Social Network Analysis and Mining (2021) 11:39
https://doi.org/10.1007/s13278-021-00738-y

ORIGINAL ARTICLE

A transformer-based architecture for fake news classification

Divyam Mehta[1] · Aniket Dwivedi[1] · Arunabha Patra[1] · M. Anand Kumar[1]

Received: 24 June 2020 / Revised: 2 February 2021 / Accepted: 15 February 2021 / Published online: 17 April 2021

Abstract

In today's post-truth world, the proliferation of propaganda and falsified news poses a deadly risk of misinforming the public on a variety of issues, either through traditional media or on social media. Information people acquire through these articles and posts tends to shape their world view and provides reasoning for choices they take in their day to day lives. Thus, fake news can definitely be a malicious force, having massive real-world consequences. In this paper, we focus on classifying fake news using models based on a natural language processing framework, Bidirectional Encoder Representations from Transformers, also known as BERT. We fine-tune BERT for specific domain datasets and also make use of human justification and metadata for added performance in our models. We determine that the deep-contextualizing nature of BERT is effective for this task and obtain significant improvement over binary classification, and minimal yet important improvement in six-label classification in comparison with previously explored models.

Keywords Natural language processing · Deep learning · Fake news classification · BERT

1 Introduction

Social media has become an integral part of our lives in recent times. People are increasingly reliant on social media as a news source, and it was also found that television has been dethroned by social media as the primary news source throughout the world (Wakefield 2016). The Pew Research Center Lichterman (2016) found that 44% of Americans get their news from Facebook. Due to the relative ease, lower cost and effective spreading opportunities through the Internet, it is all too easy for people who are interested in advertising their news articles. It is no surprise that the accuracy of the news we see online is at times, questionable, in spite of its popularity and leads to the rapid spread of online misinformation (Vis 2014). Fake news refers to verifiable, untrue and misleading news which are propagated in any medium by malicious individuals for the intent of influencing people's behaviour. The uncontrolled dissemination of fake news through the Internet can have disastrous results on individuals as well as society as a whole. A Vox study (Resnick 2018) found that 'Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information'. It also found that 'the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information'. Fake news may be politically motivated—as in the case of the US presidential election of 2016, where the widespread reach of the fake news resulted in 871,000 likes, comments and shares on Facebook, which was significantly greater than other election-related real news circulating at the time and played an undeniable effect on the result of the election (Silverman 2016). Fake news is often targeted towards the consumers to misleading them for financial gain, for selling fraudulent products or to defame certain brands, and affects the trust and reliability of the consumers on the brand (Chen and Cheng 2020). Fake news is frequently used to manipulate the behaviour of the stock market and trades in an unethical way (Rapoza 2017). Given the influence of news, therefore, it should come as no surprise that a large number of people consider that fake news can be used as a potential weapon (Handley 2018).

✉ M. Anand Kumar
m_anandkumar@nitk.edu.in

Divyam Mehta
divyammehta@yahoo.co.in

Aniket Dwivedi
aniket.dwivedi1299@outlook.com

Arunabha Patra
a.patra5678@gmail.com

[1] Information Technology Department, NITK Surathkal, Mangalore 575025, India

It is, therefore, imperative to develop tools to detect and control the spread of fake news. To remit the impact of fake news on society, personnel are employed by few organizations to conduct a manual fact-checking procedure to verify claims; however, this process is laborious. The problem of detecting fake news accurately presents itself with several challenges—firstly, that fake news is often intentionally written to mislead viewers, so it can be difficult to conclude that a piece of news is fake without fact-checking the content (Shu et al. 2017). Secondly, the kind of fake news can be very varied, ranging across a wide array of topics (Shu et al. 2017). Thirdly, when users interact with fake news content, the data generated can become large, unstructured and noisy—making it difficult to make conclusions from it (Tang et al. 2014). For example, Bourgonje et al. (2017) focused on using headlines for detecting stances of news paragraphs, with respect to fake news especially clickbait titles.

In this work, we attempt to detail and address the various concepts related to transfer learning and propose an architecture to classify fake news. Approaches focusing on text classification using word embeddings like word2vec (Mikolov et al. 2013) and Glove (Bojanowski et al. 2017) seem promising, however, leave room for improvement because they employ fixed vector values rather than context or usage-based vector values for the words. To build upon this, we use contextual word embeddings, because the context of the events is important while judging the authenticity of the news. This is done in language models like ELMo and BERT, which improved performance in a variety of NLP tasks. BERT is built upon ELMo and is the first deeply bidirectional, unsupervised language representation. Wang (2017) showed effective classification of hyperpartisan news using BERT and ELMo, with BERT outperforming ELMo significantly. Considering the above points, we decided to go ahead and implement our models with BERT as the base architecture.

We explored both binary as well as multi-label classification using our proposed model. Using BERT requires nominal pre-processing and we have tried two different versions of BERT in this paper. In our experiments, we observed significant improvement in binary classification metrics and minor yet crucial improvement with multi-label classification.

2 Related work

Fake news detection has gained interest in recent times due to the easy availability of online news from social media. The problem is usually approached either from a text classification perspective like (Volkova et al. 2017; Ahmed et al. 2017), where they use traditional machine learning techniques to the statements of the article to classify them as fake or real. Zhang et al. (2018) proposed a method based on TF-IDF and word2vec embeddings for detection of fake news text classification. However, the performance of this approach was found to vary across datasets, possibly because word2vec has context-blind fixed vector values, suggesting a room for improvement by using context-sensitive word embeddings.

This problem could also be approached from a propagation-based perspective where focusing on the features of the social media outlet itself, such as likes, dislikes or retweets, leads to detection of clickbait content. The use of logistic regression and crowd sourcing algorithms has been promising towards the detection of fake news without considering the news content (Eugenio et al. 2017). A paper focused on curbing the spread of fake news by Vo and Lee (2018) proposes an approach to identify and encourage a certain kind of user-base called ‘guardians’ whose purpose is to observe and conduct regular fact-checking, and a new fact-checking URL recommendation system, and together was able to outperform state-of-the-art models. However, the limitation was the active use involvement by human employees, which is a more costly and slower approach.

Similar work has been done in the very related areas of fake reviews and clickbait detection. Nandimath Jyoti (2017) mentioned that buyers go through an evaluation of reviews before buying an item and reduces the counterfeiting feedback based on opinion mining through text classification. Feng et al. (2012) categorized different types of clickbait are categorized using classical machine learning algorithms and showed that some of them are highly correlated with non-factual claims. Crawford et al. (2015) worked on online spam detection from reviews using NLP and text classification algorithms.

Deep learning for misinformation detection on online social networks: a survey and new perspectives (Lichterman 2016) is a comprehensive article underlining many aspects of deep learning usage in fake news detection. Under the category of discriminative models, it brings out the importance of contextual and sequence information, showing how recurrent neural network or long short-term memory-based models perform better than convolutional neural networks. Lichterman (2016) also mentions some relevant works in generative models such as generative adversarial network or variational autoencoders which show promise, but are yet to be fully explored for our task. We realized an opportunity to employ transformer models, as an uncharted approach for fake news detection.

Another way to solve this problem has been experimented by Shu et al. (2018) where they use graphs to track the path through which fake news is propagated in the network and the users who share it, and make a judgement on the credibility of the news. However, it has the limitation of being blind to the actual content of the news. Extracting features from user activities as well as network information has been shown to be effective at spammer detection, and those accounts can then further be targeted to control the spread of fake news (Hu et al. 2014a). This is especially important because a large amount of the propagation of fake news is done by bots without verification (Hu et al. 2014b). A major limitation in statistical approaches to fake news detection has been the size of the labelled benchmark datasets, which was alleviated when Wang (2017) presented a public manually labelled dataset with 12.8k data points, an order of magnitude higher than what we had before and investigated into detecting them using CNN-based models. However, the absence of justification or elaboration of news content leaves scope for improvement in the results. Alhindi (2018a) further improved on it by extending the dataset using justification from the news articles and tried out techniques like SVM, and Bi-LSTM for classifying the news as true or false. A limitation of that approach was that the claim and justification refer to temporal information are harder to model by the relatively shallow approaches they used in the paper.

Deep learning is also a promising approach towards the detection and classification of fake news. Kaliyar et al. (2020) proved the superiority of using deep neural networks as opposed to traditional machine learning algorithms in the detection. The use of deep diffusive neural networks for the same task has been demonstrated in Zhang et al. (2019). The use of support vector machines has been shown to be superior as compared to Naive Bayes network (Aphiwongsophon and Chongstitvatana 2018). Various state-of-the-art approaches have been compared (Kumar et al. 2019) and proposed a novel deep learning approach based on CNN + bidirectional LSTM ensemble network with attention mechanism that was found to outperform the state-of-the-art results.

In comparison for binary label classification, we reference the following works:

- Alhindi et. al experiment with both logistic regression (LR) and support vector machines (SVM) with linear kernel. For the basic representation of the claim/statement (S condition), they experimented with unigram features, tf-idf weighted unigram features and Glove word embeddings. They chose to use bidirectional long short-term memory (BiLSTM) architectures that have been shown to be successful for various related NLP tasks such as textual entailment and argument mining. For the S condition they use BiLSTM (size 32) to model the statement preceded Glove pre-trained word embeddings, a 100-dimensional embedding layer. The output of the BiLSTM layer is passed to a softmax layer.
- Yang et. al propose an unsupervised learning framework, UFD, which utilizes a probabilistic graphical model to model the truths of news and the users' credibility. An efficient collapsed Gibbs sampling approach is proposed to solve the inference problem.
- Wang (2017) used five baselines: a majority baseline, a regularized logistic regression classifier (LR), a support vector machine classifier (SVM), a bidirectional long short-term memory networks model (Bi-LSTMs) and a convolutional neural network. For LR and SVM, they used the LIBSHORTTEXT toolkit, which was shown to provide very strong performances on short text classification problems. For Bi-LSTMs and CNNs, they used TensorFlow for the implementation. They used pre-trained 300-dimensional word2vec embeddings from Glove to warm-start the text embeddings. They strictly tuned all the hyperparameters on the validation dataset. The best filter sizes for the CNN model were (2, 3, 4). In all cases, each size had 128 filters. The dropout keep probabilities was optimized to 0.8.
- Balwant et. al use a POS tagging and word embeddings assigned to each word and later each statement is aggregated to a different 300-dimensional embedding. This is sent through a pre-trained word vector layer (Glove) followed by a Bi-LSTM layer. The output vector is then concatenated with a CNN output fed with user profile information. The concatenated layer goes through a softmax classifier to make final predictions.

3 Dataset analysis

In evaluating the performance of our proposed architecture for fake news detection, we have used two datasets, LIAR dataset (Vo and Lee 2018) and LIAR PLUS dataset (Vaswani et al. 2017). LIAR dataset is a collection of political news obtained from fact-checking website 'POLITIFACT' along with various metadata and a mention of its degree of its veracity, split into six labels, these being 'true', 'mostly true', 'half true', 'mostly false', 'false' and 'pants-fire'.

In regard to the context of each label, 'true' indicates a completely true statement. 'Mostly true' suggests that overall the statement is correct with some minor incorrect details. 'Half true' indicates that certain facts may be correct, but there are a lot of incorrect facts as well. 'Barely true' means that the statement in question is for the most part incorrect with minor details that might be true. 'False' label is an incorrect statement. 'Pants on Fire' refers to a completely incorrect fact with no basis in truth. Table 1 displays a sample statement along with each of the label, showing a glimpse of what the dataset looks like.

The 14 columns of the dataset are ID, label, statement, subject, the speaker, their job title, state info, speaker's political affiliation, 'true' counts, 'half true' counts, 'mostly true' counts, 'mostly false' counts, 'false' counts, 'pants-fire' counts and the venue/location. The LIAR dataset has around 12,791 records split between 3 tsv files, namely 'train.tsv', 'test.tsv' and 'valid.tsv' which are used for training, testing and validation, respectively. 'train.tsv' has 10,240 records, 'test.tsv' has 1267 and 'valid.tsv' has 1284. There is not much imbalance between the different labels. Also used is an extension of the LIAR dataset, which is called the LIAR PLUS dataset. This includes all the columns of the original LIAR but extends it further by adding a justification column which provides human justification as to why the statement has been classified to its label.

This human justification is obtained from the 'Our Ruling' section present in every 'POLITIFACT' article. However, every sentence which contains the verdict or any verdict related words are removed. These include words like 'true', 'false', 'inaccurate', 'correct' among 25 others. This human explanation is shown to have significantly improved performance over LIAR. Presented above are the word clouds and label counts for the two largest political organizations form the dataset, namely the 'Democratic Party' and the 'Republican Party'.

Figure 1a and b shows the word cloud of the statements from the dataset based on whether it was made by a Democrat or a Republican, respectively. These are two major political parties in the USA, and hence, they are used to bring up some common themes that they talk about. There is not much imbalance between the different labels in the dataset, instances of each label range from 2,638 to 2,063, apart from 'Pants on Fire' label which has 1050 instances. This can be observed from Fig. 1c and d each of which show a distribution of the statement labels across the two parties.

4 Methodology

4.1 Data pre-processing

In order to effectively use text data for natural language processing, we start with data pre-processing. Tokenization is the process of splitting out text strings into a list of tokens that can later be used for lexical analysis, and our data fields are converted to tokens. Stop words are the common words in any language which do not provide much context and hence are best left out. So they make take up unnecessary computation power and time. We remove stop words such as 'but', 'and' and 'or' as well as punctuation marks like ',', ';' among others.

4.2 Transformer

A transformer architecture converts a sequence to another sequence, but the key difference is that it works without any recurrent networks, but using only attention-focusing mechanisms.

Both the encoder and decoder are both made up of a six-layer stack, and each layer is followed by two sublayers under it. In order, we have a self-attention mechanism, a fully connected feed forward network and then a normalization layer. The decoder additionally implements a third sublayer which then implements multi-head attention onto the result of the previous stack. The attention mechanism makes use of local information effectively to decide the next sequence by focusing on the important part of the information.

In the absence of recurrent memory networks like RNN or LSTM, the positional encoding of the words becomes significant. The inputs and target sentences are embedded into a multidimensional space. The self-attention function weighs the relative relevance of the set of input embeddings to the output embeddings. The final feed forward network present in both the encoder and the decoder performs some additional processing on the values. The final linear and softmax layers in the ending decoder are used to generate the output probabilities (Vaswani et al. 2017).

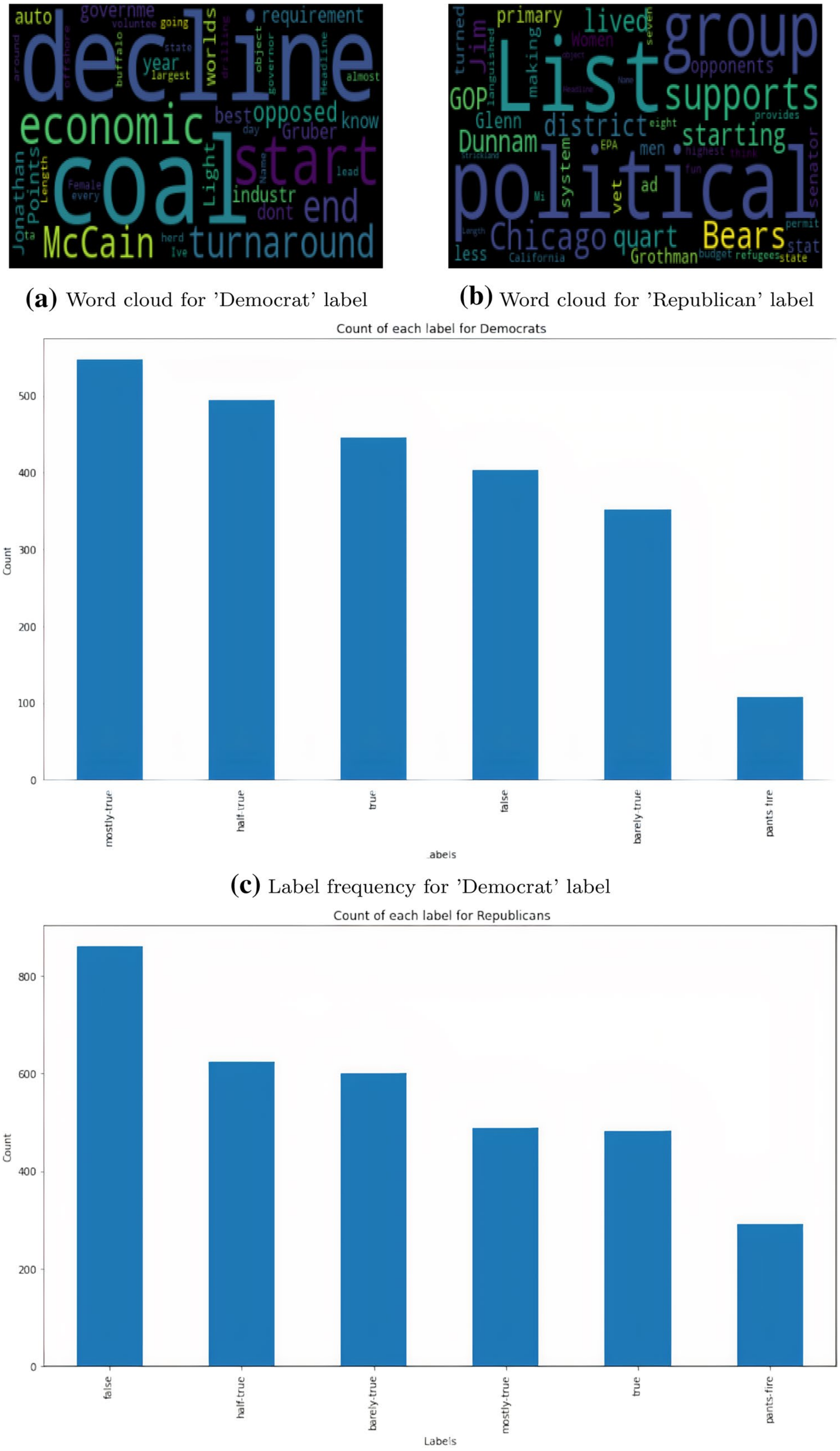
Thus, we see that the encoder and decoder blocks in transformers are just multiple identical decoders and encoders stacked on top of each other. The attention mechanism implemented in transformer blocks helps to focus on the relevant parts of the sequence, rather than all at once. Transformers have been shown to be out performing state-of-the-art models for language translation tasks and are both generalizable to other tasks, and parallelizable (Vaswani et al. 2017).

4.3 Transfer learning and BERT

Most neural networks, whether shallow or deep, share a similar phenomenon—once the learning is complete, there's a little scope for generalization to different tasks. Transfer learning is an optimization over a single or isolated learning process, which allows generalizing the knowledge learnt from one task to another task. The result is a faster and more optimized learning process. Firstly, a base network is trained, and secondly, the target network based on the knowledge transferred from the base. Transfer learning has been proven effective to a variety of machine learning tasks, including text sentiment classification (Wang and Mahadevan 2011) and multi-language text classification (Zhou et al. 2014). BERT uses transfer learning from pre-training on corpus.

Bidirectional Encoder Representations from Transformers is an advanced language model for natural language processing which was developed by the researchers of Google in

Fig. 1 Dataset analysis for the largest labels, 'Democrat' and 'Republican'



(a) Word cloud for 'Democrat' label

(b) Word cloud for 'Republican' label

(c) Label frequency for 'Democrat' label

(d) Label frequency for 'Republican' label

2018 (Alhindi et al. 2018b). It is based on transformers and uses the power of transfer learning. BERT uses the concept of transfer learning as it is already provided pre-trained on a huge text corpora—and it can be either used as the pre-trained version or fine-tuned to our dataset. BERT model is pre-trained on BookCorpus which is made up of over 800 million words and a huge chunk of English language Wikipedia, which is around 2,500 million words (Balwant 2019). The architecture of BERT is shown in Fig. 3.

BERT is essentially a trained, multi-directional transformer encoder stack. BERT comes in 2 sizes—'BERT large', which contains 24 encoder layers, and 'BERT base', which contains 12 encoder layers. There exist 22 BERT models as of now, each adhering to different situations and languages like English, Japanese, Chinese and German among others. In the implementation of our triple branched BERT network, we use 'bert-base-uncased' which contains 12 encoder layers, 768 hidden, 12 heads and 110M parameters. It is trained exclusively on lower case English text.

The standard way for using language modelling for NLP tasks was to use pre-trained word embeddings like word-2vec, where the words are converted to representative vectors and trained on neural networks as mathematical elements. However, with the concept of transfer learning,

Table 1 Sample text for each of the six labels

ID	Sample text	Label
1	Over the past 5 years the federal government has paid out 601 million in retirement and disability benefits to deceased former federal employees	True
2	The federal minimum wage is worth about 20% less than it was when Ronald Reagan gave his first address to a joint session of Congress	Mostly true
3	Tom Ganley has two Fs from the Better Business Bureau and over 160 complaints in just 3 years	Half true
5	Any state tax law has to start in the House and the renewal of the state hospital bed tax this year started in the Senate, which is unconstitutional	Barely true
4	U.S. Rep. Lloyd Doggett is the most liberal man in the United States Congress	False
6	President Obama has the lowest public approval ratings of any president in modern times	Pants on fire

Fig. 2 BERT input representation

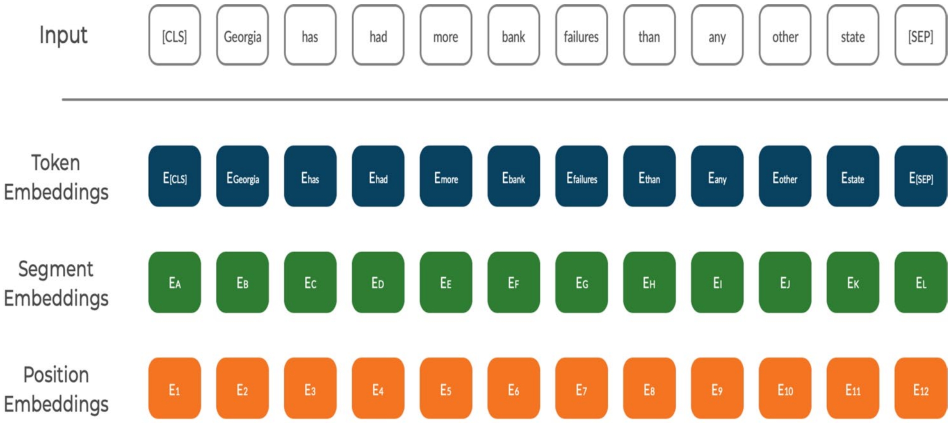
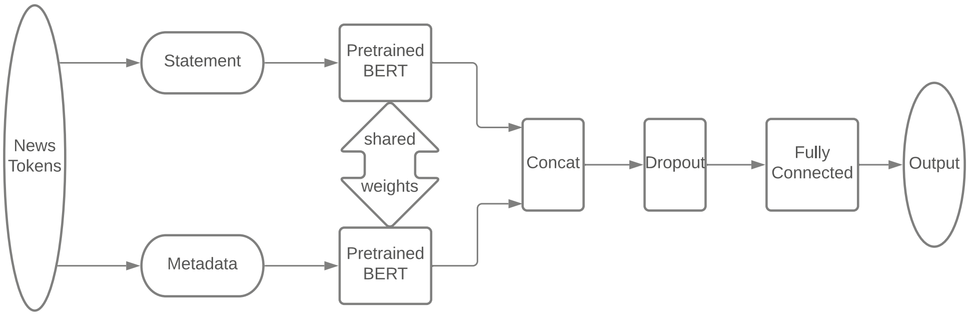


Fig. 3 Proposed architecture for double BERT network using LIAR



BERT offers a feature-based approach, where it is possible to transfer the knowledge gained from the pre-training to a supervised downstream task. This is called the fine-tuning approach.

A special [CLS] token, where CLS stands for classification, is added to the first input token. Sentence pairs are grouped together into a single representation. The way to differentiate sentences from each other is using a special separate token [(SEP)] as shown in Fig. 2. A learned embedding is appended to every token to mark if it belongs to sentence A or B, as shown in Fig. 3. Similar to the encoder of the transformer, BERT takes as input a sequence of words, and they flow up the network, through the self-attention and feed-forward networks before it is passed to the next encoder. And each position gives an output vector having size 768 in BERT Base (Devlin et al. 2018).

In our proposed architecture, we use pre-trained BERT layers as integral parts of the Neural Network which are fed the data of the statements, the metadata, and if available, the human justification.

4.4 Proposed architecture

As we discussed before, BERT input is supposed to take a single sentence as input, and according to the makers, BERT was not pre-trained to handle multiple SEP tokens between sentences nor does it have a third token type for this purpose.

BERT large and BERT base both come in case sensitive and insensitive variations. In the case of fake news detection, it does not make sense that the capitalization of letters has anything to do with the fakeness of news, so the added complexity of having a much larger vocabulary has been avoided and we used the uncased variant.

Now, the LIAR dataset contains important information of two kinds—the news content, the metadata and the LIAR PLUS dataset contains important information of three kinds—the news content, the human-extracted justification, and the metadata. Because of the structure of BERT, we cannot take two or three different inputs. Since the human justification is not part of the news content, to preserve the context of each sentence we cannot simply add it as part of the news. However, it has been repeatedly proved that not taking advantage of more information in the form of either justification or metadata significantly reduces the classification accuracy of machine learning models (Wang 2017; Long et al. 2017; Alhindi et al. 2018b). Therefore, we decided to use multiple BERT models with shared weights between them in order to handle multiple inputs and take advantage of the additional justification and metadata.

We use a multiple branch BERT as base network with shared weights. We selected BERT in our base network due to it being state of the art in terms of language modelling and translation. We leverage pre-trained BERT weights and fine-tune them for classifying news as fake or real. When we consider LIAR dataset, since it does have justification, we construct a double BERT architecture which accepts statement tokens in first branch and the metadata in the second, as shown in Fig. 3. Here the information available in the metadata includes the subject, speaker's job title, the state information, the speaker's political affiliations, previous history of the counts for each of the labels of veracity and the venue/location.

In the case of LIAR plus dataset, the first BERT branch accepts the news statement tokens. The second BERT branch accepts the human justification tokens, and the third BERT branch accepts the metadata tokens. This is shown clearly in Fig. 4.

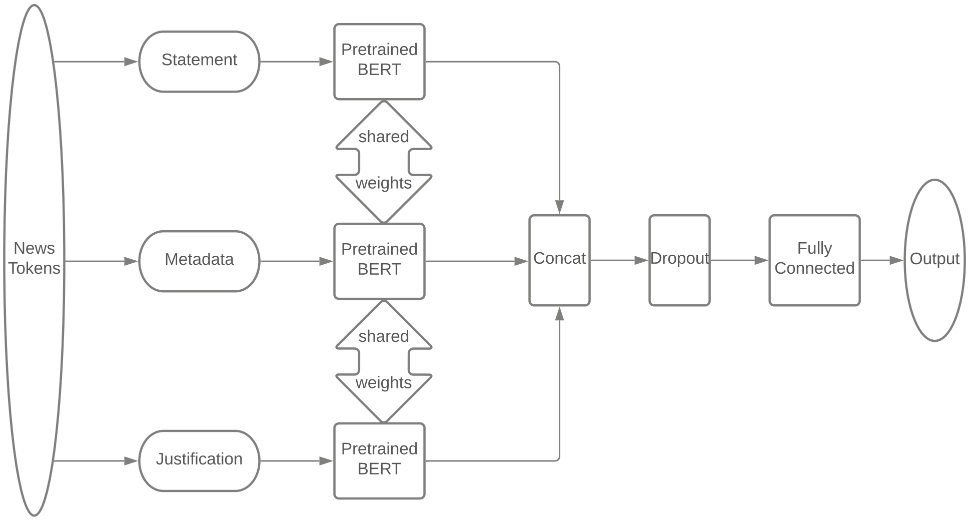
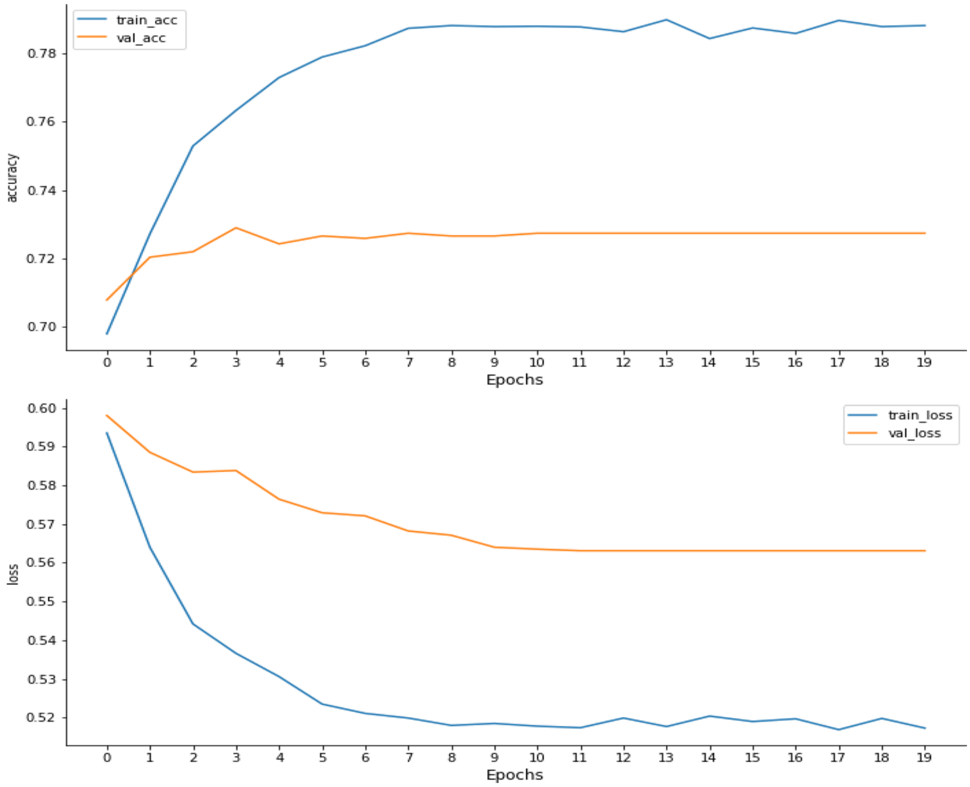


Fig. 4 Proposed architecture for triple BERT network using LIAR PLUS

Fig. 5 Training accuracy, validation accuracy and loss plots when trained using LIAR with binary labels



The sequence size of input for each BERT branch was chosen depending on the median number of words in the branches. From the dataset analysis, we concluded that the appropriate input sizes of each of the branches would be 64 for the news content branch, 32 for the metadata branch because there's no inputs longer than that and 256 for the justification branch, because justifications are usually longer in length. This helps to optimize memory use. The output of each BERT layer branch is a one-dimensional tensor of shape (768). Those three output tensors are concatenated. To the output of the concat layer, we add a weighted reliability factor which increases the output activation between the false and verified news. This factor judges the authenticity of a news source based on columns 9 to 13 of the dataset which contain the previous counts for each of the labels of the given news source, such as 'true', 'mostly true', 'half true', 'mostly false', 'false' and 'pants on fire'. This is then passed through a linear fully connected layer following a dropout layer which gives binary output logits. After that the softmax activation gives the final output probabilities. This whole network is retrained using aforementioned fake news datasets. Even BERT parameters are set as trainable. However, they are not changed drastically as their pre-training is performed on mammoth datasets. The newer layers are tuned well since they are initialized in fine-tuning stage, and also since they are close to the downstream end of the network.

5 Results and analysis

To test our Triple branch BERT network, we use two datasets, namely LIAR dataset and LIAR PLUS dataset. Also along with this, we also check state-of-the-art solutions to the problem of fake news detection by comparing our accuracy against support vector machine (SVM), logistic regression (LR) and a bidirectional long short-term memory neural network (Bi-LSTM).

We test our proposed models with the LIAR and LIAR PLUS datasets under two situations, with binary labels and with six-way labels. In binary labels condition, the 'true', 'half true' and 'mostly true' labels are clubbed into one common one, 'true'. Similarly, 'false' is an amalgamation of the labels 'mostly false', 'false' and 'pants-fire'. However, in six-way condition, we test keeping each label as they are. In each condition, the model is trained for 20 Epochs each.

Figures 5 and 6 represent the training accuracy, validation accuracy and loss plots when trained using LIAR and LIAR plus with binary labels. In regard to original LIAR dataset, peak validation accuracy was achieved after the third epoch for binary classification and after the fifth epoch for six-way classification. For LIAR-PLUS dataset, peak accuracy for binary labels was achieved after fourth epoch and for six-way, after the fifth epoch. The details of our final hyper-parameter tuned models are :

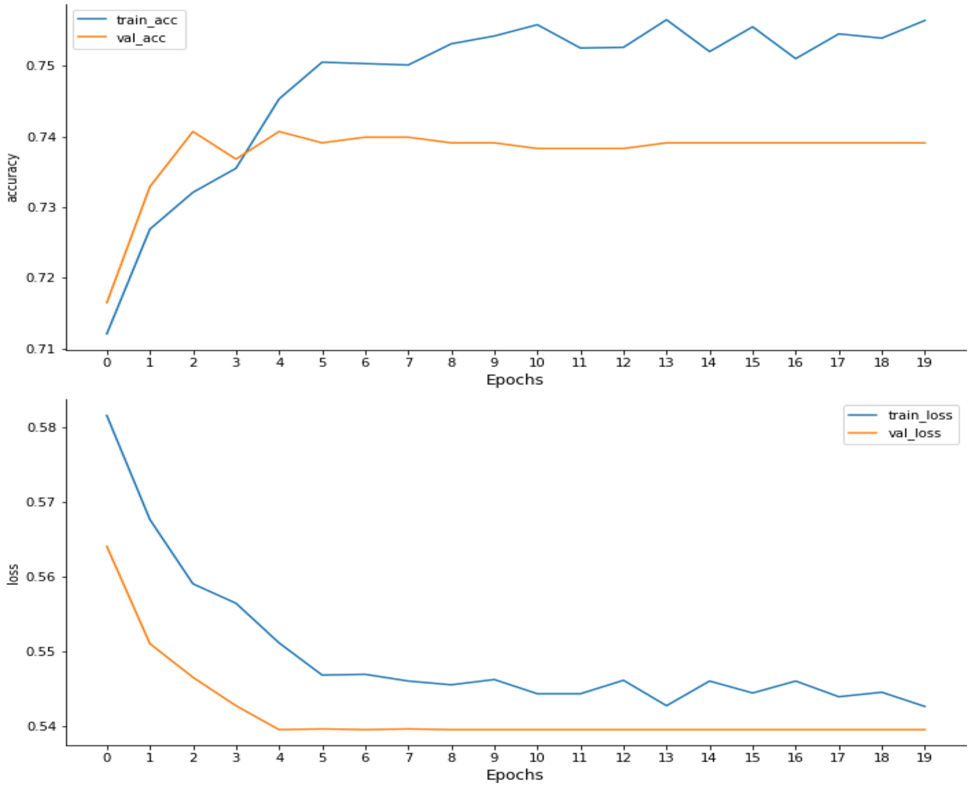


Fig. 6 Training accuracy, validation accuracy and loss plots when trained using LIAR PLUS with binary labels

- Optimizer: Adam
  - Learning rate for BERT layers: 10^{-5}
  - Learning rate for succeeding layers: 10^{-4}
- LR scheduler: Step LR with step size = 3 and gamma = 0.1
- Activation function: GeLU for BERT hidden layers, Softmax for output layer
- Size of model: 12 bert layers (self-attention, feed-forward, normalize) +1 dropout layer and 1 Fully Connected layer
- Dropout rate: 0.1 default bert
- No. of parameters: 111,179,648
- BERT: Hidden layer size—768 ; Maximum position embeddings—512 ; Intermediate layer size—3072

Tables 2 and 3 show the comparison of fake news classification accuracy with binary labels and six labels. When comparing results with existing solutions, we can see that the BERT networks are able to outperform SVM, LR and LSTM on their own. When it comes to binary classification,

Table 2 Comparison of fake news classification accuracy with binary labels

Model	Var	
	Val	Test
Alhindi et al. (2018b)		
LR	0.69	0.67
SVM	0.66	0.66
Bi-LSTM	0.71	0.70
P-BiLSTM	0.70	0.70
Yang et al. (2019)		
UFD-Bayesian network model	0.75	N/A
Double BERT network	0.72	0.72
Triple BERT network	**0.75**	**0.74**

Highest values are in bold

Table 3 Comparison of fake news classification accuracy with six labels

Model	Var	
	Val	Test
Wang (2017)		
LR	0.257	0.247
SVM	0.258	0.247
Bi-LSTM	0.223	0.233
CNN	0.260	0.270
Hybrid CNN	0.247	0.274
Balwant (2019)		
Bi-LSTM based on POS tags	0.204	0.274
Double BERT network	0.358	0.352
Triple BERT network	**0.360**	**0.371**

Highest values are in bold

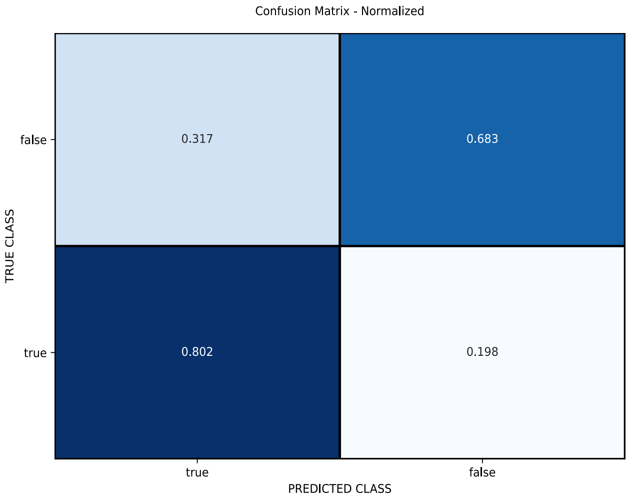


Fig. 7 Confusion matrix for binary label classification

it is able to compete with state-of-the-art implementations proposed by Yang et al. (2019), by providing similar metrics. In regard to six-label classification, we are able to improve metrics proposed by Wang (2017) and Balwant (2019) by significant advancements of about 9%.

It is also to be noted that accuracy for LIAR PLUS dataset is higher than LIAR dataset due to the human justification extension that LIAR PLUS presents. The increase in accuracy is due to the usage of the statement, justification and all the metadata from the datasets with the triple BERT network. The addition of the weighted reliability scores to the output of the concat layer of the network is also expected to be a vital factor in improvement in accuracy. The model is able to learn much more under binary labels with the help of these modifications. Table 4 shows the performance of different metrics for the proposed fine-tuned BERT models. When comparing the double BERT and Triple BERT, 5.3% of accuracy improvement in six-class classification and 2.7% for binary model. It reflects the role of the justification feature in the Fake news dataset.

Figures 7 and 8 illustrate the confusion matrix of binary and six-label classification results of the triple BERT model. Figure 7 represents the confusion matrix for binary classification. It indicates that most of the true cases are identified correctly and only a minimal percentage are false negatives. However, we see a lesser contrast in case of false cases, where even through the distinction is clear, it is not as stark as the true cases.

Figure 8 clearly shows the less accuracy on true, barely true and pants on fire classes. The true cases are mostly misclassified as half true and mostly true. The barely true classes are misclassified as half true and false classes. The pants on fire class is mostly misclassified as false case. This goes to show that although the model is not as precise as it

Fig. 8 Confusion matrix for six-label classification

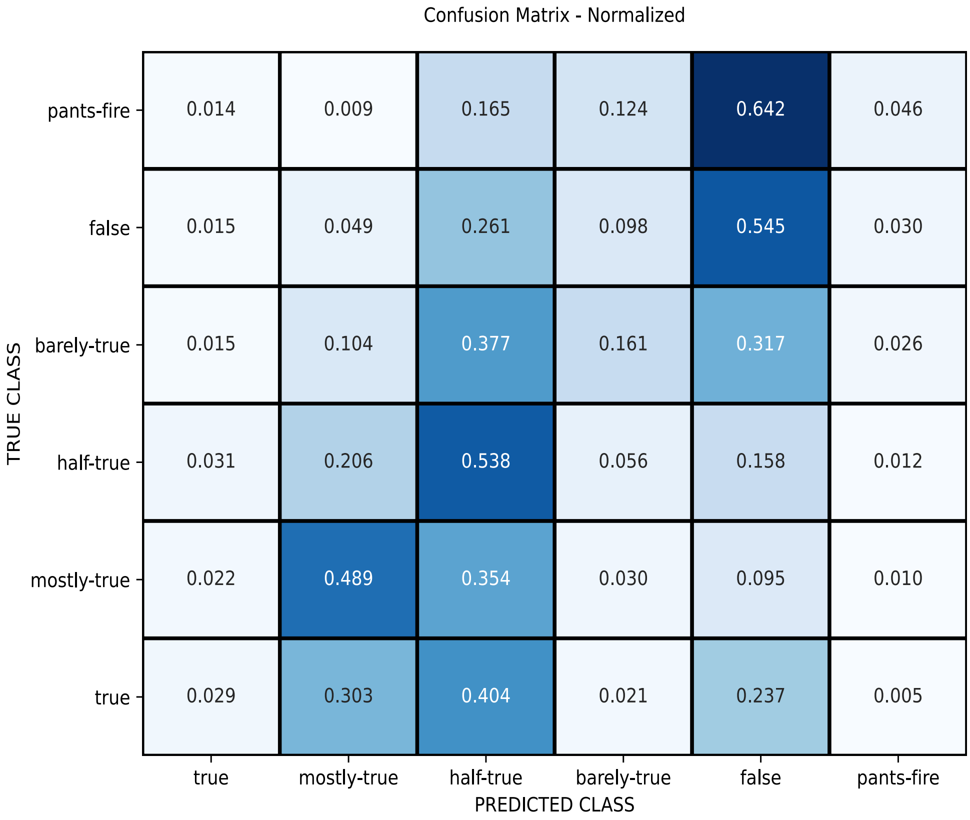


Table 4 Metrics for the proposed fine-tuned BERT models

Metric	Double BERT	Triple BERT
For binary classification		
Precision	**0.712**	0.699
Recall	0.802	**0.854**
F1	0.754	**0.768**
F2	0.782	**0.817**
Accuracy	0.72	**0.74**
For six-label classification		
Accuracy	0.352	**0.371**
Hamming loss	0.649	0.650

Highest values are in bold

should be, it does a commendable job at classifying statements to an adjacently similar class. We can also observe in this matrix that there's fairly visible boundary between the four true-category classes and the two false-category classes. This leads us to believe that the model holds up and can perform reliably in multi-label classifications as well.

6 Conclusion and future work

In today's hyper-digital world where the rapid spread of false information can have immense real-world ramifications, solutions to classify news and other statements as true or untrue will be crucial. This work outlines the importance of such solutions, details existing works, where there is area for improvement and provides a novel solution by introducing the multiple BERT layer neural network for fake news detection. The networks are tested against two state-of-the-art datasets, namely LIAR and LIAR PLUS. Accuracy of its classification is then compared against existing implementations of fake news classifiers to observe significant improvement in some areas.

In the future, this architecture can be tested on different application domains and it could improve existing benchmarks as well. Our proposed model could be explored and tested with different set-ups over the coming days and we hope to achieve better performance than the current state. We plan to continue working on this model and progress by:

1. Exploring different combinations of features in concatenated BERT models. We also aim to implement this over different datasets that hopefully have more features available.
2. Trying to achieve results by multiple BERT models (4-BERT possibly).
3. Tuning the hyper parameters of both BERT and subsequent layers and making a detailed study on its effects.

References

Ahmed H, Traore I, Saad S (2017) Detecting opinion spams and fake news using text classification. Secur Priv. https://doi.org/10.1002/spy2.9

Alhindi T (2018a) Where is your evidence: improving fact-checking by justification modeling. In: Proceedings of the first workshop on fact extraction and verification (FEVER), Brussels, Belgium, pp 85–90

Alhindi T, Petridis S, Muresan S (2018b) Where is your evidence: improving fact-checking by justification modeling. In: Proceedings of the first workshop on fact extraction

Aphiwongsophon S, Chongstitvatana P (2018). Detecting fake news with machine learning method, pp 528–531. https://doi.org/10.1109/ECTICon.2018.8620051

Balwant MK (2019) Bidirectional LSTM Based on POS tags and CNN architecture for fake news detection. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT). https://doi.org/10.1109/ICCCNT45670.2019.8944460

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146

Bourgonje P, Schneider JM, Rehm G (2017) From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In: Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism, pp 84–89

Chen ZF, Cheng Y (2020) Consumer response to fake news about brands on social media: the effects of self-efficacy, media trust, and persuasion knowledge on brand trust. J Prod Brand Manag 29(2):188–198

Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H (2015) Survey of review spam detection using machine learning techniques. J Big Data 2(1):1–24

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Eugenio T et al (2017) Some like it hoax: automated fake news detection in social networks. arXiv preprint arXiv:1704.07506

Feng S, Banarjee R, Choi Y (2012) Syntactic stylometry for deception detection. In: ACL'12

Handley L (2018) Nearly 70 percent of people are worried about fake news as a 'weapon,' survey says. Retrieved from https://www.cnbc.com/2018/01/22/nearly-70-percent-of-peopleare-worried-about-fake-news-as-a-weapon-survey-says.html

Hu X, Tang J, Gao H, Liu H (2014a) Social spammer detection with sentiment information. In: ICDM'14

Hu X, Tang J, Liu H (2014b) Online social spammer detection. In: AAAI'14, pp 59–65

Kaliyar RK, Goswami A, Narang P, Sinha S (2020) FNDNet–a deep convolutional neural network for fake news detection. Cogn Syst Res 61:32–44

Kumar S, Asthana R, Upadhyay S, Upreti N, Akbar M (2020) Fake news detection using deep learning models: a novel approach. Trans Emerg Telecommun Technol 31(2):e3767

Lichterman J (2016) Nearly half of US adults get news on Facebook, Pew Says. URL: http://www.niemanlab.org/2016/05/pew-report-44-percent-of-us-adults-get-news-onfacebook

Long Y et al (2017) Fake news detection through multi-perspective speaker profiles. In: Proceedings of the eighth international joint conference on natural language processing, vol 2: short papers, pp 252–256

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems, Lake Tahoe, NV, USA, 5–10, pp 3111–3119

Nandimath JN, Katkar BS, Ghadge VU, Garad AN (2017) Efficiently detecting and analyzing spam reviews using live data feed. Int Res J EngTechnol (IRJET) 4(2):1421–1424

Rapoza K (2017) Can 'fake news' impact the stock market? RealClearMarkets, Forbes

Resnick B (2018) False news stories travel faster and farther on Twitter than the truth, Vox.(Erişim: 09.09. 2019). https://www.vox.com/science-and-health/2018/3/8/17085928/fake-news-study-mit-science

Silverman C (2016) This analysis shows how viral fake election news stories outperformed real news On Facebook. BuzzFeed News, BuzzFeed News. www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook. Accessed 16 Nov 2016

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. SIGKDD Explor Newslett 19(1):22–36

Shu K, Bernard H, Liu H (2018) Studying fake news via network analysis: detection and mitigation

Tang J, Yi C, Huan L (2014) Mining social media with social theories: a survey. ACM SIGKDD Explor Newslett 15(2):20–29

Vaswani A et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 6000–6010

Vis F (2014) 10. The rapid spread of misinformation online. World Economic Forum. Retrieved from http://reports.weforum.org/outlook-14/top-ten-trends-categorypage/10-the-rapid-spread-of-misinformation-online/

Vo N, Lee K (2018) The rise of guardians: fact-checking URL recommendation to combat fake news. In: The 41st international ACM SIGIR conference on research & development in information retrieval, pp 275–284

Volkova S, Shaffer K, Jang JY, Hodas N (2017) Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on Twitter. In: ACL

Wakefield J (2016) Young using social media to access news. BBC News. www.bbc.com/news/uk-36528256. Accessed 15 Jun 2016

Wang WY (2017) "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, vol 2, Short Papers, pp 422–426

Wang C, Mahadevan S (2011) Heterogeneous domain adaptation using manifold alignment. In: Proceedings of the 22nd international joint conference on artificial intelligence, vol 2, pp 541–546

Yang, S et al (2019) Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI conference on artificial intelligence, vol 33

Zhang S, Wang Y, Tan C (2018) Research on text classification for identifying fake news. In: IEEE 2018 international conference on security, pattern analysis, and cybernetics (SPAC). https://doi.org/10.1109/SPAC46244.2018.8965536

Zhang J et al (2019) FAKEDETECTOR: effective fake news detection with deep diffusive neural network. In: 2019 36th IEEE international conference. https://doi.org/10.1109/ICDE48307.2020.00180

Zhou JT, Tsang IW, Pan SJ, Tan M (2014) Heterogeneous domain adaptation for multiple classes. In: International conference on artificial intelligence and statistics, pp 103–1095

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