An Effective Hybrid Text-Based Approach to Identify Fake News on Social Media

Imtiez Fliss[a] and Hamza Bargougui
National School of Computer Sciences, Manouba University, Tunisia

Keywords: Text-Based Fake News Detection, BERT (Bidirectional Encoder Representations from Transformers), Hybrid Deep Learning Techniques.

Abstract: Because of their low cost, simplicity of access, and quick dissemination, social media are today one of the primary information sources for millions of people worldwide. However, this is at the expense of dubious credibility and a large danger of being exposed to ”fake news,” which is deliberately designed to mislead readers. In light of this, in this paper we propose a novel method for identifying bogus news based on the text content. This method is founded on a mix of BERT (Bidirectional Encoder Representations from Transformers) and deep learning techniques (Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM)). Promising results are seen when the proposed approach is compared to other models.

1 INTRODUCTION

Today’s world is one that is constantly evolving. For millions of individuals, social media is the best source of information and is quickly taking over as the preferred news source. They have the power to make decisions on matters of politics, society, and the economy, but they may also harm the reputations of people and organizations while advancing certain persons, teams, or organizations.

There is a ton of information available, some of it accurate, some not. Most of the time, this data is instantly shared without being thoroughly checked. In recent years, the phenomenon of fake news (Quandt et al., 2019; Kalsnes, 2018) has become widespread, facilitated by social networks. There are several uses for these false reports. Others have an impact on political decisions and public perception of financial markets through altering the reputation of businesses and institutions on the Internet. Some are developed merely to boost clicks and traffic to a website.

Therefore, it is crucial to make sure that information is accurate, keep an eye on internet news, and create tools to counteract fake news. For this purpose, We can use social media deep learning (Bengio, 2019; LeCun, 2016) and can facilitate correct analysis of information types. On the other hand, using semantic analysis of circulating data is very interesting. This allows us to recognize relevant information in context and classify it accordingly.

In this context, to simplify accurate analysis of the type of information published and to distinguish between false news and true news, the purpose of this paper is to present a novel method based on deep learning categorization of texts shared on social media.

This approach combines the use of BERT method as a feature encoder with convolutional neural networks (CNN) (Kim, 2014) and Long Short Term Memory (LSTM) (Allcott and Gentzkow, 2017) for text classification. Finally, this approach put into practice, and its accuracy is evaluated against other models taken from the literature.

The remainder of this paper is structured as follows. While, Section 2 introduces related work, Section 3 presents the proposed approach for detecting fake news. Section 4 contains experiments and discussion. Finally, some concluding remarks and perspectives are presented.

2 RELATED WORK

Before presenting existing fake news detection models, it is important to define what fake news is. So, we will first identify the exact characteristics that define fake news. Next, we will focus on the main fake news detection models.

[a] https://orcid.org/0000-0003-2229-7004

Fliss, I. and Bargougui, H.
An Effective Hybrid Text-Based Approach to Identify Fake News on Social Media.
DOI: 10.5220/0011674400003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 374-380
ISBN: 978-989-758-623-1; ISSN: 2184-433X

2.1 Fake News Definition

The term ”Fake News” began to be used frequently after the 2016 US election. According to (Allcott and Gentzkow, 2017; Shu et al., 2017; Zhou et al., 2019; Zhou and Zafarani, 2020), fake news are news articles that are intentionally and verifiably false and can misinform readers. There are two key characteristics of this definition: veracity and intent. First, fake news involves false information that can be verified intrinsically. Second, fake news is created with the dishonest purpose of misleading readers. This definition has been widely adopted by (Klein and Wueller, 2017; Lazer et al., 2018; Mustafaraj and Metaxas, 2017; Potthast et al., 2017). According to (Ahmed et al., 2017), the term Fake News is based on four main components: creator, target victim, news content and social context.

- Creator: creators of fake news online can be a human being or not.
- Target victims: Depending on the objectives of the news, targets can be students, voters, parents, elderly people.
- News content: News content refers to the body of the news.
- Social context: the social context indicates how the news are disseminated on the Internet. Social context analysis includes the user, network analysis (how online users are involved in the news) and analysis of the dissemination pattern.

2.2 Fake News Detection Models

There are numerous models for spotting bogus news (Kaur et al., 2019). We focus in this section on recent works looking at the detection of fake news based on news content. In (Vivek et al., 2017), to acquire linguistic attributes for each of the articles, authors employed the LIWC (Linguistic Analysis and Word Count) package. Z-score normalization was used to normalize each feature. Then, focusing on how well the algorithms performed for the test set, they developed a number of machine learning models based on well-known methods like logistic regression, support vector machines, random forests, decision trees, and k-neighbors classifier. The Support Vector Machine algorithm outperformed the others in terms of prediction accuracy.

Ahmed et al. (Ahmed et al., 2018) also examined the performance of six different algorithms that categorize information as false or true using N-gram traits and TF-IDF scores. The dataset included both real and false comments on political news stories and product evaluation comments. A model that they created can be used in two different contexts. Their study demonstrates that utilizing TF-IDF and SVM to rank news articles was the most effective method.

Convolutional neural network (CNN) models have also been deployed in detecting fake news. Authors in (Xu et al., 2020) propose a fake news detection solution to the problems occurring during an analysis of short texts namely, CNN_Text_Word2vec.CNN_Text_Word2vec introduces a word2vec neural network model to train distributed word embeddings on every single word.

Authors in (Nicole et al., 2018) use the deep neural network introduced by (Kim, 2014) to learn to detect fake news. It has been proven that this works well for text classification. All of the text’s words are a one-hot representation, which is a sparse vector with only one item equal to 1 to represent each word. This network’s first layer is a pre-trained word2vec embedding, which converts each word into a 1,000-dimensional representation with close distances between words with comparable meanings. the introduction follows a method for identifying the patterns in an article that are most beneficial for classifying it as fake or legitimate news. This entails returning the network’s output to the article in order to identify the phrases that are ”most fake” and ”most real.”

On the other hand, BERT-based models had been successfully applied to the fake news detection task; FakeBERT is a new technique Kaliyar et al. (Kaliyar et al., 2021) created to identify bogus news in the early 2020s. By merging several parallel blocks of the single-layer CNNs with BERT, this method is effective contrarily to (De Sarkar et al., 2018) that looked at a text sequence in a unidirectional way.

In (Verma et al., 2022), the Message Credibility (MCred) framework was proposed. This research makes use of both local and global text semantics to detect bogus news. The Bidirectional Encoder Representations from Transformers (BERT) and Convolutional Neural Networks (CNN) used in this framework are combined to provide global text semantics based on the relationship between words in sentences and local text semantics based on N-gram characteristics.

3 PROPOSED APPROACH FOR THE DETECTION OF FAKE NEWS

Our method focuses on taking contextual understanding into account when determining whether text posts are fake news or true news. The text classification process displayed in Figure 1 serves as a major inspiration for the key elements of our approach.

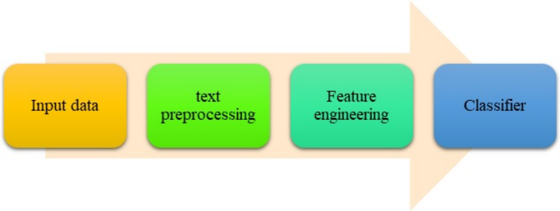

Figure 1: Overview of a text classification pipeline.

Our contribution consists in the choice of adequate models for each step.

In fact, after selecting the dataset to be used, we proceed to the fundamental step of preprocessing the inputted text.

We need to extract the important characteristics of the text (Facebook post, tweets, or other) in order to better classify it. We want to understand the context of the words in this step. In this paper, we rely on a document vocabulary representation that can capture the context of a word in a document, take into account semantic and syntactic similarity, and consider the relationship with other words. As a result, we chose BERT as a technique for the feature encoding step.

We can use deep learning to its full potential during the classification phase. Fake news detection models are trained to comprehend the context of the text and determine if the material is fake news or not. Long or short texts are both acceptable for analysis. If the messages are brief, the CNN, a subclass of deep and feed-forward artificial neural networks, might be an excellent choice for classifying fake news. As it shows how well it extracts regional and position-invariant features. Another deep learning recurrent neural network model based on Long Short Term Memory(LSTM) performs better for long texts since it can learn the long-term reliance of the text.

Therefore, a hybrid CNN-LSTM model is proposed for the purpose of classifying fake news in this work in order to combine the benefits of CNN and LSTM.

3.1 An Overview of the Proposed Model

Our BERT + CNN + LSTM model is composed of seven layers illustrated in Figure 2 and detailed as follows:

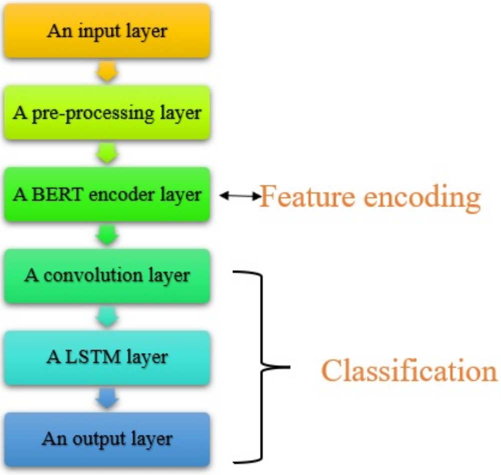

Figure 2: The proposed LSTM Model.

1. **An Input Layer:** which takes the plain text as input.
2. **A Pre-Processing Layer:** which takes the plain text as input and the result of the pre-processing is a batch of input sequences of fixed length for the encoder.
3. **A BERT Encoder Layer:** which takes the output of the previous layer as input and the result is a matrix of 128 vectors represents a text such that each vector of dimension 512 represents a sequence of text, then the output shape is (128, 512). It is a vector representation of each word in the text which takes into account the context and a vector representation for each text.
4. **A convolution Layer :** of dimension 1 that applies a filter to the input to detect and group features.The output shape is (128, 32)
5. **A LSTM Layer:** that retains long text important information.The output shape is (32).
6. **An Output Layer:** which represents the classifier that takes the output of the LSTM layer and the result are two binary classes: Fake or real news.The output shape is (2).

3.2 Text Pre-Processing Step

The first step is the dataset preparation step which includes:

- loading a dataset which is the input to our system.
- Basic pre-processing: When pre-processing the data, we:
  - Remove punctuation marks (such as @,!)
  - Remove the extra white space in the text.
- the division of the dataset into training and validation sets.

3.3 Feature Encoding Step

Once the data has been cleaned, formal feature encoding methods can be applied. In general, texts and documents are unstructured data sets. However, these unstructured text sequences need to be converted into structured form. The goal of text representation is to convert pre-processed texts into a form that the computer can process.

In our work, we propose to use a method in which each word of a vocabulary is mapped to a real vector considering its context to facilitate subsequently the detection of fake news. In this context, we propose to use the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). It is a language representation model that uses encoder-decoder architecture. The general structure of these neural networks processes language in sequence and the input passes through a layer called an "encoder" where the words are communicating information with each other for the model to generate a semantic structure. For the model to perform a task such text classification, the flow of information then passes through a "decoder" where it performs the reverse operation while it recovers the information from the "encoder". This type of neural network structure is known as Transformers (Gillioz et al., 2020) in NLP. The BERT model has two versions. The version we use in this work is BERTBASE with 12 layers with hidden layer sizes of 768 and 12 attention layers.The parameters are approximately 110M. Self-attention layers use dot-product similarity scores to augment each token's representation with information from other tokens that are similar or contextually informative. Crucially, self-attention layers are fully parallels because there is no recursiveness. BERT produces contextualized word embeddings where the vector for the word is computed based on the context in which it appears.

BERT differs from its predecessors (pre-trained NLP models) in the way it is pre-trained. This pre-training is unsupervised. BERT has been pre-trained using a self-monitoring technique on a sizable corpus of multilingual data. This indicates that it is an automated process that generates input and labels from texts that have merely been pre-trained on raw texts without any sort of human labeling. BERT is pre-trained on a large dataset consisting of texts from English Wikipedia pages (2500 million words) as well as a set of books (800 million words). This pre-training is done in two steps:

- *Step 1: MLM or Masked Language Modeling* Language Modeling is a common NLP task that consists of predicting the next word given the beginning of the sentence. The principle of Masked Language Modeling is to predict ”masked” tokens from the other tokens in the sequence.The model selects a sentence at random from the input, masking 15% of words, and then executes the entire sentence while predicting the words that were hidden. This is different from conventional endogenous neural networks (RNN), which often only observe one word at a time, and from isolated models like the GPT (Generative Pre-trained Transformer), which internally conceals tokens in the future. This allows the model to pick up on the sentence's dual representation.
- *Step 2: NSP or Next Sentence Prediction* This step consists of predicting whether or not a certain sequence A is followed by a certain sequence B. The NSP step allows BERT to understand the inter-dependencies between the sentences that follow each other. This enriches the final representation of the text. Using two masked sentences as input, the model is trained to connect the two sentences. It occasionally corresponds to subsequent sentences in the original text, but not always. The model should then foretell whether two sentences will follow one another. As a result, the model has acquired knowledge about how languages are internally represented in the training organization, which can be used to extract functions useful for subsequent work. For instance, if we have a data set with labeled sentences, we can train using functions from BERT-models generated as input.

The principle of using BERT is quite simple: it is ”already” pre-trained on a large quantity of data, we modify it for a precise task and then (re)train it with our own data. This allows BERT, in addition to its main function of text modelling in text classification tasks, to classify texts thanks to fine-tuning (Zhang et al., 2020). Fine-tuning is the process of adding a neural network to the output of BERT as needed. In our approach, we settle for BERT as a feature encoding technique.

3.4 Classification Step

According to (Mahrishi et al., 2020), when it comes to tackling real-world issues, deep neural networks do better than traditional machine learning techniques. It has been shown that neural network models are capable of impressive performance in text classification. For this kind of task, the two most popular designs are convolutional neural networks (CNN) and recurrent neural networks (RNN). In this paper, we propose combining the benefits and advantages of both architectural models to develop a new model for text classification. The convolutional neural network applies a non-linear activation function to the convolutional operation's output, and after pooling the data for classification, it uses a full connection layer. Filter, also referred to as the kernel function, is the fundamental component of convolutional operation.

The Long Short Term Memory network (LSTM), on the other hand, is a unique kind of recurrent neural network (RNN) that has the capacity to learn long-term dependencies. Cell state, the central component of the LSTM, allows for the addition or deletion of information from cells and the selective letting of information via the door mechanism. Remember gate, input gate, and output gate are the three gates that make up an LSTM. The input gate chooses what data should be updated to the cell state after the forget gate has decided which data should be removed from the cell state. The cell state is updated after establishing these two locations. The network's ultimate output is determined by the output gate.

In this instance, the LSTM layer keeps track of crucial text information (especially long ones). The classification process is completed by making a final determination regarding the text's authenticity using the output from the LSTM layer and the information that was retained.

4 RESULTS

We will first present the dataset we used before revealing the findings.

4.1 Dataset

To validate our approach, we used the dataset WELFake proposed in (Verma et al., 2021). It is a dataset of 72,134 news articles, 35,028 of which are true and 37,106 of which are false. The authors combined four popular news datasets (Kaggle, McIntire, Reuters, and BuzzFeed Political) to prevent classifier overfitting and to provide more text data for better ML training.

The dataset has four columns: serial number (beginning at 0), title (about the text news heading), text (about the news content), and label (0 = fake and 1 = real). The distribution of real and fake news is depicted in Figure 3. The dataset includes both lengthy and concise texts. The distribution of the news word count is shown in Figure 4.

70% of the data were used for training, 30% were used for testing, and 30% of the training set was used for validation.

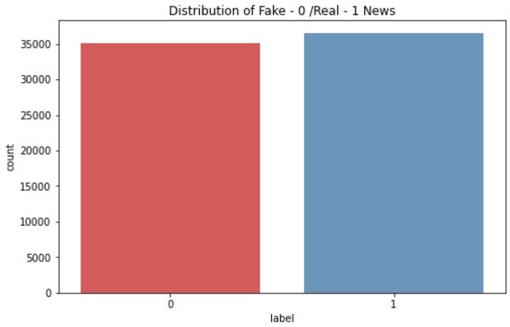

Figure 3: Distribution of real and fake news in the dataset.

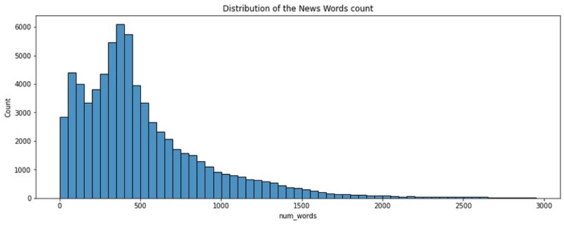

Figure 4: Distribution of the news word count in the dataset.

We start by preprocessing the dataset before moving on to feature encoding and determining whether or not the data is fake news. We outline the experimental setup and findings in the section that follows.

4.2 Experimental Settings

In this experiment, we use the following parameters.

- **Number of Filters:** 32 filters of size 8
- **Number of Units:** 32
- **Activation Function:** Softmax
- **Optimizer:** Adamw provided by TensorFlow
- **Loss Function:** cross-entropy. The cross-entropy compares the model's prediction with the label which is the true probability distribution. The cross-entropy goes down as the prediction gets more and more accurate. It becomes zero if the prediction is perfect.
- **The Evaluation Metric:** accuracy. Accuracy is the fraction of predictions our model got right.
- **Number of Epochs:** 5 (We use Stochastic Gradient Descent to get the best value of epochs).
- **Batch Size:** 16
- **Learning Rate:** $3e^{-5}$
- **Drop Out:** 0.1

4.3 The Obtained Experimental Results

The model based on BERT+CNN+LSTM was trained over 5 epochs, obtaining an accuracy of 1 and a loss of 2.4586e-4 on the training set, and an accuracy of 1 and a loss of 5.7634e-4 on the validation and test sets. Figure 5 shows the accuracy and loss curve for the model evaluated on the training and validation sets.

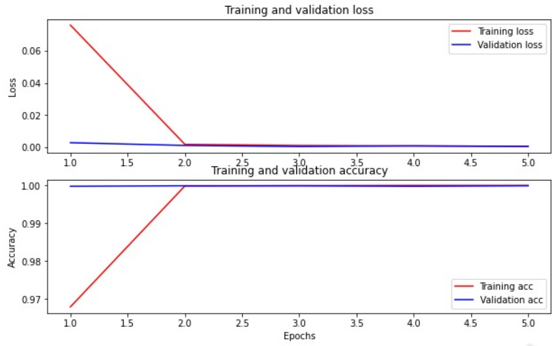

Figure 5: Evaluation of the BERT + CNN + LSTM model.

We note that the loss is minimal and the accuracy value is interesting.

4.4 Discussion

To validate the obtained experimental results, we tested the same dataset with other models. In Table 1, we sum up the accuracy of our model with other models of literature.

Table 1: Comparison to existing works.

Reference	Model	%Accuracy
(Vivek et al., 2017)	LinguisticAnalysis+SVM	87
(Ahmed et al., 2018)	TF-IDF+SVM	92
(Nicole et al., 2018)	DNN	93.50
(Kaliyar et al., 2021)	BERT+CNN	98.9
(Verma et al., 2022)	BERT+CNN+N-gram	99.01
Our model	BERT+CNN +LSTM	100

We find that our model surpasses all existing conventional methods, including TF-IDF for feature encoding and machine learning models for classification, like Ahmed et al. (Ahmed et al., 2018) who, with their best method, attained an accuracy of 92%.

Additionally, our model outperforms recent methods that use BERT for feature encoding and neural network models for classification, such as those proposed by Kaliyar et al. (Kaliyar et al., 2021) and Verma et al. (Verma et al., 2022), who respectively achieved accuracies of 98.90% and 99.01% .

5 CONCLUSION

In this paper, we propose developing a text mining and learning-based approach for detecting fake news. As a result, we proposed a method for classifying texts as true or false that combines CNN and LSTM deep learning with the BERT text feature encoding technique. When compared to other models, the experimental results of the proposed technique are extremely promising.

To identify fake news in this study, we relied solely on the text's content. We intend to expand on this approach in future works by taking into account additional factors such as the source (the original author), the source's characteristics (false profiles, robots, malicious people, etc.), and so on.

REFERENCES

Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer.

Ahmed, H., Traore, I., and Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Bengio, Y. (2019). La révolution de l'apprentissage profond. *Interstices*.

De Sarkar, S., Yang, F., and Mukherjee, A. (2018). Attending sentences to detect satirical fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3371–3380.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gillioz, A., Casas, J., Mugellini, E., and Abou Khaled, O. (2020). Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Kalsnes, B. (2018). Fake news. In *Oxford Research Encyclopedia of Communication*.

Kaur, P., Boparai, R. S., and Singh, D. (2019). A review on detecting fake news through text classification. *Int. J. Electron. Eng.*, 11(1):393–406.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Klein, D. and Wueller, J. (2017). Fake news: A legal perspective. *Journal of Internet Law (Apr. 2017).*

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. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

LeCun, Y. (2016). L'apprentissage profond, une révolution en intelligence artificielle. *La lettre du Collège de France*, (41):13.

Mahrishi, M., Hiran, K. K., Meena, G., and Sharma, P. (2020). *Machine Learning and Deep Learning in Real-Time Applications*. IGI global.

Mustafaraj, E. and Metaxas, P. T. (2017). The fake news spreading plague: was it preventable? In *Proceedings of the 2017 ACM on web science conference*, pages 235–239.

Nicole, O., Sophia, L., Georgios, E., and Xavier, B. (2018). The language of fake news: Opening the black-box of deep learning based detectors. In *32nd Conference on Neural Information Processing Systems*.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Quandt, T., Frischlich, L., Boberg, S., and Schatto-Eckrodt, T. (2019). Fake news. *The international encyclopedia of journalism studies*, pages 1–6.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Verma, P. K., Agrawal, P., Amorim, I., and Prodan, R. (2021). Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Verma, P. K., Agrawal, P., Madaan, V., and Prodan, R. (2022). Mcred: multi-modal message credibility for fake news detection using bert and cnn. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13.

Vivek, S., Rupanjal, D., Darshan, S., Karthik, R., , and Isha, G. (2017). Automated fake news detection using linguistic analysis and machine learning. In *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation(SBP-BRiMS)*, pages 1–3.

Xu, D., Tian, Z., Lai, R., Kong, X., Tan, Z., and Shi, W. (2020). Deep learning based emotion analysis of microblog texts. *Information Fusion*, 64:1–11.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2020). Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837.