Misinformation detection in news text: Automatic methods and data limitations

Fatemeh Torabi Asr

(wrong information) and `disinformation' (wrong information with the intention to deceive) more accurate terms for scienti c research as compared to `fake news', which is frequently used in political discourse and media stories (for de nitions, see Habgood-Coote, 2019; Tandoc Jr et al., 2018; Wardle and Derakhshan, 2017). The subject of our study is false information in news text, regardless of the distributing source's intention, thus misinformation in its general sense, which can manifest itself as rumours and hoaxes, propaganda, or even false information in mainstream news publications.

The research community turned its attention to the phenomenon in 2015 and 2016 (Connolly et al., 2016; Perrott, 2016), with two comprehensive studies published in 2018 (Lazer et al., 2018; Vosoughi et al., 2018). The latter in particular clearly established the danger of misinformation: Fake news stories are particularly dangerous because they not only tend to reach a larger audience, but also penetrate into social networks nearly 10 times faster than fact-based news (Vosoughi et al., 2018).

The current trends to combat the misinformation problem take three main approaches: educate the public, carry out manual checking, or perform automatic classi cation. Educating the public involves encouraging readers to check the source of the story, its distribution (who has shared it, how many times), or to run it by fact-checking websites. This is certainly necessary, but it will not be enough. Organized manual checking, before or after publication, is a possibility, but it is also not a realistic solution, given the fast spread of misinformation that Vosoughi et al. (2018) found. Approaches from machine learning, computational linguistics, and natural language processing (NLP) show promise, in that they can perform automatic classi cation and can help complement the e orts of fact-checking sites. The promise is that we will be able to detect fake news stories automatically, before they have a chance to spread and do harm. The process of fact-checking can be modeled as a series of NLP tasks, from identifying claims and rumours to comparing information and producing fact-checking verdicts and justi cations (Guo et al., 2022). In this paper, we explore the deployment of a speci c NLP task, text classi cation.

One of the important challenges in automatic misinformation detection using modern NLP techniques is data (Asr and Taboada, 2018, 2019). Annotation of fake news is a resource-demanding and particularly sensitive task because of the wide spectrum of public opinions about who exactly is a reliable source, including established news organizations. The majority of automatic systems built to identify fake news rely on training data (news articles) labelled with respect to the credibility or the general reputations of the sources, i.e., domains/user accounts (Fogg et al., 2001; Horne et al., 2018; N rregaard et al., 2019; Rashkin et al., 2017; Volkova et al., 2017; Yang et al., 2017). Even though some of these studies try to identify fake news based on linguistic cues, what they eventually model is the publisher's general writing style (e.g., common writing features of the publishing websites) rather than the linguistic similarities of the articles containing false information.

For example, Rashkin et al. (2017) collected news articles from websites that they categorized as general publishers of Hoax, Propaganda, Satire or Trusted (mainstream) news. They showed that a classi er trained on news articles from some of these websites could identify news from other websites from the same category, thus learning the general linguistic characteristics of each type of publisher. Detecting the style of a news article in terms of belonging to coarse categories such as Satire, Propaganda, Hoax or Trusted mainstream outlets is an interesting task, but not exactly what we would like to do in our battle against fake news. The goal of our paper is to pursue a slightly di erent and hypothetically more di cult task, namely detecting, based on linguistic properties, whether or not a news article contains false information. This is useful, because the approach could then work across di erent sites, regardless of publisher.

In terms of methodology, we focus on a content-based approach to news text classi cation. Rather than using contextual metadata such as user activity features, network cues, or credibility of the publishing sources, we assess the feasibility of detecting misinformation by examining the

content of the article, i.e., the text itself. This puts our work in the category of style-based fake news detection, as opposed to context-based or knowledge-based detection (Potthast et al., 2018) and in the area of language-based detection (Lugea, 2021). The hypothesis behind our approach is that deception in news has its own style, i.e., a language for misinformation. If the language of news articles with true vs. false content is different, then we should be able to detect misinformation

surface characteristics such as document length and n-gram frequency to speci c types of semantic classes (e.g., subjectivity and emotion markers), syntactic features (e.g., depth of syntactic tree and frequency of each part of speech) and discourse-level features (Afroz et al., 2012; Conroy et al., 2015; Horne and Adali, 2017; Pérez-Rosas and Mihalcea, 2015; Rashkin et al., 2017; Rubin et al., 2015; Ruchansky et al., 2017; Volkova et al., 2017). Some of these studies have been characterized as stylometric (Przybyla, 2020), in that they use the style of the language as an indicator of misinformation.

Algorithms deployed in this type of supervised learning are often Support Vector Machines (SVMs), with a feature engineering and feature selection process. Performance in these approaches tends to plateau as data increases, showing that features are useful with smaller amounts of data, but performance increases stall at some point as amount of available data increases. Therefore, these methods are considered to have an important limitation (Ng, 2011).

A second set of studies use modern neural network models. In cases where large amounts of data are available, deep neural network models tend to achieve more impressive results. Deep learning has, in general, taken over many natural language processing tasks, at least in domains where large-scale training data is available. Deep learning models in NLP usually rely on word vectors and embedded representations. Although it is possible to extract embeddings from domain data, most methods rely on pre-trained embeddings (Le and Mikolov, 2014; Pennington et al., 2014). Models in deep learning include Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Attention models (Conneau et al., 2017; Lai et al., 2015; Le and Mikolov, 2014; Medvedeva et al., 2017; Yang et al., 2016; Zhang et al., 2015). They perform slightly di erently, depending on the task and the type of data. In general, the task tends to be a binary classi cation task (i.e., is this text X or Y?), in our case whether the text in question is an instance of fake news/misinformation or an instance of reliable, fact-based news. For this task, what RNNs do is encode sequential information in the articles, modeling short text semantics. CNNs are composed of convolution and pooling layers, providing an abstraction of the input. CNNs are useful in tasks where presence or absence of features is a more distinguishing factor than their location or order, and work well when classifying longer text. For instance, CNNs are helpful in sentiment analysis of product reviews, where the distinguishing features between positive and negative reviews may be the relative frequency or presence of positive and negative words (Dos Santos and Gatti, 2014; Kucharski, 2016; Ouyang et al., 2015).

The issue with many of these studies is the data collection methodology. Many of t4(Ma-c6dH.)]TJgThegr3

annotations for stance (Thorne et al., 2018), or articles that were modified to be made untrue (Pérez-Rosas et al., 2017). The dataset from Shu et al. (2020) comes closest to the requirements for this task, as it was crawled from fact-checking websites (but not validated after scraping).

In summary, although a few different types of datasets exist, none of them contain a large enough number of both fake and legitimate news articles, which is the type of data that we need to learn

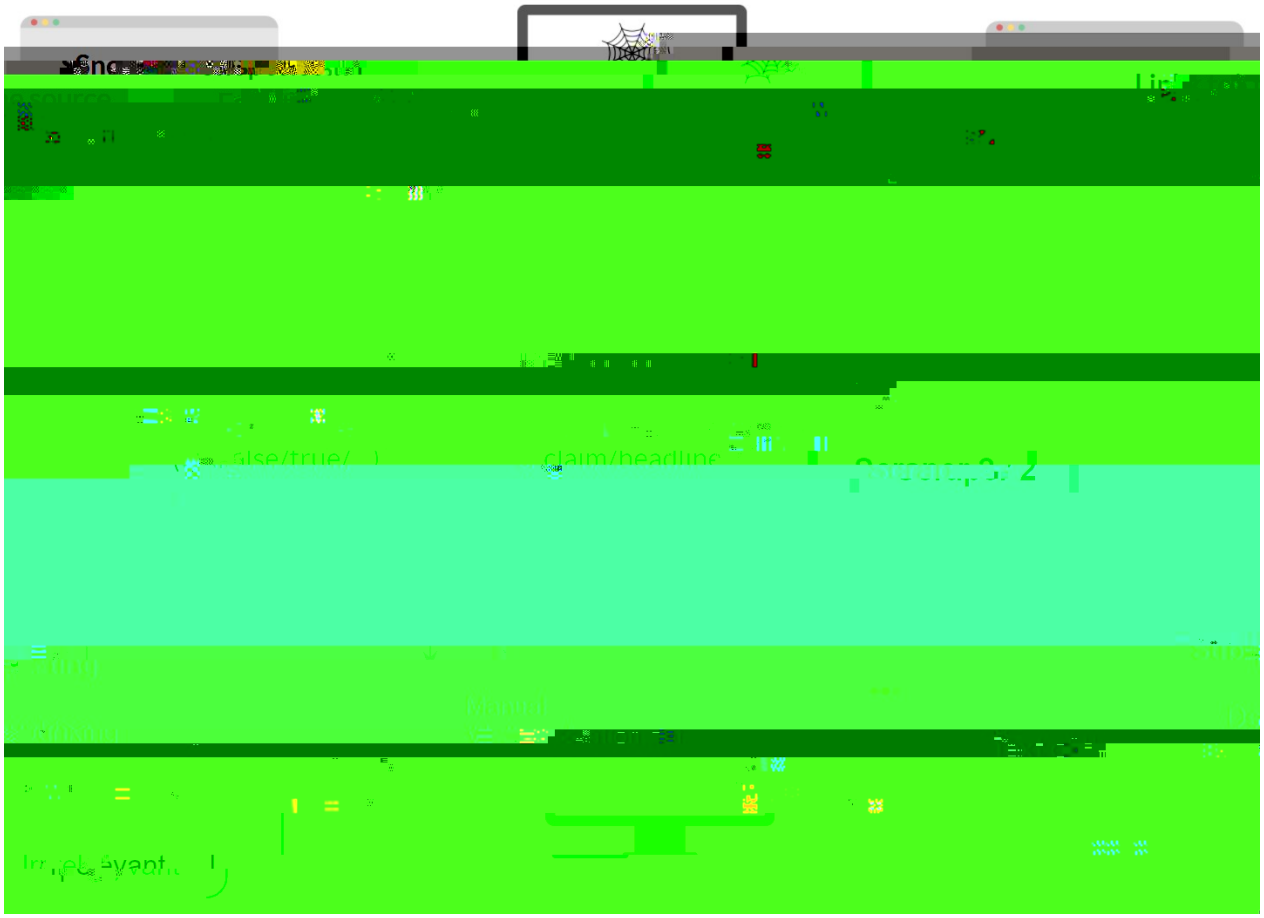

Figure 1: Data scraping and validation process

of a Snopes article is the investigation of a claim, rumour, or story that has been contested. Snopes then discusses the content of that contested story, including a number of links, some to the source of the false story, some to sites that debunk the story or other links for extra information on the topic. Our task, then, was to follow those links and, for instance, in the case of a \false" label on the Snopes site, to find the link containing the full article with the false story. The entire process involved, for each claim (\true" or \false"), scraping the discussed claim, the veracity scoring of the claim according to the Snopes labelling system, and the links to the source of the claim. Given that the scraped data might be noisy (e.g., other links on the webpage might be harvested but do

can provide useful information about a text. Studies on deceptive text and fake news have used such features to try and nd distributional di erences between lies and truthful statements (Biyani et al., 2016; Perez-Rosas et al., 2017; Perez-Rosas and Mihalcea, 2015; Rashkin et al., 2017; Rubin et al., 2016). We consider a large set of semantic lexicons, where words

Table 1: Important features for distinguishing true from false news articles; only overlaps between Buzzfeed and Snopes Silver datasets are included here. Shaded rows are features of false news articles.

Feature category	Feature	Ratio false/true	P-value
Surface	num_punc/num_char	0.926844562	0.001331224
Semantic lexicon	comparative_forms	0.947737051	0.020227509
	negative_HuLui	1.133089752	0.00267177
	negative_mpqa	1.192098625	0.000864263
	modal_adverbs	1.298092007	0.005319897
	manner_adverbs	1.35698137	0.001742446

This can be attributed to the use of direct quotations or self-mention of the reporter in true news, as opposed to repeated mention of other entities in false news, perhaps a sign of othering certain groups of people (Riggins, 1997).

Extracting the most important n-grams is a challenging task due to the large number of such features. Moreover, by examining the individual n-grams one can hardly infer a general pattern regarding the differences they reveal between false and true news articles. In order to examine the n-grams most specific to each dataset, we first combined the Snopes Silver and the Buzzfeed USE corpora and then extracted 100 n-grams that occurred in less than half of all the documents (to avoid corpus-specific stop words) but in more than three documents within the combined corpus. We then analyzed these unigrams the same way as other feature types: we calculated the false to true proportion of each unigram and filtered those with highest and lowest values and $ap - value < 1$ according to the Recursive Feature Elimination method. The list of most discriminating unigrams based on this technique is provided in Table 2.

The majority of high-score unigrams marking the true news articles are focused around the topic of the US election. This is expected, given the fact that most true examples in the combined dataset come from the Buzzfeed USE corpus, which includes news related to the US presidential candidates and events around the 2016 election. The Snopes data includes a variety of topics and most false articles come from this dataset; that is why the high-score unigrams in the false class come from a more diverse and general vocabulary, as it is evident from the table. Now, this imbalance in terms of topic vocabulary between the two classes of news articles may raise a challenge for building predictive models based on the presented data: If we train a model on this dataset, the classifier may overfit to fine-grained lexical features rather than high-level properties of the text and this may result in weak generalization and low accuracy on collections of false/true news articles with a different topic distribution. We will discuss this further in the next section.

Table 2: TF-IDF unigram features with highest proportion in true (top) vs. false news (bottom) within the combined corpus (Snopes Silver and Buzzfeed USE)

Feature	Avg. in true	Avg. in false	Ratio false/true	P-value
debate	0.066639	0.011568	0.173589	1.18E-50
voters	0.043444	0.009592	0.220792	2.48E-31
clinton	0.114134	0.031633	0.277158	2.71E-72
presidential	0.065025	0.019048	0.29294	8.90E-56
campaign	0.075448	0.022253	0.294945	1.77E-52
republican	0.062131	0.019052	0.306639	1.74E-46
hillary	0.074586	0.023268	0.311959	5.28E-53
donald	0.087009	0.032008	0.367873	1.74E-62
trump	0.150655	0.063265	0.419935	4.73E-67
vote	0.036512	0.018449	0.505276	1.03E-08
election	0.041777	0.021478	0.514117	7.42E-12
today	0.025932	0.036286	1.399304	4.85E-04
year	0.044547	0.06537	1.467454	8.41E-10
come	0.02743	0.040592	1.479861	3.47E-06
family	0.029711	0.044435	1.495613	2.58E-05
home	0.026483	0.041189	1.555327	9.89E-06
school	0.021147	0.034088	1.611911	2.88E-04
world	0.03383	0.056416	1.667639	5.27E-10
use	0.027951	0.049615	1.775065	1.07E-09
children	0.023254	0.043545	1.872602	3.24E-08
old	0.024174	0.047418	1.96152	3.68E-13

5.1. Models

Feature-based model . We use a Support Vector Machine Classi er (SVM) from the scikit-learn python library with all the linguistic features that we introduced and analyzed in the previous section. These features include surface text features, TF-IDF scored n-grams, semantic category features, syntactic features (parts of speech counts), and readability scores. Both the TF-IDF vectorizer and the SVM classi er had a set of parameters that we tuned through cross validation on training data. The best values for parameters of the TF-IDF vectorizer were max-df=0.5, min-df=5, n-gram-range=(1,2) and sublinear-tf=True. Best parameter values for the SVM were penalty=\l2", tol=1e-3 and others set to default.

with 1 million steps for 40 epochs, with batch size of 256 on 8 GPUs for 6.5 days. Due to the large number of parameters and layers in a BERT network and the high risk of over tting, we ne-tune this pre-trained BERT model. To ne-tune, we train the entire pre-trained model on our training data and feed the output to a softmax classi er to compute logits. We optimize the main hyperparameter values, including the dropout rate, batch size, optimizer learning rate, and the number of epochs based on the average Area Under Curve (AUC) score from cross-validation on training data. The best values for the parameters of the BERT-based model were dropout=0.35, batch_size=12, learning_rate=1e-5 (Adam optimizer). We utilized the maximum sequence length of 512 and tried between 4 and 15 training epochs. With the smaller training data, higher epoch numbers (around 12) gave us better results (lower validation and training loss). However, with larger training data size the performance plateaued after about 5 epochs, so we keep that as the accepted parameter. It is also worth mentioning that smaller training data are more prone to su er from over tting, resulting in larger validation losses with the same hyperparameter values. Table C.7 includes our best range of hyperparameters for the BERT-base model.

5.2. Data preparation

While building a predictive model for any detection task, it is important to spend time preparing data by removing noise and balancing the samples for the prediction classes. Our observation during feature analysis of the datasets showed a topic imbalance across datasets and, most importantly, across target classes (false and true news articles), as we have shown in other work (Asr and Taboada, 2019). Therefore, we decided to consider two di erent data scenarios in our text classi cation experiments. Table 3 shows how we sample the Buzzfeed USE and Snopes Silver datasets for preparation of training data with two approaches. We consider two training data scenarios, one with a small balanced training data and another with a relatively large but mixed training data. For sampling the small and balanced dataset, we randomly picked 64 items from each class within the Buzzfeed USE dataset, because its false class only contains this many items. This was in principle to make sure that data from a focused topic (the US election) is represented in both false and true classes in the balanced dataset. We take a similar approach in sampling from the Snopes Silver data, by picking 259 items from each class. The total number of items in our small and balanced dataset is 646 news articles. Second, we consider a larger training dataset, that is, we put together all true and false news articles from the Snopes and Buzzfeed datasets and then sample 1,300 items per class from this collection, which totals 2,600 news articles. This dataset is about four times larger than our small sample, but it is unbalanced with respect to the distribution of topics across false and true news articles.

As test data, instead of sampling from the same data sources, which may result in an arti cially high accuracy, we consider three separate test datasets. These are all datasets that have been manually checked for the content veracity of individual news articles. The rst obvious choice is the Snopes Gold data that we described in the section on data collection (Section 3). Apart from having been veri ed manually, this dataset has the nice feature of including non-overlapping news headlines with the Snopes Silver data. Performance on this dataset would tell us about the generalization power of a model to new topics and headlines.

Table 3: Number of samples taken from each collection (Snopes Silver and BuzzFeed USE) to prepare our two di erent training datasets: the small balanced & large mixed samples

Dataset	False items	True items	Small balanced sample	Large mixed sample
Snopes Silver	1,585	259	2 259	2 1,300
Buzzfeed USE	64	1,090	2 64	
Total	1,649	1,349	646	2,600

in our training data, so it may re ect the generalization power of the models better; furthermore, this dataset contains a balanced number of false and true news articles on pre-selected matched topics (for instance, the same number of false and true articles about Jenifer Aniston's personal life!). Performance of our models on this last test dataset would be representative of cross-domain classi cation performance on real data.

number of items from both classes in the training data, one possible explanation for the classi er's bias could be that the items of the false class were a better representative of the language data that we see in the test sets; in other words, these items could have covered a larger number of topics, more varied vocabulary and writing styles. This distributional characteristic can be due to the more diverse sources of online news scraped for the fake items than for real items (mainstream trusted news) in Rashkin's data.

Overall, the low F-scores obtained on the Snopes Gold and Perez Celebrity data provides some evidence that reputation-based data collection may not be the best strategy when the target task is to detect false from true content. While the classi er seems to be good at detecting fake news (as

6. Conclusion

We have investigated the problem of misinformation in news text from a linguistic perspective, using Natural Language Processing and text classi cation techniques. The contributions can be summarized as the following:

- We built a dataset of false and true news articles by scraping the Snopes fact-checking pages, tracking the links to the original publisher of the news headlines and collected the body text. We also used crowdsourcing to verify the alignment between each news article and the headline labelled for veracity by the fact-checker, to make sure the data is of good quality. The Snopes Silver collection contains 1,844 texts; it has been introduced in our previous work and a small sample of it, i.e, the Snopes Gold, was used in our previous experiments (Asr and Taboada, 2019). The complete collection with crowdsourced stance data will become available upon the publication of the current manuscript.

- We analyzed the above dataset and the Buzzfeed USE dataset (from our previous work) for linguistic features indicative of false content and provided signi cant tests on what types of features were most discriminatory between false and true news articles.

- We conducted experiments on automatic misinformation detection using a variety of text classi cation techniques. By doing so, we established a new baseline for this NLP task and clari ed the type of data and features that can o er the best accuracy both in within-domain and cross-domain predictions.

Our experiments show that the veracity and linguistic characteristics of a text are correlated, but high-quality training data is required to develop an accurate and scalable misinformation detection system. In particular, data should be well-distributed across topics and sources, balanced across di erent levels of factuality, and reliably labelled based on individual articles rather than the reputation of publishing sources, because dubious websites may publish or republish factual news articles, making the data noisy.

In terms of the machine learning techniques, we found that the classic feature-based SVM model was superior across all data scenarios. Especially in a small but balanced training data scenario, the models showed a more robust behavior, i.e., they generalized better on the test news articles with unseen headlines and claims.because dubd36 -W8t (b)363(w)895eabandea-382(b) m(w)895eblues. Bstudnop

Our new dataset and the properties we found for a quality dataset based on repeated experiments contribute to opening up the bottleneck in the NLP approach to misinformation detection, but more

Table A.5: Discriminative features in Buzzfeed data

Feature Category	Feature	Ratio false/true	P-value
Surface	num_punc/num _char	0.824111622	1.07E-07
Semantic lexicon	comparative_forms.txt	0.847129838	0.000312741
Semantic lexicon	negative_mpqa.txt	1.170890719	0.039175394
Semantic lexicon	negative-HuLui.txt	1.191929819	0.003812285
Semantic lexicon	assertiveshooper1975.txt	1.204042781	0.025607042
Semantic lexicon	factives_hooper1975.txt	1.210492088	0.075483432
Semantic lexicon	modal_adverbs.txt	1.250324586	0.08518351
Semantic lexicon	neutral_mpqa.txt	1.262506856	0.008025674
Semantic lexicon	manner_adverbs.txt	1.285541793	0.070899584

Appendix C. Fine-tuned BERT model optimized hyperparameters

Table C.7: Fine-tuned BERT model hyperparameters

Hyperparameters	Tested range	Best range
Sequence Length	256 - 512	360 - 512
Number of epochs	3 - 15	4 - 12
Batch size	4 - 16	8 - 12
Dropout rate	0 - 0.5	0.25 - 0.35
Learning rate (Adam optimizer)	1E-6 - 1E-4	1E-5 - 5E-5
Warm-up steps	0 - 500	0 - 500

References

Afroz, S., Brennan, M., Greenstadt, R., 2012. Detecting hoaxes, frauds, and deception in writing style online. In: Proceedings of IEEE Symposium on Security and Privacy. San Francisco, pp. 461{475.

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. Journal of Economic Perspectives 31, 211{236.

Asr, F. T., Taboada, M., 2018. The data challenge in misinformation detection: source reputation vs. content veracity. In: Proceedings of the First Workshop on Fact Extraction and VERi cation (FEVER),Conference on Empirical Methods in Natural Language Processing. Brussels, pp. 10{15.

Asr, F. T., Taboada, M., 2019. Big data and quality data for fake news and misinformation detection. Big Data & Society, January{June 2019: 1{14.

Biyani, P., Tsioutsiouliklis, K., Blackmer, J., 2016. \8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In: Proceedings of the 30th AAAI Conference on Arti cial Intelligence. pp. 94{100.

Church, K. W., Chen, Z., Ma, Y., 2021. Emerging trends: A gentle introduction to ne-tuning. Natural Language Engineering 27 (6), 763{778.

Conneau, A., Schwenk, H., Barrault, L., LeCun, Y., 2017. Very deep convolutional networks for text classi cation. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, pp. 1107{1116.

Connolly, K., Chrisa s, A., McPherson, P., Kirchgaessner, S., Haas, B., Phillips, D., Hunt,

Mukherjee, A., Venkataraman, V., Liu, B., Glance, N., 2013. Fake review detection: Classi cation and analysis of real and pseudo reviews. Technical Report UIC-CS-2013{03, University of Illinois at Chicago, Tech. Rep.

Ng, A., 2011. Why is Deep Learning taking o ? Tech. rep., Coursera.
URL https://www.coursera.org/lecture/neural-networks-deep-learning/why-is-deep-learning-taking-off-praGm

N rregaard, J., Horne, B. D., Adal , S., 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In: Proceedings of AAAI Conference on Web and Social Media. Munich, pp. 630{638.

Ouyang, X., Zhou, P., Li, C. H., Liu, L., 2015. Sentiment analysis using convolutional neural network. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. IEEE, pp. 2359{2364.

Pennebaker, J. W., Boyd, R. L., Jordan, K., Blackburn, K., 2015. The development and psychometric properties of LIWC 2015. Technical report, University of Texas at Austin.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, pp. 1532{1543.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R., 2017. Automatic detection of fake news. arXiv preprint arXiv:1708.07104.

Pérez-Rosas, V., Mihalcea, R., 2015. Experiments in open domain deception detection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, pp. 1120{1125.

Perrott, K., 2016. A fake news on social media in uenced US election voters, experts say. ABC 26.
URL http://www.abc.net.au/news/2016-11-14/fake-news-would-have-influenced-us-election-experts-say/8024660

Post, M., Bergsma, S., 2013. Explicit and implicit syntactic features for text classi cation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Vol. 2. So a, pp. 866{872.

Potthast, M., Kiesel, J., Reinartz, K., Bevendor , J., Stein, B., 2018. A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, pp. 231{240.

Przybyla, P., 2020. Capturing the style of fake news. In: Proceedings of AAAI Conference on Arti cial Intelligence. Vol. 34. New York, pp. 490{497.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., Choi, Y., 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, pp. 2921{2927.

Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D., 2013. Linguistic models for analyzing and detecting biased language. In: Proceedings of the Conference of the Association for Computational Linguistics. So a, pp. 1650{1659.

Wardle, C., Derakhshan, H., 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Report, Council of Europe.

Wilson, T., Wiebe, J., Ho mann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, pp. 347{354.