applied sciences

MDPI

Article

exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT)

Heejung Jwa [1], Dongsuk Oh [2], Kinam Park [3], Jang Mook Kang [4] and Heuiseok Lim [1,*]

[1] Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea; jwaheejung@korea.ac.kr
[2] Human-Inspired AI & Computing Research Center, Korea University, Seoul 02841, Korea; inow3555@gmail.com
[3] Creative Information & Computer Institute, Korea University, Seoul 02841, Korea; spknn@korea.ac.kr
[4] Department of Convergence Contents, Global Cyber University, Seoul 02841, Korea; honukang@gmail.com
* Correspondence: limhseok@korea.ac.kr; Tel.: +82-2-3290-2396

Received: 4 September 2019; Accepted: 24 September 2019; Published: 28 September 2019

Abstract: News currently spreads rapidly through the internet. Because fake news stories are designed to attract readers, they tend to spread faster. For most readers, detecting fake news can be challenging and such readers usually end up believing that the fake news story is fact. Because fake news can be socially problematic, a model that automatically detects such fake news is required. In this paper, we focus on data-driven automatic fake news detection methods. We first apply the Bidirectional Encoder Representations from Transformers model (BERT) model to detect fake news by analyzing the relationship between the headline and the body text of news. To further improve performance, additional news data are gathered and used to pre-train this model. We determine that the deep-contextualizing nature of BERT is best suited for this task and improves the 0.14 F-score over older state-of-the-art models.

Keywords: fake news; fake information; fake news detect; fake news challenge; fake news classification; deep learning

1. Introduction

Fake information appears in a variety of forms, including videos, audio, images, and text. Furthermore, fake information in text form can be classified as news, social network services, speeches, documents, and so on. This study proposes a model for fake news detection by focusing on text-based fake news. Fake news has recently become a widespread problem worldwide. Fraudulent or falsified information can spread rapidly and become a problem if readers fail to detect, at a glance, whether or not the given information is fake news.

In 2015, the International Fact-Checking Network (IFCN) was established by Poynter, an American media education agency. IFCN observes fact check trends and provides training programs for fact checkers. In addition, various efforts have been undertaken to prevent the spread of fake news by providing a code of principles that fact check organizations around the world can use. Politifact (https://www.politifact.com) and snopes (https://www.snopes.com) developed a Fake news detection tool to classify the level of fake news in stages based on the presented criteria. However, these tools are time-consuming and expensive as they require manual work and judgment. Therefore, a model that automatically detects fake news is required.

There are several tasks involved in detecting fake news. Fake news challenge stage 1 (FNC-1) (www.fakenewschallenge.org) involves classifying the stance of a body text from a news article relative

to a headline. The body text may agree, disagree, discuss or be unrelated to the headline. The Web Search and Data Mining (WSDM) 2019 fake news challenge (www.kaggle.com/c/fake-news-pair-classification-challenge) detects fake news by classifying the title of a news article. Given the title of a fake news article A and the title of a coming news article B, participants were asked to classify B into one of the three categories [1–3]. In addition, there are other tasks involved in fake news detection using Social Network Services (SNS) data. This task of the challenge (www.clickbait-challenge.org) is to develop a classifier that rates how clickbaiting a social media post is [4–6]. Finally, one of the tasks in the artificial intelligence research and development challenge (www.ai-challenge.kr) in Korea involves the detection of unrelated contextual news content in the news bodytext. This is done to prevent a situation in which the reader is exposed to unintended information.

However, the results of these challenges and tasks do not exhibit good performance and can be improved. In addition, word embedding is an important factor for improving the performance of the model. word2vec [7] and fastText [8] were previously used for word embedding. These do not exhibit good performance because they employ fixed vector values rather than fluid vector values for words. To complement this, we use contextual word embedding. Typical models include Embeddings from Language Models (ELMO) [9], BERT [10], and Generative Pre-Training (GPT) [11]. ELMO uses a bi-LSTM structure [12] and a feature-based approach that requires the application of many hyperparameters. The feature-based approach includes a pre-trained language representation of an additional feature of networks that perform specific tasks. On the other hand, BERT and GPT used a transformer structure and a fine-tuning approach to minimize the hyperparameters. The fine-tuning approach involves reducing the parameter of a specific task as much as possible and slightly changing the pre-trained parameters by training downstream tasks. We considered the above factors and applied fake news data to the BERT model, allowing us to better analyze the meaning of the news article.

In this paper, we propose BAKE, an automatic fake news detection model that improves upon BERT by mitigating the data imbalance problem. Furthermore, we also propose a model that incorporates extra unlabeled news copora into BAKE, which we term exBAKE.

Our key contributions are summarized as follow:

- We use BERT [10], which was the first to study on fake news detection using a headline-body text dataset. BERT includes pre-training language representations developed by Google.
- We recognize that the data are unstable, and therefore to develop the BAKE model to classify the data using weighted cross entropy (WCE).
- We include CNN and Daily Mail news data for BAKE pre-training. Greater amounts of news data are used to detect fake news more efficiently.
- Finally, we evaluate the performance of the proposed model exBAKE, and demonstrate that it performs better than other models that use FNC-1 data.

The remainder of this paper is organized as follows. In Section 2, we present an overview of the related works. Section 3 presents the data used in this study. In Section 4, we analyze the structures of the proposed model, and, in Section 5, we present the experiments and their results. Finally, we conclude the paper and highlight several future research directions in Section 6.

2. Related Works

Fake news detection has been examined using several methods in accordance with the scope and format of available fake news data and technical approaches [13–15]. Tasks to be performed on fake news datasets include verifying whether the headline matches the body text, finding mismatched sentences in the body text, and identifying dissemination of fake news over SNS. In this study, we use a data set for the first task. Such methods generally employ techniques like deep learning [16,17], machine learning [18,19], or rule-based methods [20] for the purpose of detection. In this study, we use deep learning to train the data.

A majority vote includes FNC-1 baseline features. The co-occurrence (COOC) of character n-grams and word from the document, headline and two lexicon based features (i.e., polarity (POLA) words and

count the number of refuting (REFU) based on small word lists used gradient-boosting baseline from which the FNC-1 organizers are provided. By observing the FNC-scores of the system performance, it is determined that both lexicon-based features and the majority vote baseline operate comparably. On the other hand, COOC exhibits comparatively better performance.

In the fake news challenge, the first place was secured by the SWEN team that used TalosComb. The TalosComb model is an average weighted model of TalosCNN and TalosTree [21]. Talostree is based on the gradient-boosted decision tree model, which consists of singular-value decomposition (SVD) [22], word count, term frequency-inverse document frequency (TF-IDF) [23], and sentiment features using word2vec embeddings. TalosCNN is based on deep convolutional neural networks and uses pre-trained word2vec embeddings. It uses several convolutional layers comprising three fully-connected ones and a final softmax layer for classification.

In the fake news challenge, the second place was secured by the Athene team that proposed a multi-layer perceptron (MLP) [24]. It extends the original model structure to six hidden layers and one softmax layer and incorporates multiple manually engineered features from the Athene team—namely, unigrams, the cosine similarity of word embeddings of verbs and nouns between document tokens and headlines, latent Dirichlet allocation, topic models based on non-negative matrix factorization, and latent semantic indexing. Moreover, the baseline features are provided by the FNC-1 organizers. These feature types either form separate feature vectors or a joint feature vector [25]. From the results obtained from the FNC-1 test dataset, it is determined that featMLP exhibits good overall performance but is still not the best. As with other systems, there were multiple differences between the performance of featMLP during development and on test datasets. Owing to the 100 new topics, which had not been included in the training dataset, the stackLSTM model was merged with a stacked long short-term memory (LSTM) [26] network and the best feature set was derived from an ablation test [27]. The DSC class performs the best in terms of stackLSTM. This is an important feature of stackLSTM. As there are a few instances of the DSG class, DSG is difficult to classify. In other words, stackLSTM correctly detects more complex negation instances.

In the fake news challenge, the third place was secured by the UCL Machine Reading (UCLMR) team. The team suggested MLP using a single hidden layer [28]. The team used a term frequency (TF) and TF-IDF to express text inputs. TF vectors were extracted from a vocabulary of the 5000 most frequently occurring words in the training set, and the TF-IDF vectors were obtained from a vocabulary of the 5000 most frequently occurring words in both the training and test datasets.

The result of using BERT in this paper is higher performance than existing models. Since news data are composed of various words and sentences, it is important to clearly understand the relationship between words for accurate analysis. BERT is designed to clearly identify the relationship between words in a sentence. BERT adopts semi-supervised learning and a language representation model that uses only the encoder portion of the transformer [29]. In particular, BERT is based on a multi-layer bidirectional transformer encoder that jointly conditions both the left and the right contexts in all layers. BERT performs pre-training using an unsupervised prediction task, which includes a masked language model (MLM) and a next sentence predictor. MLM is about understanding context first and then predicting words. First, we randomly mask several tokens at 15% probability from the word piece applied input. The input is included in the Transformer structure to predict the masked words based on the context of the surrounding words. Through these processes, BERT understands the context more accurately. The next sentence predictor is for identifying the relationship between sentences. This task is important for language understanding tasks such as Question Answering (QA) or Natural Language Inference (NLI). BERT includes a binarized next sentence prediction task, which combines the two sentences in corpus with the original sentence. This model structure allows BERT to perform very well in various NLP tasks. We use BERT base model, which includes a base model and a large model, and, depending on the size of the model, we use a different number of layers for the transformer block, a hidden size, and self-attention heads. Data used in the BERT model comprises 800 M words from the Book Corpus and 2500 M words from Wikipedia.

3. Data

In this study, we included the CNN (www.cnn.com) and Daily Mail (www.dailymail.co.uk) datasets (https://github.com/abisee/cnn-dailymail) for additional learning during BERT's pre-training stage to improve its detection capabilities. The data that the training and testing processes employ are used to fine-tune news articles. BERT's pre-training exhibits good performance in other prior natural language processing (NLP) tasks [30–32]. However, the data used in BERT model is based on 2500 M words of general data obtained from Wikipedia and 800 M words from the Book Corpus. While this data includes a wide range of information, it still lacks detailed information on individual domains. To address this shortcoming, news data have been added at the pre-training level in this study to improve its fake news detection capabilities. Summarization data [33–36] from CNN included approximately 90,000 documents and 380,000 questions (with a vocabulary size of 118,497), while the Daily Mail dataset included 197,000 documents and 879,000 questions (with a vocabulary size of 208,045). The CNN articles were collected from the period between April 2007 and the end of April 2015 from the CNN website. Daily Mail articles were collected from the period between June 2010 and the end of April 2015 from the Daily Mail website [37].

FNC-1 data were used for fine-tuning. The training set comprises corresponding pairs of headlines and body texts, including the appropriate class label for each pair. In addition, the test set comprises pairs of the headlines and body texts without class labels to help evaluate the systems. In total, 2587 headlines and 2587 body texts were used, and the data can be found at the FNC-1 github (https://github.com/FakeNewsChallenge/fnc-1).

4. Methods

The model we propose is presented in Figure 1. It is comprised primarily of two parts. In the fine-tuning process, we used WCE [38–40] to classify the dataset into four groups: Agrees (AGR), Disagrees (DSG), Discusses (DSC), and Unrelated (UNR). Even though this is essentially a BERT model, we call it BAKE because we first applied it to the task of fake news detection in our case. In the pre-training process, we experimented with extra the CNN and Daily Mail news data to our BAKE model, creating the exBAKE model.

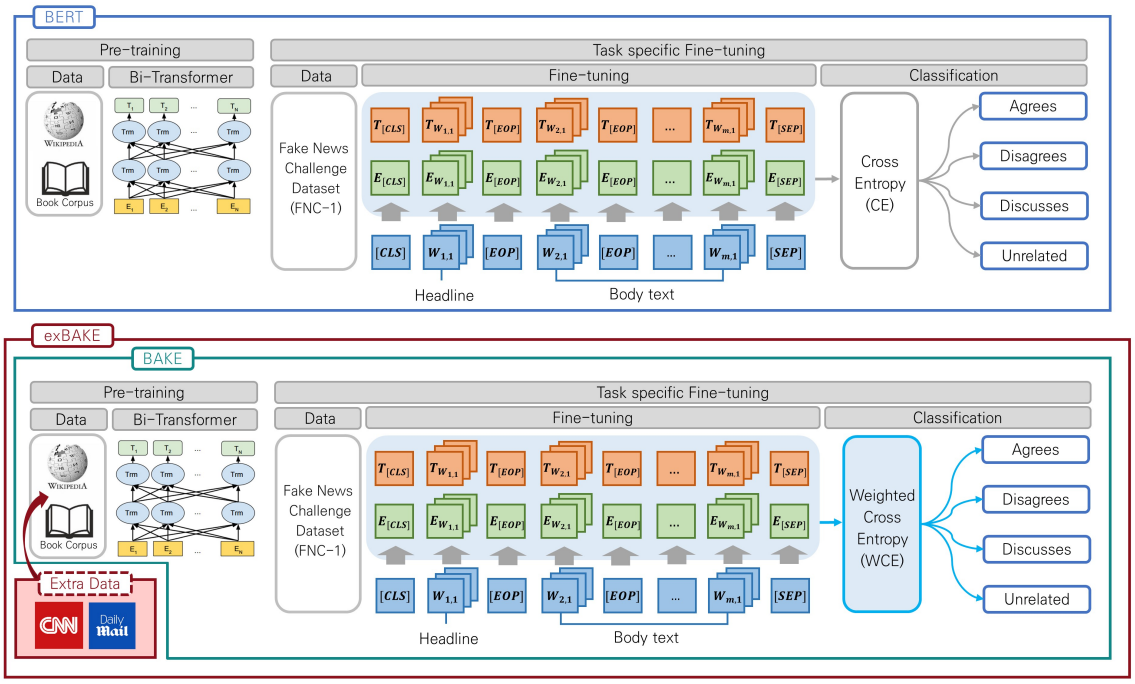

Figure 1. This is a figure that describes BAKE and exBAKE proposed by us.

In other words, the data were classified into four multi-classes using Linear and Softmax layers, and WCE was used as training loss. This training loss, as indicated below, is characterized by different weight values according to the corpus statistics and the label distribution analyzed for each class. The FNC-1 dataset is unbalanced to AGR 7.4%, DSG 2.0%, DSC 17.7%, and UNR 72.8%. Cross Entropy (CE) [41,42] is the most widely used loss function. It is a method of calculating the amount of information existing between two probability distributions, the true probability P and the predicted probability Q:

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} w_i \cdot L_{CEi}, \tag{1}$$

$$w_i = \frac{\sum_{i=1}^{N} x_i}{x_i}, \tag{2}$$

$$L_{CEi} = -[y_i log(p_i) + (1 - y_i) log(1 - p_i)]. \tag{3}$$

We denote news sentences containing headline and body text by si = wi, 1, ... wi, li and denote the input of BERT by x = ([CLS], s1, [EOP], s2, [EOP], ..., [SEP]). As in the input, the news sentences, including the headlines and body texts, were divided based on the tag (EOP) (i.e., End Of Paragraph).

5. Experiment

5.1. Evaluation Method

The FNC organizers proposed the hierarchical evaluation metric FNC, but they did not consider the fact that the FNC-1 dataset is very imbalanced. Achieving a high score in FNC is not difficult, since only performing well on the majority task (UNR) and randomly predicting on the others would still lead to a good score. Therefore, the hierarchical evaluation metric FNC is inappropriate for validating the document-level stance detection task [25].

In this study, we used the macro-averaged F1-score (F1) evaluation method [43]. F1 can be interpreted as a weighted average of precision and recall. F1 calculates the metrics for each label and obtains the unweighted mean. It is obtained using the formula given below:

$$Precision = \frac{Precision1 + Precision2}{2}, \tag{4}$$

$$Recall = \frac{Recall1 + Recall2}{2}, \tag{5}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{6}$$

5.2. Comparative Model

The results are reported in Table 1. We prove that BERT is best suited for this task due to its deep contextualizing nature. We already outperformed our previous models by applying BERT, and further improved BAKE performance by using WCE, and exBAKE is a state-of-the-art result of learning more news data.

We compare the performance of our methods with previous approaches in Table 1. BAKE and exBAKE surpass the performance of the previous state-of-the-art (stackLSTM) by 0.125 and 0.137 F1 scores, respectively. The proposed methods also surpasses BERT, which is already a strong baseline. This indicates that the use of WCE is crucial for showing competitive results in fake news detection. The exBAKE model has shown the best overall F1 score, which suggests that incorporating extra knowledge from large news corpora is beneficial to this task.

Table 1. Model performance. We improve the 0.14 F-score over prior state-of-the-art results.

Models	F1	AGR	DSG	DSC	UNR
Majority vote	0.210	0.0	0.0	0.0	0.839
TalosComb [21]	0.582	0.539	0.035	0.760	0.994
TalosTree [21]	0.570	0.520	0.003	0.762	0.994
TalosCNN [21]	0.308	0.258	0.092	0.0	0.882
Athene [25]	0.604	0.487	0.151	0.780	**0.996**
UCLMR [28]	0.583	0.479	0.114	0.747	0.989
featMLP [24,25]	0.607	0.530	0.151	0.766	0.982
stackLSTM [25,27]	0.609	0.501	0.180	0.757	0.995
BERT	0.656	0.651	0.145	**0.839**	0.989
BAKE	0.734	0.667	0.463	0.822	0.986
exBAKE	**0.746**	**0.684**	**0.501**	0.813	0.988
Upper bound	0.754	0.588	0.667	0.765	0.997

Overall, our method was able to achieve state-of-the-art performance in three out of four news categories. Furthermore, the proposed methods surpassed the upper bound (the score of human annotators), in AGR and DSC. The performance difference is most dramatic in the minority categories, again demonstrating that WCE plays an important role in overcoming the data imbalance problem. This comes in the price of a small drop in performance in the majority categories. Still, the performance on these majority categories are superior or comparable to previous state-of-the-art.

6. Conclusions

A majority of the data collected for fake news detection are written in English. As the spread of fake news has a negative impact on society, several studies have been conducted, and numerous technologies have been introduced to deal with such falsified texts. It is imperative to enable readers to distinguish between real and fake news.

In this study, we proposed an improved exBAKE model by using pre-training based on a BERT model to accurately understand the contents of such articles. The results indicate that the model worked best on the FNC-1 dataset, which detected fake news by analyzing the relationships between headlines and the corresponding body texts of news articles.

No automated tools had previously been created to check the authenticity of a news article in real time. Our proposed model will help readers and other journalists to avoid having to manually go through the process of distinguishing fake news from real news.

In the future, we will experiment with various cases of fake news detection tasks using the pre-trained BERT model proposed in this study. We only analyzed the relationship between the headline and the body text of an article. Further experimentation is needed to apply data from other fake news detection tasks to BERT model, which will use additional news data in the pre-training phase.

Author Contributions: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft preparation, and writing—review and editing, H.J. and D.O.; validation, supervision, resources, project administration and funding acquisition, K.P., J.M.K. and H.L.

Acknowledgments: This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (IITP-2018-0-00705, Algorithm design and software modeling for judge fake news based on artifical Intelligence. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. NRF-2017M3C4A7068189).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pham, L. Transferring, Transforming, Ensembling: The Novel Formula of Identifying Fake News. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019.
2. Liu, S.; Liu, S.; Ren, L. Trust or Suspect? An Empirical Ensemble Framework for Fake News Classification. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019.
3. Yang, K.C.; Niven, T.; Kao, H.Y. Fake News Detection as Natural Language Inference. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019.
4. Omidvar, A.; Jiang, H.; An, A. Using Neural Network for Identifying Clickbaits in Online News Media. In *Annual International Symposium on Information Management and Big Data, Lima, Peru, 3–5 September 2018*; Springer: Cham, Switzerland, 2018; pp. 220–232.
5. Zhou, Y. Clickbait detection in tweets using self-attentive network. *arXiv* **2017**, arXiv:1710.05364.
6. Grigorev, A. Identifying clickbait posts on social media with an ensemble of linear models. *arXiv* **2017**, arXiv:1710.00399.
7. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. 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 December 2013; pp. 3111–3119.
8. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
9. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf (accessed on 20 September 2019).
12. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
13. Wang, W.Y. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.00648.
14. Ruchansky, N.; Seo, S.; Liu, Y. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; ACM: New York, NY, USA, 2017; pp. 797–806.
15. Kochkina, E.; Liakata, M.; Augenstein, I. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv* **2017**, arXiv:1704.07221.
16. Popat, K.; Mukherjee, S.; Yates, A.; Weikum, G. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *arXiv* **2018**, arXiv:1809.06416.
17. Yang, Y.; Zheng, L.; Zhang, J.; Cui, Q.; Li, Z.; Yu, P.S. TI-CNN: Convolutional neural networks for fake news detection. *arXiv* **2018**, arXiv:1806.00749.
18. Rasool, T.; Butt, W.H.; Shaukat, A.; Akram, M.U. Multi-Label Fake News Detection using Multi-layered Supervised Learning. In Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, Perth, Australia, 23–25 February 2019; ACM: New York, NY, USA, 2019; pp. 73–77.
19. Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; Liu, H. Unsupervised fake news detection on social media: A generative approach. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
20. Feng, S.; Banerjee, R.; Choi, Y. Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers—Volume 2, Jeju Island, Korea, 8–14 July 2012; pp. 171–175.

21. Sean, B.; Doug, S.; Yuxi, P. Talos Targets Disinformation with Fake News Challenge Victory. 2017. Available online: https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html (accessed on 20 September 2019).
22. De Lathauwer, L.; De Moor, B.; Vandewalle, J. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1253–1278. [CrossRef]
23. Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, Piscataway, NJ, USA, 3–8 December 2003; Volume 242, pp. 133–142.
24. Davis, R.; Proctor, C. Fake news, real consequences: Recruiting neural networks for the fight against fake news. 2017. Available online: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761239.pdf (accessed on 20 September 2019).
25. Hanselowski, A.; PVS, A.; Schiller, B.; Caspelherr, F.; Chaudhuri, D.; Meyer, C.M.; Gurevych, I. A retrospective analysis of the fake news challenge stance detection task. *arXiv* **2018**, arXiv:1806.05180.
26. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Hermans, M.; Schrauwen, B. Training and Analysing Deep Recurrent Neural Networks. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2013; pp. 190–198.
28. Riedel, B.; Augenstein, I.; Spithourakis, G.P.; Riedel, S. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv* **2017**, arXiv:1707.03264.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Diaz, M.; Ferrer, M.A.; Impedovo, D.; Pirlo, G.; Vessio, G. Dynamically enhanced static handwriting representation for Parkinson's disease detection. *Pattern Recognit. Lett.* **2019**, *128*, 204–210. [CrossRef]
31. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. Biobert: Pre-trained biomedical language representation model for biomedical text mining. *arXiv* **2019**, arXiv:1901.08746.
32. Liu, Y.; Lapata, M. Text Summarization with Pretrained Encoders. *arXiv* **2019**, arXiv:1908.08345.
33. See, A.; Liu, P.J.; Manning, C.D. Get to the point: Summarization with pointer-generator networks. *arXiv* **2017**, arXiv:1704.04368.
34. Liu, Y. Fine-tune BERT for Extractive Summarization. *arXiv* **2019**, arXiv:1903.10318.
35. Liu, L.; Lu, Y.; Yang, M.; Qu, Q.; Zhu, J.; Li, H. Generative adversarial network for abstractive text summarization. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
36. Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. *arXiv* **2017**, arXiv:1705.04304.
37. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2015; pp. 1693–1701.
38. Yang, K.; Lee, D.; Whang, T.; Lee, S.; Lim, H. EmotionX-KU: BERT-Max based Contextual Emotion Classifier. *arXiv* **2019**, arXiv:1906.11565.
39. Aurelio, Y.S.; de Almeida, G.M.; de Castro, C.L.; Braga, A.P. Learning from imbalanced data sets with weighted cross-entropy function. *Neural Process. Lett.* **2019**, 1–13, doi:10.1007/s11063-018-09977-1. [CrossRef]
40. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: New York, NY, USA, 2017; pp. 240–248.
41. Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; Srebro, N. Exploring generalization in deep learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5947–5956.

42. Zhang, Z.; Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 8778–8788.
43. Shang, W.; Huang, H.; Zhu, H.; Lin, Y.; Qu, Y.; Wang, Z. A novel feature selection algorithm for text categorization. *Expert Syst. Appl.* **2007**, *33*, 1–5. [CrossRef]