2021 shared task. In: Chakraborty, T., Shu, K., Bernard, H.R., Liu, H., Akhtar, M.S. (eds.) *Combating Online Hostile Posts in Regional*

- Languages during Emergency Situation, pp. 189–200. Springer, Cham (2021)
12. Ghasemi, R., Asl, A.A., Momtazi, S.: Deep Persian sentiment analysis: cross-lingual training for low-resource languages. *J. Inf. Sci.* (2020)
 13. Ghayoomi, M., Mousavian, M.: Deep transfer learning for covid-19 fake news detection in Persian. *Exp. Syst.* (2022). <https://doi.org/10.1111/exsy.13008>
 14. Giachanou, A., Zhang, G., Rosso, P.: Multimodal multi-image fake news detection. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 647–654. IEEE (2020)
 15. Goldani, M.H., Momtazi, S., Safabakhsh, R.: Detecting fake news with capsule neural networks. *Appl. Soft Comput.* **101**, 106991 (2021). <https://doi.org/10.1016/j.asoc.2020.106991>
 16. Goldani, M.H., Safabakhsh, R., Momtazi, S.: Convolutional neural network with margin loss for fake news detection. *Inf. Process. Manag.* **58**(1), 102418 (2021)
 17. Gôlo, M., Caravanti, M., Rossi, R., Rezende, S., Nogueira, B., Marcacini, R.: Learning textual representations from multiple modalities to detect fake news through one-class learning. In: Proceedings of the Brazilian Symposium on Multimedia and the Web, Association for Computing Machinery, New York, NY, USA, WebMedia'21, pp. 197–204, <https://doi.org/10.1145/3470482.3479634> (2021)
 18. Greene, D., Cunningham, P.: Practical solutions to the problem of diagonal dominance in kernel document clustering. In: Proceedings of 23rd International Conference on Machine Learning (ICML'06), ACM Press, pp. 377–384 (2006)
 19. Hakak, S., Alazab, M., Khan, S., Gadekallu, T.R., Maddikunta, P.K.R., Khan, W.Z.: An ensemble machine learning approach through effective feature extraction to classify fake news. *Futur. Gener. Comput. Syst.* **117**, 47–58 (2021)
 20. Jahanbakhsh-Nagadeh, Z., Feizi-Derakhshi, M.R., Ramezani, M., Rahkar-Farshi, T., Asgari-Chenaghlu, M., Nikzad-Khasmakhi, N., Feizi-Derakhshi, A.R., Ranjbar-Khadivi, M., Zafarani-Moattar, E., Balafar, M.A.: A model to measure the spread power of rumors. *arXiv preprint arXiv:2002.02020* (2020)
 21. Janze, C., Risius, M.: Automatic detection of fake news on social media platforms. In: PACIS, p. 261 (2017)
 22. Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H.: exbake: automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). *Appl. Sci.* **9**(19), 4062 (2019)
 23. Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools Appl.* **80**(8), 11765–11788 (2021)
 24. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., et al.: The science of fake news. *Science* **359**(6380), 1094–1096 (2018)
 25. Lin, S., Wu, X., Chawla, N.V.: motif2vec: semantic-aware representation learning for wearables' time series data. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 1–10 (2021)
 26. Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., Lu, X.: A two-stage model based on bert for short fake news detection. In: International Conference on Knowledge Science, Engineering and Management, pp. 172–183. Springer (2019)
 27. Liu, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., Zhou, J.: GCDT: a global context enhanced deep transition architecture for sequence labeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 2431–2441, <https://doi.org/10.18653/v1/P19-1233> (2019b)
 28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: a robustly optimized BERT pretraining approach. [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019c)
 29. Liu, Z., Wang, J., Du, X., Rao, Y., Quan, X.: Gsmnet: global semantic memory network for aspect-level sentiment classification. *IEEE Intell. Syst.* **36**(5), 122–130 (2020)
 30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
 31. Momtazi, S., Lindenberg, F.: Generating query suggestions by exploiting latent semantics in query logs. *J. Inf. Sci.* **42**(4), 437–448 (2016). <https://doi.org/10.1177/0165551515594723>
 32. Momtazi, S., Torabi, F.: Named entity recognition in Persian text using deep learning. *Signal Data Process.* **16**(4), 93–112 (2020)
 33. Müller, M., Salathé, M., Kummervold, P.E.: Covid-twitter-BERT: a natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* (2020)
 34. Munikar, M., Shakyia, S., Shrestha, A.: Fine-grained sentiment classification using BERT. In: 2019 Artificial Intelligence for Transforming Business and Society (AITB), vol. 1, pp. 1–5. IEEE (2019)
 35. Oliveira, S., Loureiro, D., Jorge, A.: Improving Portuguese semantic role labeling with transformers and transfer learning. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp. 1–9 (2021)
 36. Ozyurt, B., Akcayol, M.A.: A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: Ss-Ilda. *Expert Syst. Appl.* **168**, 114231 (2021). <https://doi.org/10.1016/j.eswa.2020.114231>
 37. Pathwar, P., Gill, S.: Tackling covid-19 infodemic using deep learning. *arXiv preprint arXiv:2107.02012* (2021)
 38. Patwa, P., Sharma, S., PYKL, S., Guptha, V., Kumari, G., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Fighting an infodemic: Covid-19 fake news dataset. In: Proceedings of the CONSTRAINT-2021 workshop, co-located with the AAAI'21 conference (2021)
 39. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
 40. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3391–3401 (2018)
 41. Rubin, V.L., Chen, Y., Conroy, N.K.: Deception detection for news: three types of fakes. *Proc. Assoc. Inf. Sci. Technol.* **52**(1), 1–4 (2015)
 42. Sabeeh, V., Zohdy, M., Al Bashair, R.: Fake news detection through topic modeling and optimized deep learning with multi-domain knowledge sources. In: Stahlbock, R., Weiss, G.M., Abou-Nasr, M., Yang, C.Y., Arabnia, H.R., Deligiannidis, L. (eds.) *Advances in Data Science and Information Engineering*, pp. 895–907. Springer, Cham (2021)
 43. Samadi, M., Mousavian, M., Momtazi, S.: Deep contextualized text representation and learning for fake news detection. *Inf. Process. Manag.* **58**(6), 102723 (2021). <https://doi.org/10.1016/j.ipm.2021.102723>
 44. Samadi, M., Mousavian, M., Momtazi, S.: Persian fake news detection: neural representation and classification at word and text levels. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* <https://doi.org/10.1145/3472620> (2021)
 45. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. In: Proceedings of the EMC2 Workshop, Co-located with NeurIPS'19 Conference (2019)

46. Shifath, S., Khan, M.F., Islam, M. et al.: A transformer based approach for fighting covid-19 fake news. arXiv preprint [arXiv:2101.12027](https://arxiv.org/abs/2101.12027) (2021)
47. Shishah, W.: Fake news detection using BERT model with joint learning. Arab. J. Sci. Eng. 1–13 (2021)
48. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
49. Shu, K., Dumais, S., Awadallah, A.H., Liu, H.: Detecting fake news with weak social supervision. IEEE Intell. Syst. **36**(4), 96–103 (2020)
50. Tasnim, S., Hossain, M.M., Mazumder, H.: Impact of rumors and misinformation on covid-19 in social media. J. Prevent. Med. Public Health (2020)
51. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003—vol. 4, Association for Computational Linguistics, USA, CONLL'03, p 142-147, <https://doi.org/10.3115/1119176.1119195> (2003)
52. Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., Gravier, G.: A survey on training and evaluation of word embeddings. Int. J. Data Sci. Anal. **11**, 85–103 (2021)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
54. Wang, W.Y.: “liar, liar pants on fire”: a new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, pp. 422–426. <https://doi.org/10.18653/v1/P17-2067> (2017)
55. Wang, Z., Ng, P., Ma, X., Nallapati, R., Xiang, B.: Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp. 5878–5882. <https://doi.org/10.18653/v1/D19-1599> (2019)
56. Wani, A., Joshi, I., Khandve, S., Wagh, V., Joshi, R.: Evaluating deep learning approaches for covid19 fake news detection. arXiv preprint [arXiv:2101.04012](https://arxiv.org/abs/2101.04012) (2021)
57. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, pp. 38–45 (2020)
58. Yang, T., Yao, R., Yin, Q., Tian, Q., Wu, O.: Mitigating sentimental bias via a polar attention mechanism. Int. J. Data Sci. Anal. **11**(1), 27–36 (2021)
59. Zamani, S., Asadpour, M., Moazzami, D.: Rumor detection for Persian tweets. In: 2017 Iranian Conference on Electrical Engineering (ICEE), pp. 1532–1536. <https://doi.org/10.1109/IranianCEE.2017.7985287> (2017)
60. Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., Cui, L.: Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.