Università degli studi di Milano-Bicocca

Advanced Machine Learning

Final Project

Fake News Detection

Authors:
Mattia Piazzalunga - 851931
Nicolò Urbani - 856213

February 18, 2024

Abstract

”Fake news” are a central issue in today’s society, a phenomenon intensifid by the advent of social media which has facilitated their dissemination. Automatically detecting them, therefore, becomes a fundamental challenge involving research. In this paper, the state-of-the-art has been achieved on two benchmark datasets for ”fake news detection”: ISOT and WELFake, achieving 100% accuracy in classifying news from the former and 97% from the latter. Additionally, a custom neural network was built, which, with a 99.96% parameter reduction, achieved accuracy close to the most performant model studied in this analysis, based on BERT. Back Translation and Dataset Combination proved unhelpful techniques in attempting to improve model generalization for this type of task.

1 Introduction

”Fake news” represent a growing problem in today’s society. “News that conveys or incorporates false, fabricated, or deliberately misleading information, or that is characterized as or accused of doing so” is reported by the Oxford English Dictionary [1], dating the term back to 1890.

This phenomenon has been exacerbated by the advent of social media and digital technologies, which have facilitated the rapid spread of unverified information. Fake news can have significant impacts on political, social, and economic issues, influencing public opinion, elections, and even trust in institutions.

In a world where fake news is an evident and central issue, according to a December 2021 survey by Eurostat [2], only 23% of individuals aged 16 to 74, in the three months prior to the survey, verified the accuracy of information or content, despite 47% of them encountering false or dubious information from news websites or social media.

Automatically detecting fake news, therefore, becomes a fundamental challenge that heavily involves research and can mitigate their impacts on society.

This report analyzes the language used in news articles in an attempt to identify and counter them. Drawing on two benchmark datasets for this important task, different NLP pipelines are employed in an attempt to process, study, and classify news based on the language involved. To achieve state-of-the-art performance on the datasets under consideration, it becomes crucial to use state-of-the-art techniques for NLP and fake news detection, analyzing which of these techniques lead to the best results in terms of prediction accuracy, while considering training and prediction times, as well as model size.

2 Datasets

Two important datasets are employed in this study, recognized benchmarks for fake news detection:

- ISOT dataset [3]. In 2017, at the "International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments," Ahmed et al. presented the ISOT dataset, generated by collecting 21.417 "Real News" from the news site Reuters.com and 23.481 "Fake News" from a dataset on Kaggle.com. The fake news were classified by a collaboration between Facebook and Politifact, a fact-checking organization operating in the USA. The articles, in general, concern political news written around 2016, with a length of at least 200 characters. For each article, the following features are available: "Article Text," "Article Type," "Article label" (fake or truthful), "Article Title," and "Article Date."

- WELFake dataset [4]. In 2021, in the journal "IEEE Transactions on Computational Social Systems," Verma et al. presented the WELFake dataset, generated by collecting 72,134 articles from various sources (i.e., Kaggle, McIntire, Reuters, and BuzzFeed Political), of which 35.028 are "Real News" and 37.106 are "Fake News." The dataset was created with a large number of news articles to prevent model overfitting and generate unbiased outputs. For each article, the following features are available: "Label," and "Text."

How to process datasets to get to prediction? Compared to general ML/AI pipelines, in NLP it is necessary to perform some additional processing stages to enable machines to understand text. Despite the learning power of current deep learning algorithms, using an NLP pipeline for data preprocessing remains a fundamental step in the learning process. Some initial steps involve, centrally, preprocessing the data present in the datasets. The articles in ISOT and WELFake are "raw", and without the use of specific preprocessing techniques, this could compromise the model's understanding of the text. It is necessary, therefore, to remove noise, as well as to normalize the text with techniques tailored to what is being analyzed:

1. Lowercasing. This step is used to convert all the text to lowercase letter and is, generally, the first preprocessing step performed.

2. Removing NaN values and merging title and text columns. This step is used to merge the "Text" and "Title" columns of the ISOT dataset and make the text itself more informative. Furthermore, it is crucial to remove any NaN values encountered during this process, as they could negatively influence ML algorithms.

3. Removing spaces before and after texts, transformation of multiple spaces into one & removing NaN.

4. Removing CDATA. The acronym CDATA (Character Data) is used in some markup languages to delimit blocks of text data that should not be interpreted as markup code. Within the articles under examination, there are many "CDATA" sections that do not aid semantic understanding and, for this reason, must be removed.

5. Removing URLs.

6. Replacement of HTML entities with standard syntactic characters. HTML entities are character sequences used in HTML documents to represent symbols, special characters, or reserved characters. *For example, the entity "&amp" corresponds to the & character in natural language.* It is crucial to replace such entities with their corresponding symbols.

7. Removing emails.

8. Removal of mentions. Many of the news articles present in the datasets have been collected from Twitter posts and, for this reason, contain mentions of user profiles that need to be removed.

9. Removing html tags. In HTML, tags are elements used to define the structure and meaning of the content on a web page. Although their content is crucial for semantic interpretation, the tags themselves are superfluous.

10. Removing special characters. Emotions, symbols, and special characters are to be removed.

11. Replacement of slang, acronyms and abbreviations. In informal language, slang, acronyms, and word abbreviations are often used, which can confuse machine learning algorithms. In this paper, we present a list associating a "translation" to slang/acronyms/terms, attempting to disambiguate their semantic meaning. The list is created ad-hoc for this paper, after gathering slang/acronyms/terms from various sources on the web.

12. Expansion of contradictions. The term "contraction" refers to multiple words abbreviated into one. *For example, "can't" is a contraction of "can not".* These "contradictions" need to be resolved to facilitate text comprehension.

13. Adding spaces after punctuation marks. Very often, in informal writing, there is a tendency to forget to add a "space" after punctuation marks before the next word. This practice can mislead tokenization algorithms and should therefore be resolved.

14. Removing punctuation marks. The correct use of punctuation marks is essential to ensure clarity and precision in written and oral communication. Well-punctuated text makes it easier for the reader to understand the meaning and intention of the author. However, punctuation marks are not useful for automatic comprehension purposes.

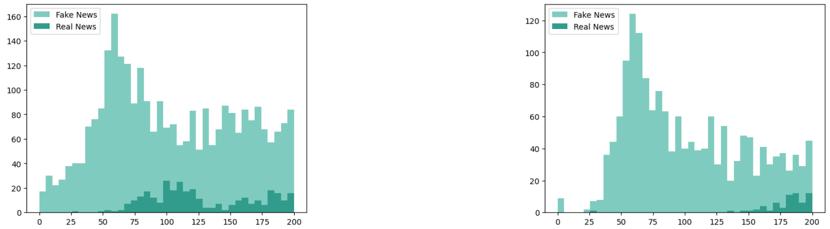
15. Transformation of multiple spaces into one.

16. Conversion of numbers into words. In writing, very often, the representation of numbers in digits or words is used indiscriminately. This, however, complicates automatic interpretation.

The highlighted steps are chosen and studied for the datasets under examination. It is important to execute them in the order in which they are listed. *For example, removing punctuation marks before removing CDATA would compromise this second point by affecting the formatting of the CDATA itself.*

An additional step, often criticized, is related to the removal of stop words. Stop words are common words that are often ignored or removed during text analysis in computational linguistics. These words are generally considered uninformative because they frequently appear in various contexts without providing significant contribution to the text's semantics. This step is often indicated as optional because the removal of stop words does not always bring benefits. Certainly, in contexts like web searching, stop words play a fundamental role, but their importance depends on the task. According to a 2019 study by Qiao et al. [5], in the presence of neural models like BERT (used in this study), contrary to common belief, even though stop words receive equal "attention" to non-stop words, their removal has no effect: the model learns their uselessness by assigning redundant attention weights to them. Therefore, in the analysis of this paper, stop words are removed.

To understand the difficulty of the task at hand, it is useful to observe the composition of the datasets.
An exploratory analysis, on the preprocessed datasets, revealed that the distribution of the target variable is balanced in both datasets, WELFake and ISOT, with approximately 50 % comprising fake news and 50% real news. The instances in the datasets are evenly distributed with respect to the two classes of the target variable. Balancing improves the model's learning, preventing bias towards the better-represented classes and promoting convergence.



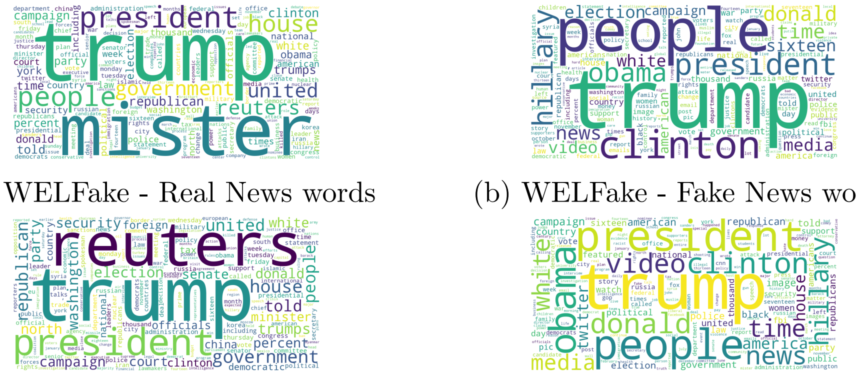
(a) WELFake

(b) ISOT

Figure 1: News length comparison

From Figure 1, it emerges that fake news tends to be longer than real news. However, there is no official scientific research that highlights this trend. In any case, it is important to pay attention to what is observed in Figure 1: one must ensure that the model does not learn to distinguish news based on their length without considering informative characteristics.
One last aspect to study, which can underline the difficulty of the problem

at hand, is the words, and their frequency, within the news.



(a) WELFake - Real News words
(b) WELFake - Fake News words
(c) ISOT - Real News words
(d) ISOT - Fake News words

Figure 2: Frequency of words

From Figure 2, it is possible to notice that the news has a predominantly political background, with some words like "Trump" which are not useful for discrimination since they are frequent in both real and fake news. In this context, it is important to mention two laws:

- Zipf's law, first expressed in 1949 by Zipf et al. in the book "Human Behaviour and the Principle of Least Effort" [6], is an empirical law that describes the distribution of word frequencies in a natural text. It is based on the observation that few words in a text are extremely common, while most words are very rare. To be more precise, the frequency of a word is inversely proportional to its rank in the list of most frequent words.

- From Zipf's law derives Luhn's analysis, described in 1958 by Luhn in the "IBM Journal of Research and Development" [7]. It is a technique aimed at identifying and extracting relevant information from large volumes of text or data. The main idea is as follows: the frequency with which some words appear in a text provides an important indication of their meaning. Additionally, their position in the sentence is another important parameter for semantic interpretation.

To summarize, rare words are discardable because they are "rare," while common words are not informative.

In this study, many of the common words were removed in the "stop word removal" process. However, the removal of words based on Luhn's analysis was not applied: the chosen embedding techniques autonomously handle this aspect.

3 The Methodological Approach

After considering the datasets and the preprocessing phase, in the central part of this analysis, it becomes crucial to employ state-of-the-art techniques in NLP and Fake News Detection to properly categorize the news under examination. The subsequent steps of the NLP pipeline, the core of this study, are enriched with models and best practices that have been experimentally demonstrated as the best for the task at hand. What are the next steps?

1. Word tokenization. Tokenization is the process of breaking down a text into smaller units called "tokens". A token can be a word, a phrase, a symbol, a single letter, or any meaningful element depending on the context and purposes of the tokenization process.

2. POS tags + Lemmatization. POS (Part-of-Speech) refers to the grammatical category of a word in a sentence. The four grammatical categories considered in this study are: nouns, verbs, adjectives, and adverbs. Assigning POS to each word in a sentence helps understand its grammatical structure. Lemmatization, on the other hand, is the process of reducing a word to its base form or canonical form, called a lemma, to address grammatical variations and verb conjugations, thereby simplifying text analysis. Some lemmatizers use the word's POS tag, in addition to its spelling, to enhance accuracy in the lemmatization process: adding information about the part of speech is helpful in understanding the different word forms based on their usage context.

3. Word embedding. Word embeddings are a particular form of representation aimed at capturing the semantic relationships between words based on their context. These embeddings are dense vector representations, where words with similar meanings are mapped to similar vectors in a high-dimensional space.

4. Modeling. The term "modeling" refers to the creation of computational

models that can understand natural language and classify news. Those used in the study will be further explored.

5. Evaluation. In NLP pipelines, evaluations are crucial for assessing the performance and effectiveness of the models and techniques used. Accuracy is used as the evaluation metric, as it is also the metric selected for an important challenge related to this type of task, the FNC (Fake News Challenge), conducted for the first time in 2016.

For the initial models built and studied, the tokenization, POS tagging, and lemmatization steps are performed sequentially. The embedding technique used is GloVe, introduced in 2014 by Pennington et al [8]. In the process of generating embeddings, GloVe involves analyzing the co-occurrence statistics of words within a corpus. The basic idea is that words that frequently appear together and have similar meanings should have similar vector representations. This way, even rare words are useful: GloVe captures their semantic relationships. The GloVe model used is trained on a Twitter corpus with 27 billion tokens and words are represented in 100 dimensions.
GloVe embedding is then incorporated as a layer in the neural networks of this study:

- Base neural network with convolutional layers - Convolutions, commonly used in convolutional neural networks (CNNs), were initially developed for image analysis but have been successfully adapted for natural language processing. In the context of NLP, text convolutions are used to extract meaningful features from textual data. The first paper associating them with NLP dates back to 2014 by Kim et al., published in the CoRR Journal [9], and they are increasingly employed in such tasks today. The model architecture comprises the following layers: Input — GloVe Embedding — Conv1D(64) — GlobalMaxPooling1D — Dropout — Conv1D(64) — Output.

- LSTM - LSTM is a variant of RNN designed to address the vanishing gradient problem, which can occur when training traditional recurrent neural networks on long temporal sequences. LSTM introduces a more complex architecture compared to standard RNNs, with a cell structure that can store, read, and write information more effectively. This allows LSTMs to capture long-term relationships in sequences. LSTMs were introduced in 1997 by Hochreiter et al. [10], but are still used in NLP tasks today, also due to a paper published in ”Interspeech” by Mikolov et al. in 2010 [11] demonstrating their effectiveness as LMs. The model architecture comprises the following layers: Input — GloVe Embedding — LSTM — Dropout — Output.

- Bi-LSTM - Presented by Graves et al. in 2013 [12], they are a variation of classic LSTMs. In a bidirectional LSTM, the model has two sets of LSTM cells: one processing the sequence in chronological order (from first to last position) and another processing the sequence in reverse order (from last to first position). This bidirectional approach allows the model to capture information from both directions, enabling a better understanding of the overall context of a sequence. The model architecture comprises the following layers: Input — GloVe Embedding — Bi-LSTM — Dropout — Output.

Another model used in this study is BERT. In 2018, Devlin et al. from Google published a paper and subsequently presented BERT at the annual meeting of the North American Chapter of the Association for Computational Linguistics [13]. It is a transformer-based language model, a class of models that has demonstrated success in NLP and beyond. The main peculiarity of BERT is its ability to understand the context of words in a sentence using a bidirectional approach. Unlike previous models, which treated words sequentially, BERT analyzes the context both left and right, allowing it to capture more complex and implicit relationships between words.

BERT is pretrained on a large corpus of text and is employed in the task at hand through fine-tuning. The BERT model is pretrained by Tensorflow on case-insensitive text, has 4 hidden layers of 512 nodes each and 8 attention heads. The model architecture comprises the following layers: Input — BERT Preprocessing — BERT Encoder — Dropout — Output.

BERT is state-of-the-art in many tasks as demonstrated by a 2019 study by Rogers et al. [14], including text classification. For BERT, tokenization is also employed, but using a tokenizer created specifically for the model: Wordpiece. Presented in 2012 at the ”2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)” by Google and studied by Schuster et al. [15], it is a tokenization method that divides words into smaller sub-units called ”word pieces”. The goal is to efficiently handle linguistic diversity and complexity. These ”word pieces” are determined based on their frequency in the training corpus and can represent more complex words through combinations. However, unlike the training of other models,

lemmatization is not applied as a preprocessing step. It does not always add value when using neural embeddings that strongly capture semantic context. Nevertheless, to evaluate the effectiveness of lemmatization with BERT, research has been conducted. A 2019 article by Kutuzov et al. [16] challenges the assumptions of most research articles on deep learning approaches to natural language processing (NLP), where it is presumed that lemmatization is not necessary, especially when powerful contextualized embeddings are used. Experiments conducted show that this is indeed true for languages with simple morphology (such as English). However, for languages with rich morphology (such as Russian), using lemmatized training data produces modest but consistent improvements in word sense disambiguation tasks. Since English datasets are available, lemmatization is not performed in training BERT.
Once the analysis on the four models is completed, to enhance their robustness and generalization capability, two techniques are tested on the WELFake trainset:

- Data augmentation. Data augmentation is a technique aimed at increasing the quantity of available training data by intelligently modifying existing data. There are various text augmentation techniques, the main ones being mentioned in a 2021 Survey Paper by Shorten et al. published in the "Journal of Big Data" [17]. One of these techniques is Back-Translation, which involves translating a sentence or text from one language to another and then translating it back to the original language. This way, new data instances can be generated for model training. Back-Translation exploits the semantic invariances encoded in supervised translation datasets to produce semantic invariances for augmentation purposes. This technique has been found effective even in the presence of neural embedding techniques, as reported in Corbeil et al.'s 2020 paper [18]. In this study, an English-French-English Back-Translation is used.

- Datasets combination. Combining datasets is useful when multiple datasets with similar characteristics or objectives are available. By merging these sets, it is possible to increase the amount of data available for training, which can lead to better model performance. In this analysis, a horizontal combination occurred by adding instances of ISOT to the WELFake dataset. There are no significant studies regarding this technique; however, it is employed in an "innovative" manner in this paper in hopes of improving the models in the task of

Fake News Detection, considering the presence of two political news datasets described by the same features.

Up to now, the study has been solely focused on attempting to achieve state-of-the-art performance for the two benchmark datasets. However, the models require many parameters, and some devices do not have enough memory to contain them. Additionally, a lot of computation is required for prediction, leading to high delays before response, as well as high battery consumption, making it less environmentally friendly.

In an attempt to achieve the same performance as the more complex models, the "knowledge distillation" technique is employed. It is useful for transferring knowledge from a large and complex model (teacher model) to a smaller and simpler model (student model). In particular, a neural network specifically designed, the student model, is created to have the minimum number of parameters, trying not to excessively impact accuracy. Meanwhile, the previously trained BERT Fine-Tuned model is used as the teacher model.

The goal is to enable the student model to mimic the behavior of the teacher model while being computationally more efficient and potentially smaller in size. The effectiveness of this technique is evidenced in numerous articles, even for NLP tasks, especially those utilizing deep neural models, as reported and studied in Sun et al.'s 2019 paper published in the book "IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)" [19].

4 Results and Evaluation

The results of the analysis are reported here.

Table 1: Model comparison.

		GloVe + CNN	**GloVe + LSTM**	**GloVe + BiLSTM**	**BERT fine tuned**
WELFake	Accurancy	0,91	0,75	0,94	0,97
	Precision	0,92	0,81	0,94	0,97
	Recall	0,91	0,75	0,94	0,9
	F1-score	0,91	0,75	0,94	0,97
	Training time	*24,5s*	*18m 32s*	*35m 44s*	*36m 28s*
ISOT	Accurancy	1,00	0,95	1,00	1,00
	Precision	1,00	0,95	1,00	1,00
	Recall	1,00	0,95	1,00	1,00
	F1-score	1,00	0,95	1,00	1,00
	Training time	*13,7s*	*12m 27s*	*23m 26s*	*21m 50s*
	# Parameters	1.036.489	1.017.257	1.482.201	28.764.162

Table 2: Compression of models.

		(old) BERT fine tuned	**Student network**	**Student network distilled**
WELFake	Accurancy	0,97	0,90	0,90
	Precision	0,97	0,91	0,91
	Recall	0,97	0,90	0,90
	F1-score	0,97	0,90	0,90
	Training time	*36m 28s*	*41,9s*	*18m 10s*
ISOT	Accurancy	1,00	1,00	1,00
	Precision	1,00	1,00	1,00
	Recall	1,00	1,00	1,00
	F1-score	1,00	1,00	1,00
	Training time	*21m 50s*	*21,4s*	*11m 24s*
	# Parameters	28.764.162	7.361	7.361

Table 3: Attempt to increase generalization capacity.

		CNN	**LSTM**	**Bi-LSTM**	**BERT**
WELFake + Back translation	Accurancy	0,83	0,73	0,86	0,92
	Precision	0,83	0,74	0,86	0,92
	Recall	0,83	0,73	0,86	0,92
	F1-score	0,83	0,73	0,86	0,92
	Training time	*3m 59s*	*27m 25s*	*53m 53s*	*52m 3s*
WELFake + Dataset combination	Accurancy	0,92	0,81	0,94	0,97
	Precision	0,92	0,82	0,94	0,97
	Recall	0,92	0,80	0,94	0,97
	F1-score	0,92	0,80	0,94	0,97
	Training time	*42s*	*29m 26s*	*54m*	*53m 3s*
	# Parameters	1.036.489	1.017.257	1.482.201	28.764.162

5 Discussion

A first aspect to note is related to the accuracy comparison between the network with Convolutional layers and LSTM 1. The fact that the former outperforms the latter is not, contrary to what one might think, an anomalous result; rather, it is an outcome found in numerous papers. A 2023 study by Ajik et al. published in the ”Journal of Information Systems and Informatics” [20] and related to ”Fake News Detection,” highlights how a CNN network achieves higher accuracy compared to LSTM, even in NLP-related tasks, and that the combination of the two can be the ”winning card.” Ajik et al.’s study is not the only article emphasizing this particularity: certainly, performance depends on the dataset, but it is worth noting that while CNNs achieve excellent results when substantial features need to be extracted, LSTMs, or RNNs in general, are suitable when considering data sequentiality and relationships. It is not the objective of this study, but it would be interesting to evaluate the advantages and limitations of networks with convolutional layers in working with natural language.

An online search reveals that the state-of-the-art of WELFake and ISOT are, respectively, 96.75% [3] accuracy and 100% accuracy [20]. In this study, as shown in table 1, both have been matched thanks, in particular, to a BERT model fine-tuned for a classification task.

Furthermore, it can be noted that false positives outnumber false negatives in all models: this data emphasizes the quality of information, not classifying false news as true, but at most vice versa, which is acceptable. In an attempt to increase the models' generalization capacity, two data augmentation techniques and dataset combinations were employed. Unfortunately, the techniques used proved unhelpful in achieving the goal, but nevertheless, the models proved effective in predictions even in the presence of synthetic data, as can be seen in table 3. Are there more effective data augmentation techniques for this task? It would be interesting to investigate in the future.

The common problem among all the mentioned models is the high number of parameters involved: at least 1 million. Considering the limited resources of devices and a more environmentally friendly informatics, a very promising model was built with only 7361 parameters. Compared to BERT, the new architecture, saving 99.96% of the parameters 2, achieves 100% accuracy for ISOT and 90% for WELFake (a loss of 7.2%). In an attempt to further attenuate this difference between BERT and the new model in terms of accuracy, Knowledge Distillation was employed using BERT as a teacher, which, however, proved superfluous in this task.

A last aspect to discuss is related to a further possible future investigation: from the exploratory analysis, it emerged that fake news tends to be generally longer and more detailed than real news. There are no, as previously highlighted, studies on this, but it could be a useful discovery in the current and ongoing attempt to find new discriminants to effectively counter fake news.

6 Conclusions

In this study, aimed at combating the increasingly prevalent "fake news" in today's society, ML models were built on two benchmark datasets: ISOT and

WELFake. After numerous experiments, the model that best classified the news is based on BERT, and it achieved an accuracy of 100% for ISOT and 97% for WELFake, matching the state-of-the-art. Even if models are already performing well, using Back Translation augmentation technique or dataset combinations, in this task are unhelpful, but other techiques can potentially lead to further improvements in its performance and generalization ability, making it more reliable and effective in real-world scenarios; it would be interesting to evaluate others in future studies. Even though models currently exhibit good performance, the utilization of Back Translation augmentation techniques or dataset combinations in this task proves unhelpful. However, other techniques may hold potential for further enhancing performance and generalization ability, rendering the system more reliable and effective in real-world scenarios. Evaluating alternative methods in future studies would be of considerable interest.

Although BERT fine-tuned proved to be the best classifier, with a 99.96% parameter savings, a model was built that achieves 100% accuracy for ISOT and 90% for WELFake, suitable for devices with limited resources and supporting a greener informatics.

The study found that fake news tends to be longer and more detailed than real news: it would be interesting to further explore this characteristic in an attempt to find new discriminants to counter fake news.

References

[1] Oxford English Dictionary. (2023, September) fake news (n.). [Online]. Available: https://doi.org/10.1093/OED/3351660493

[2] Eurostat, "How many people verified online information in 2021?" December 2021. [Online]. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20211216-3

[3] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfake: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021.

[4] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Intelligent, Se-*

cure, and Dependable Systems in Distributed and Cloud Environments, I. Traore, I. Woungang, and A. Awad, Eds. Cham: Springer International Publishing, 2017, pp. 127–138.

[5] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, "Understanding the behaviors of bert in ranking," 2019.

[6] G. K. Zipf, *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, 1949.

[7] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[8] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[9] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[11] T. Mikolov, M. Karafiát, L. Burget, J. H. ernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:17048224

[12] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," *CoRR*, vol. abs/1303.5778, 2013. [Online]. Available: http://arxiv.org/abs/1303.5778

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*,

2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[14] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how BERT works," *CoRR*, vol. abs/2002.12327, 2020. [Online]. Available: https://arxiv.org/abs/2002.12327

[15] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5149–5152.

[16] A. Kutuzov and E. Kuzmenko, "To lemmatize or not to lemmatize: How word normalisation affects elmo performance in word sense disambiguation," *ArXiv*, vol. abs/1909.03135, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202540040

[17] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of Big Data*, vol. 8, no. 1, p. 101, 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00492-0

[18] J.-P. Corbeil and H. A. Ghadivel, "Bet: A backtranslation approach for easy data augmentation in transformer-based paraphrase identification context," 2020.

[19] H. Sun, X. Tan, J.-W. Gan, S. Zhao, D. Han, H. Liu, T. Qin, and T.-Y. Liu, "Knowledge distillation from bert in pre-training and fine-tuning for polyphone disambiguation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 168–175.

[20] E. Ajik, G. Obunadike, and F. Echobu, "Fake news detection using optimized cnn and lstm techniques," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 1044–1057, Aug. 2023. [Online]. Available: https://www.journal-isi.org/index.php/isi/article/view/548