FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German

Mattern, J.; Qiao, Y.; Kerz, E.; Wiechmann, D.; Strohmaier, M.

DOI
10.18653/v1/2021.fever-1.9

Publication date
2021

Document Version
Final published version

Published in
FEVER : Fact Extraction and VERification

FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German

Justus Mattern[1], **Yu Qiao**[1], **Elma Kerz**[1], **Daniel Wiechmann**[2], **Markus Strohmaier**[1]
[1]RWTH Aachen University
[2]University of Amsterdam
{justus.mattern,yu.qiao}@rwth-aachen.de
elma.kerz@ifaar.rwth-aachen.de
D.Wiechmann@uva.nl
markus.strohmaier@cssh.rwth-aachen.de

Abstract

As the world continues to fight the COVID-19 pandemic, it is simultaneously fighting an 'infodemic' – a flood of disinformation and spread of conspiracy theories leading to health threats and the division of society. To combat this infodemic, there is an urgent need for benchmark datasets that can help researchers develop and evaluate models geared towards automatic detection of disinformation. While there are increasing efforts to create adequate, open-source benchmark datasets for English, comparable resources are virtually unavailable for German, leaving research for the German language lagging significantly behind. In this paper, we introduce the new benchmark dataset FANG-COVID consisting of 28,056 real and 13,186 fake German news articles related to the COVID-19 pandemic as well as data on their propagation on Twitter. Furthermore, we propose an explainable textual- and social context-based model for fake news detection, compare its performance to "black-box" models and perform feature ablation to assess the relative importance of human-interpretable features in distinguishing fake news from authentic news.

1 Introduction

The online availability and rapid dissemination of 'disinformation' and 'fake news' – cover terms for various types of false, inaccurate, or misleading information, typically related to emerging and time-sensitive events – have become a global challenge over the past 10 years (Tucker et al., 2018). Never has the scale of this challenge been clearer than in the corona crisis: as the world continues to fight the coronavirus disease 2019 (COVID-19) pandemic, it is simultaneously fighting an 'infodemic' – a flood of disinformation and spread of conspiracy theories leading to the division of society (Orso et al., 2020; Solomon et al., 2020).

In order to combat the spread and consumption of disinformation, it is of essential importance to develop diagnostic and predictive analysis models that can be used to understand how and why disinformation is created and spread as well as to uncover hidden and unexpected aspects of disinformation content. A key aspect that has a significant impact on the detection of disinformation is the existence of high-quality benchmark datasets. While there are increasing efforts to create adequate, open-source benchmark datasets for English, comparable resources are virtually unavailable for German (but see Vogel and Jiang (2019b)). Accordingly, research on disinformation detection for German is clearly lagging behind.

In order to fill the research gap, the main goal of this paper is to introduce FANG-COVID[1], an entire-article COVID-19 benchmark dataset for the German language, designed based on existing English language benchmark datasets and labeled using distant supervision.

Furthermore, we present the results of benchmark experiments on the dataset using explainable textual- and social context-based models for fake news detection and compare their performance to those of classifiers based on a fine-tuned BERT language model. Finally, we conduct feature ablation experiments to assess the relative importance of human-interpretable features in distinguishing fake news from authentic news.

This paper is structured as follows: In section 2, we provide a concise overview of existing datasets and approaches for fake news detection, focusing on work related to COVID-19 and the German language. In section 3, we introduce the FANG-COVID dataset. In section 4, we propose an explainable fake news detection system, compare its performance to a black-box model and introduce the ablation algorithm used to analyze the importance of certain features, before summarizing the results and our contribution in section 5.

[1]https://github.com/justusmattern/fang-covid

Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER) at EMNLP 2021, pages 78–91

2 Related Work

In recent years, with the growing recognition of the importance of understanding and detecting fake news, increasing efforts have been made to combat disinformation, resulting in the creation of large-scale annotated datasets for fake news detection for the English language (Ahmed et al., 2018; Shu et al., 2020b; Wang, 2017; Thorne et al., 2018). For the German language, first steps towards creating suitable datasets have been made (Vogel and Jiang, 2019a; Vogel et al., 2020). To date, however, existing datasets are too small to enable the development of state-of-the-art deep learning models for fake news detection. Current approaches to automatic detection of English fake news (Kula et al., 2020; Wang et al., 2018; Ruchansky et al., 2017; Qiao et al., 2020) have taken into account both textual news content and also social context based on social media data (for recent reviews of disinformation detection, see Oshikawa et al. (2020), Shu et al. (2020a), Zhou and Zafarani (2018) and Shu et al. (2017)). Language-based approaches have an advantage over knowledge-based or propagation-based approaches[2], because (1) they enable near real-time feedback (proactive rather than retroactive), i.e. they are not restricted to being applied only *a posteriori* (Potthast et al., 2017) and (2) they are scalable. In the remainder of this section, we provide a concise overview of the existing COVID-19 fake news datasets and approaches to detect disinformation regarding the virus.

2.1 COVID-19 Fake News Datasets in English

To the best of our knowledge, there exist currently four suitable and publicly available datasets for COVID-19 fake news detection in English (see Table 2 for an overview). These datasets differ with respect to the type of news source they consist of (entire articles vs. individual claims/tweets), data size, the imbalance of the class distribution (real/fake) and the type of information available for fake news detection (linguistic and/or social context).

Patwa et al. (2020) introduced the Covid19 FN dataset in connection with the CONSTRAINT shared task for Covid-19 fake news detection. This dataset consists of 5,600 correct tweets gathered from reliable sources such as government accounts, medical institutes and news channels as well as 5,100 manually fact-checked false statements from various sources such as news articles, press releases and social media posts.

FakeCovid (Shahi and Nandini, 2020) is a multilingual dataset that consists of 5,182 manually fact-checked news articles from 40 languages of which 2,116 are English and 47 are German.

ReCOVery (Zhou et al., 2020a) is a dataset containing 2,029 news articles as well as data for 140,820 tweets and the respective user profiles sharing these articles. Its labels were obtained based on the reliability of the news publisher.

Lastly, CoAID (Cui and Lee, 2020) comprises different types of news items, i.e. claims, news articles and social media posts. Specifically, it consists of 204 fake and 3,565 real fact-checked news articles referring to the pandemic, 28 wrong and 454 true claims about the virus, 926 social media posts about it and data about Tweets reposting these news, as well as data about user engagements with each tweet. With 94% of its news item representing authentic news, it is also the most imbalanced of the datasets. Veracity labels for real news articles were obtained on a publisher level, i.e. data were scraped from reliable publishers such as ScienceDaily or WHO. Fake articles were gathered using links to false articles from fact-checking websites. For claims, WHO information and additional reliable sources were used to identify statements known to be false.

Not only are the existing datasets scarce, they are also limited in terms of the amount of text and types of information they provide and, at times, are highly imbalanced, introducing inductive bias. No COVID-19 datasets are available for the German language.

2.2 COVID-19 Fake News Detection

Given the nature of the available Covid-19 datasets, most attempts to detect fake news in this domain have been language-based. The best performing approach in the CONSTRAINT shared task on the Covid19 Fake News dataset was proposed by Glazkova et al. (2020), who obtained text representations using the COVID-Twitter-BERT language model (Müller et al., 2020) that was specifically trained on 160M English tweets about the virus

[2]Knowledge-based fake news detection es techniques from information retrieval to determine the veracity/truthfulness of news item, whereas propagation-based approaches use network analysis to determine the credibility of news sources at various stages, i.e. when being created, published online and spread via social media.

Table 1: Existing datasets for Covid-19 fake news detection. Engagement refers to data about reactions to and shares of news on social media whereas user data refers to information about user profiles sharing news articles such as how frequently they post, their follower count, etc.

Dataset	Language	Size	Fake / Real	Engagement	Profiles	Data Source
Covid19 FN	English	10,700	0.47 / 0.53	-	-	Various
FakeCovid	Various	5,182	0.20 / 0.80	-	-	Articles
ReCOVery	English	2,029	0.33 / 0.67	✓	✓	Articles
CoAID	English	4,251	0.06 / 0.94	✓	-	Articles, Claims
FANG-COVID	German	41,242	0.32 / 0.68	✓	✓	Articles

and achieved an F1-score of 98.69%. High detection performance (F1-score of 96.7%) was also reached based on concatenated text representations obtained by XLNet (Yang et al., 2019) and Latent Dirichet Allocation (LDA) in a feed forward neural network (Gautam et al., 2021) (see Patwa et al. (2021) for comprehensive overview of the approaches proposed in the CONSTRAINT shared task). Experiments conducted on the FakeCovid dataset used a pretrained BERT model to represent the textual content of English news articles from FakeCovid and obtained an F1-score of 0.78 (Shahi and Nandini, 2020). For ReCOVery, the SAFE method (Zhou et al., 2020b), an approach that involves training a neural network to specifically detect discrepancies in visual and textual news content, outperformed approaches based on text-based CNN, achieving an F1-score of 0.833 for the detection of real and 0.672 for the detection of fake news. Cui and Lee (2020) conduct experiments on the CoAID dataset using various different classification models that have proven to be successful for general fake news detection. Notably, some of the best-performing models on general fake news detection datasets such as CSI (Ruchansky et al., 2017) and SAME (Cui et al., 2019) proved to be less effective on CoAID, achieving F1-scores of 0.228 and 0.340, respectively, relative to the best performing model dEFEND (Shu et al., 2019a) with an F1-score of 0.581. This further demonstrates the need for research on specific critical events like the current pandemic.

3 Introducing the FANG-COVID dataset

In this section, we provide a detailed description of the methodology used to compile the FANG-COVID dataset and present its composition. The collection and labelling of news articles was based on news publisher bias lists compiled by media professionals, following the approach taken in Kiesel et al. (2019). For gathering true news articles, we relied on three established, reputable mainstream newspapers whose quality is recognized by media experts (Wellbrock, 2011). For the collection of fake news articles, we utilized independent fact-checking organizations such as Correctiv[3] and NewsGuard[4]. The latter published a list in January 2021 containing over 30 German news sources that are classified as unreliable news publishers based on criteria such as the frequency of false information being posted, the responsible presentation of information and sources and the clear separation of opinion and fact [5]. Based on this data and further research on its listed sources via trustworthy newspapers such as *Frankfurter Allgemeine Zeitung*, *SPIEGEL* or *ZEIT*, we selected ten publishers whose articles we would label as 'fake'. An overview of the distributions of news articles and tweets across publishers is presented in Table 3. For every one of our news sources, we scraped the header, date of publication and textual content of articles that were published between February 2020 to mid March 2021 and contained one of the following Covid-related keywords: *Corona, Covid, Infektion, Lockdown, Impfen, Impfung, Impfstoff*. The selection of news based on the keywords ensured that all articles are related to events surrounding the coronavirus pandemic. The collected textual data were cleaned and preprocessed by transforming dates of publication from handwritten formats to the uniform datetime format and the removal of HTML tags. A schematic representation of the procedure used to compile FANG-COVID is shown in Figure 1.

In total, the FANG-COVID dataset comprises 41,242 news articles: 28,056 articles from three publishers are labeled as 'real' and 13,186 articles

[3]https://correctiv.org/

[4]https://www.newsguardtech.com/

[5]https://www.newsguardtech.com/ratings/rating-process-criteria/

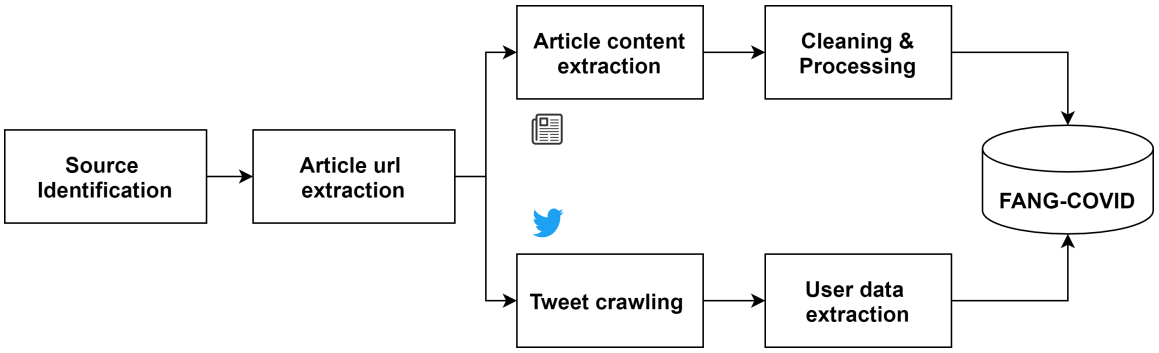


Figure 1: Procedure applied to compile the FANG-COVID dataset

are labeled as 'fake' based on the reliability of their source (see Tables 2 and 3 for details on the composition of the dataset). On average, news articles in FANG-COVID contain 803 word tokens and 43.8 sentences, with similar distributions for fake and real news articles. With regards to political orientation, the majority of the fake news publishers exhibited tendencies towards right wing populism, as common topics by these publishers involve allegations against immigrants (a topic that has been especially prevalent in Germany since the refugee crisis in 2015). The articles from unreliable sources are also characterized by a higher proportion of articles favouring the right-wing populist party "Alternative für Deutschland" (AfD) (see Figure 4 in the Appendix).

To facilitate research on the diffusion of fake news in social networks, FANG-COVID also contains rich information relating to the articles' spreading on social media. These data were collected using the snscrape[6] library, a Python scraper for social networking services. We gathered all tweets referring to any of the articles in the dataset, as well as reactions to these in the form of likes or retweets and additional information, such as which device the tweet was posted from as well as user-related data, such as the follower count, the number of posts or the date a user joined Twitter. Each news article is associated with a JSON array containing all tweets sharing its url, ordered by the time they were posted. Each tweet is represented as a JSON object containing specific information about each tweet, such as the number of likes and retweets as well as extensive data about the user sharing each tweet such as their follower count, their number of posts since joining Twitter as well as visual information in the form of their profile picture and banner. Overall, FANG-COVID contains social media information related to the news articles from a total of 363,393 tweets. On average, a real news article from our dataset was shared 8.4 times while a fake news article was shared 9.7 times. Out of all users sharing articles from our dataset, 52,289 distinct user profiles reposted real news articles while 7,861 users shared news articles labeled as fake.

4 *FANG-COVID:* Benchmark Experiments

In this section, we first describe the textual and social context representations used in the fake news detection model and the experimental setup, before presenting the empirical results of classification models trained on interpretable features and contextualized word embeddings. Specifically, in order to establish benchmarks for fake news detection on the FANG-COVID dataset and to provide initial insights into the features and approaches that prove to be effective for the detection of German fake news regarding COVID-19, we evaluate detection models based on interpretable Term Frequency Inverse Document Frequency (TF-IDF) and linguistic complexity features and compare their performance to classifiers based on a fine-tuned BERT language model with and without social context information.

4.1 Textual Representations

Interpretable representation: Our proposed interpretable text representation $T_i = t_t \oplus t_c$ consists of two concatenated vectors t_t, t_c where t_t is a 600-dimensional TF-IDF vector measuring the frequencies of single words and bigrams in a given text and t_c is a 718-dimensional vector capturing the linguistic complexity of a given text. This vector is generated using CoCoGen, a computational tool that implements a sliding window technique to calculate within-text distributions of feature scores (see

[6]https://github.com/JustAnotherArchivist/snscrape

Table 2: Text size statistics of real articles, fake articles and the whole dataset. Tokenization was performed using the Stanford CoreNLP library.

Measure	Real	Fake	All
Average number of tokens per article	784	803	790
Average number of sentences per article	50.40	43.82	48.30
Average number of tokens per sentence	15.68	18.32	16.36
Total number of types in corpus	424,529	270,961	543,720

Table 3: Distribution of articles and associated tweets across publishers along with average article length (in words)

Publisher	Number of Articles	Average article length	Number of Tweets	Number of distinct users sharing articles
Reliable Publishers:				
Sueddeutsche Zeitung	6,749	673.3	29,009	11,709
Tagesspiegel	11,623	699.7	75,100	22,248
ZEIT	9,684	1,023.7	130,945	34,264
Unreliable Publishers:				
AnonymousNews	229	689.3	8,078	1,368
Compact-Online	1,077	708.9	10,211	1,276
Contra-Magazin	998	691.2	2,621	181
FreieWelt	1,403	367.3	5,023	877
Journalistenwatch	3,675	683.1	61,299	3,480
Kopp-Report	204	1,017.0	1,932	436
Politikstube	602	239.5	3,896	649
Pravda-TV	1,117	1,612.6	4,575	496
RT-DE	2,442	504.7	15,986	2,131
Rubikon News	1,439	2,132.3	14,718	2,637

recently published papers that use this tool, (Ströbel et al., 2018; Kerz et al., 2020b,a; Qiao et al., 2020)). In contrast to the standard approach implemented in other tools for automated text analysis that rely on aggregate scores representing the average value of a feature in a text, the sliding-window approach generates a series of measurements representing the 'local' distributions of scores. A sliding window can be conceived of as a window of size ws, which is defined by the number of sentences it contains. The window is moved across a text sentence-by-sentence, computing one value per window for a given feature. The series of measurements faithfully captures a typically non-uniform distribution of features within a text and is referred here to as a 'contour'. To compute the value of a given feature in a given window m ($w(m)$), a measurement function is called for each sentence in the window and returns a fraction (wn_m/wd_m). The series of measurements generated by CoCoGen captures the progression of language performance within a text for a given indicator and is referred here to as a 'complexity contour' (see Figure 2 for illustration). CoCoGen uses the Stanford CoreNLP suite (Manning et al., 2014) for performing tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic parsing (Probabilistic Context Free Grammar Parser (Klein and Manning, 2003)). For German, CoCoGen currently supports 118 linguistic feature scores that fall into five categories: (1) features of syntactic complexity (N=5), (2) features of lexical density, sophistication and variation (N=17), (3) features of morphological complexity (N=27), (4) information-theoretic features (N=1) and (5) LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2010) features (N=68).

Thus, for a given news article, CoCoGen outputs a list $S = [s_1, s_2, ..., s_l]$ of 118-dimensional sentence representations where l is the length of the input text in sentences. These sentence representations serve as input for a 2-layer bidirectional LSTM-network (Graves and Schmidhuber, 2005) with a hidden state size of 150. t_c consists of the concatenated outputs from each last hidden state layer from both directions as well as a 118-dimensional vector containing the sum $s_1 + ... + s_l$

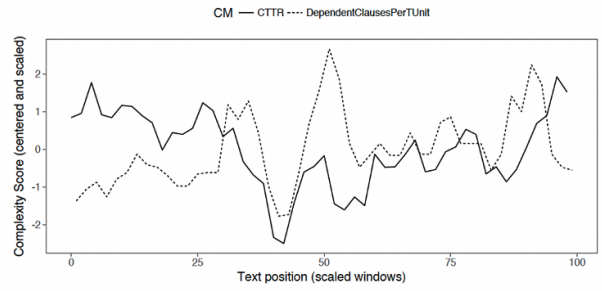


Figure 2: Schematic representation of 'complexity contours' for two out of 118 complexity measures investigated

of all feature values for each sentence.

Black-Box representation: As a textual black-box representation T_b, we use a pretrained German[7] BERT model (Devlin et al., 2019) and fine-tune six of its twelve encoding layers for our classification task. We use BERT's 768-dimensional output for its classification token [CLS] as our text representation. For texts of lengths greater than 512 tokens, we only keep the first 512 to serve as input to BERT as this method has been shown not to perform meaningfully worse than hierarchical methods (Sun et al., 2019).

4.2 Social Context Representation

As social context has been shown to be a valuable source of information for the detection of fake news (Shu et al., 2019c; Zhou and Zafarani, 2019), we employed a recurrent neural network-based representation for our social media data input for our classifiers, following an approach proposed in Ruchansky et al. (2017). For every article, a list of Twitter post representations $P = [p_1, p_2, ..., p_n]$ is constructed where p_i is an eight-dimensional vector representing user interactions with a tweet sharing the article as well as information about the user posting it. The list is ordered by the time the tweets were posted. The values contained in each vector are the tweet's number of likes, number of replies, number of retweets and its number of quotes as well as the poster's follower count, number of friends, total number of tweets and number of tweets they liked. The post representations serve as input to a two-layer bidirectional LSTM network with a hidden state size of 12. The final social context vector S is constructed by concatenating the four last hidden states from both layers in both directions as well as a five-dimensional vector containing values for the total number of tweets posting the corresponding news article as well as the total number of likes, replies, retweets and quotes to these posts. Thus, S is of dimension 53.

4.3 Experimental Setup

In this section we report on benchmark experiments conducted with different combinations of feature representations to assess how an explainable textual- and social context-based model compares to a black box model based on contextualised word embeddings in a binary classification task. Specifically, we evaluate the following four feature combinations: (1) Our interpretable textual representation T_i consisting of the TF-IDF vector and the linguistic complexity representation obtained from CoCoGen features, henceforth referred to as **TF-IDF + CoCoGen**, (2) the interpretable textual representation T_i and social context representation S, henceforth referred to as **TF-IDF + CoCoGen + Social Context**, (3) the black-box textual representation T_b obtained from BERT, henceforth referred to as **BERT** and (4) the black-box textual representation T_b and social representation S, henceforth referred to as **BERT + Social Context**.

For all feature combinations, we concatenate the individual feature vectors and use a three-layer perceptron to classify news articles as fake or real. We use binary cross entropy loss as our loss function:

$$\mathcal{L}(p, t) = -t \log(p) - (1 - t) \log(1 - p)$$

where t is the target value defined as 1 for fake news and 0 for articles labeled as real and p is the probability of a given news article being fake.

We tested our models using two data splits. First, we perform 5-fold cross validation over news articles from the whole dataset, including articles from all publishers in the train and test set. Second, we perform experiments on a randomly chosen publisher-separated split in which none of the publishers from the training set are included in the test set in order to determine whether the models learn to identify publishers-specific clues.

4.4 Ablation Study

In order to assess the relative importance of our neural networks' features for the detection of fake news, we performed a feature ablation study on the results from the 5-fold cross validation split using a modified version of an algorithm introduced by Díaz-Villanueva et al. (2010): For each

[7] https://deepset.ai/german-bert

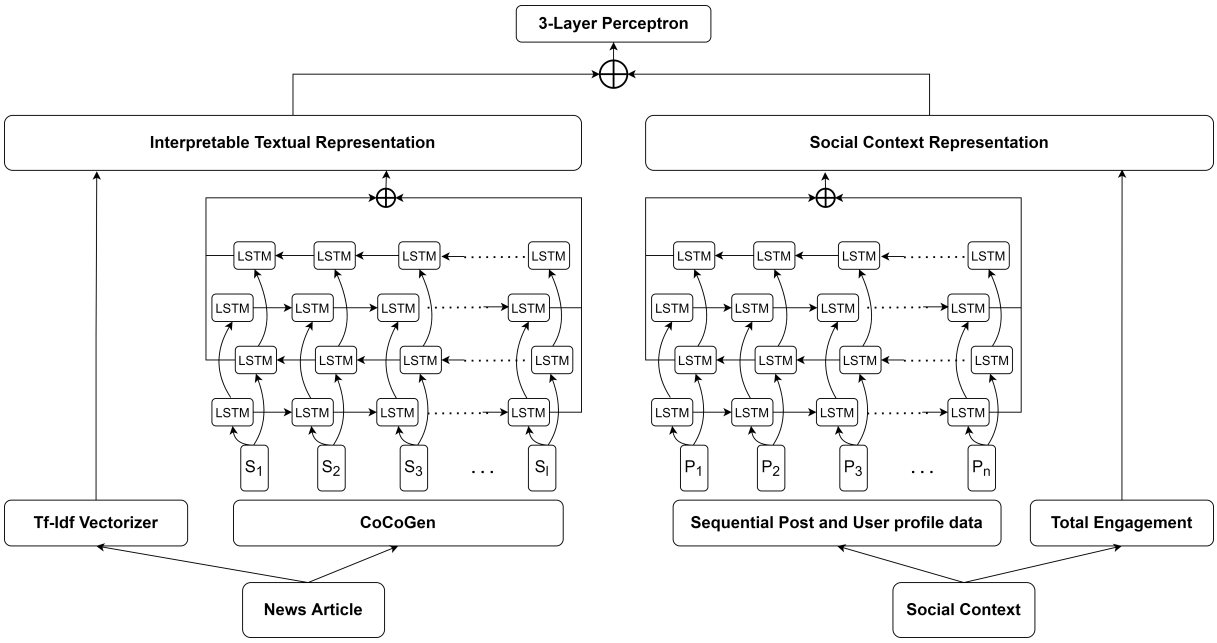


Figure 3: Architecture of proposed interpretable TF-IDF + CoCoGen+Social Context fake news detection model

step t, a neural network M_t is trained on a set of training data consisting of feature groups $F = \{f_1, f_2, ..., f_{D_t}\}$ where $f1, ..., f_{D_t}$ are the remaining feature groups at the current step, whose importance rank is to be determined. Let X_t denote the test dataset at time step t consisting of the feature set F_t and X_t^i denote the test dataset X_t with the values of features f_i being set to their average across the training dataset. Furthermore, let $acc(X)$ denote the classification accuracy of the given neural network M_t on test set X. For each time step t, the sensitivity score (Moody, 1994; Utans and Moody, 1991) $S_{i,t}$ for feature group f_i is computed as follows:

$$S_{i,t} = acc(X_t) - acc(X_t^i)$$

The most important feature group $f_{\hat{i}}$ at step t can be found by:

$$f_{\hat{i}} : \hat{i} =_{i:f_i \in F_t} (S_{i,t})$$

and the importance rank of feature group $f_{\hat{i}}$ is set as

$$Rank_{\hat{i}} = t$$

For the next step $t+1$, feature group $f_{\hat{i}}$ is dropped from the training and test dataset that are used to train and evaluate the neural network:

$$F_{t+1} = F_t - \{f_{\hat{i}}\}$$

The process is repeated until step t' where $|F_{t'}| = 1$

5 Results

We outline our full classification results in Table 4. Our best performing models are the BERT and BERT + Social Context models which achieve superior results on both the 5-fold validation split as well as the publisher-seperated split. Notably, the proposed model relying solely on interpretable features achieves results that are comparable to those of the black box model relying on BERT embeddings. While the chosen social context representation could not add additional predictive power to the model solely relying on BERT embeddings, its addition to the interpretable text embeddings improved the model performance by a large margin on both data splits. The results of the models on the publisher-separated split suggest that both our interpretable models as well as the BERT-based models have learned publisher-specific features. These results indicate that future work should take measures that ensure that the models solve the task they are trained for, rather than learning stylistic features of specific sources (see Baly et al., 2020, for a related finding in the domain of prediction of political ideology).

For our interpretable model based on social context and textual data, the feature group of highest importance is the data provided about a user sharing an article, followed by CoCoGen features measuring the syntactic complexity of a given news article. Specifically, profiles sharing real news articles

Table 4: Results of fake news detection performance on FANG-COVID

Model	Accuracy	Precision	Recall	F1-Score
5-fold validation:				
TF-IDF + CoCoGen	0.888	0.839	0.798	0.817
TF-IDF + CoCoGen + Social Context	0.937	0.913	0.887	0.900
BERT	0.981	0.966	0.970	0.968
BERT + Social Context	0.981	0.957	0.976	0.966
Publisher-seperated split:				
TF-IDF + CoCoGen	0.758	0.590	0.524	0.555
TF-IDF + CoCoGen + Social Context	0.801	0.649	0.653	0.651
BERT	0.820	0.676	0.685	0.680
BERT + Social Context	0.824	0.672	0.695	0.683

Table 5: Results of feature ablation study for Tf-IDF+CoCoGen+Social Context, evaluated on 5-fold validation: The sensitivity of the most important feature at time step t is the difference between the accuracy before drop and after drop

Step	Most important feature	Accuracy before drop	Accuracy after drop
1	Twitter profile data	0.941	0.718
2	Syntactic complexity	0.901	0.600
3	Tweet engagement	0.880	0.711
4	Lexical features	0.875	0.810
5	TF-IDF	0.850	0.791
6	LIWC-features	0.827	0.798
7	Morphological features	0.791	0.752
8	Infotheoretic features	0.749	0.384

have significantly higher average follower counts, but a lower amount of friends than those sharing fake news, indicating that our model learned to recognize specifically reputable official sources and institutions that predominantly shared real news. Other than expected, our model furthermore capitalized on fake news being on average more syntactically and lexically complex than reliable news articles which contradicts previous research from the English language suggesting that fake news are designed to be specifically easy to read, thus avoiding longer sentences and sophisticated language (Horne and Adali, 2017). Surprisingly, TF-IDF vectors merely rank as the 5th strongest feature and LIWC features that have shown to be a powerful feature for English fake news detection (Qiao et al., 2020; Shu et al., 2019b) rank as the 6th most important feature group out of eight. The full ablation results can be found in Table 5.

6 Conclusion

The development of effective diagnostic and predictive models needed to combat the spread and consumption of disinformation requires high-quality benchmark datasets. In this paper, we introduced the FANG-COVID dataset, a new dataset for fake news detection in German, consisting of 40k+ real and fake news articles related to the COVID-19 pandemic along with data on their propagation on Twitter. To facilitate fake news related research, the dataset is publicly available to the research community on GitHub.

We further conducted benchmark experiments on the data that incorporated both uninterpretable and interpretable features to language-based fake news detection in conjunction with social media engagement and profile data. The results of these experiments indicate that models that combine interpretable language features with social network data can achieve comparable detection performance relative to those based on contextualized word embeddings. This finding is important in the light of the growing recognition of the need to move away from pure black-box models towards interpretable models for solving practical problems, in particular in the context of critical industries, including healthcare, criminal justice, and news (Rudin, 2019). This is due to the fact that human experts

in a given application domain need both accurate but also understandable models (Loyola-Gonzalez, 2019).

Potentially, our work can be expanded upon with the addition of further linguistic features that can increase the predictive power of an interpretable model to match that of black box models. Furthermore, a wide range of data from the social context information provided in FANG-COVID can be integrated to build more sophisticated fake news detection models and to study the propagation of German fake news on social media. To address issues associated with the high correlation between the source of a news article and its class, approaches integrating adversarial objectives (Baly et al., 2020) could be evaluated.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Ramy Baly, Giovanni Da San Martino, James R. Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. *CoRR*, abs/2010.05338.

Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv e-prints*, page arXiv:2006.00885.

Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. Same: Sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 41–48, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wladimiro Díaz-Villanueva, Francesc J. Ferri, and Vicente Cerverón. 2010. Learning improved feature rankings through decremental input pruning for support vector based drug activity prediction. In *Trends in Applied Intelligent Systems*, pages 653–661, Berlin, Heidelberg. Springer Berlin Heidelberg.

Akansha Gautam, Venktesh V, and Sarah Masud. 2021. Fake news detection system using xlnet model with topic distributions: Constraint@aaai2021 shared task. *CoRR*, abs/2101.11425.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2020. g2tmn at constraint@aaai2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. *CoRR*, abs/2012.11967.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398.

Elma Kerz, Fabio Pruneri, Daniel Wiechmann, Yu Qiao, and Marcus Ströbel. 2020a. Understanding the dynamics of second language writing through keystroke logging and complexity contours. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 182–188.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020b. Becoming linguistically mature: Modeling english and german children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430.

Sebastian Kula, Michał Choraś, Rafał Kozik, Paweł Ksieniewicz, and Michał Woźniak. 2020. Sentiment analysis for fake news detection by means of neural networks. In *Computational Science – ICCS 2020*, pages 653–666, Cham. Springer International Publishing.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Prismatic Inc, Steven J. Bethard, and David Mcclosky. 2014. The stanford corenlp natural language processing toolkit. In *In ACL, System Demonstrations*.

John Moody. 1994. Prediction risk and architecture selection for neural networks. In *From Statistics to Neural Networks*, pages 147–165, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on twitter. *CoRR*, abs/2005.07503.

Daniele Orso, Nicola Federici, Roberto Copetti, Luigi Vetrugno, and Tiziana Bove. 2020. Infodemic and the spread of fake news in the covid-19-era. *European Journal of Emergency Medicine*.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 42–53, Cham. Springer International Publishing.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: COVID-19 fake news dataset. *CoRR*, abs/2011.03327.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A language-based approach to fake news detection through interpretable features and BRNN. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 14–31, Barcelona, Spain (Online). Association for Computational Linguistics.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Gautam Kishore Shahi and D. Nandini. 2020. Fakecovid - a multilingual cross-domain fact check news dataset for covid-19. *ArXiv*, abs/2006.11343.

Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H. Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020a. Combating disinformation in a social media age. *WIREs Data Mining and Knowledge Discovery*, 10(6):e1385.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019b. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020b. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188. PMID: 32491943.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2019c. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Daniel H Solomon, Richard Bucala, Mariana J Kaplan, and Peter A Nigrovic. 2020. The "infodemic" of covid-19.

Marcus Ströbel, Elma Kerz, Daniel Wiechmann, and Yu Qiao. 2018. Text genre classification based on linguistic complexity contours using a recurrent neural network. In *MRC@ IJCAI*, pages 56–63.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political*

polarization, and political disinformation: a review of the scientific literature (March 19, 2018).

J. Utans and J. Moody. 1991. Selecting neural network architectures via the prediction risk: application to corporate bond rating prediction. In *Proceedings First International Conference on Artificial Intelligence Applications on Wall Street*, pages 35–41.

Inna Vogel, Jeong-Eun Choi, and Meghana Meghana. 2020. Topic related corpus for fake news detection.

Inna Vogel and Peter Jiang. 2019a. Fake news detection with the new german dataset "germanfakenc". In *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*, pages 288–295.

Inna Vogel and Peter Jiang. 2019b. Fake news detection with the new german dataset "GermanFakeNC". In *International Conference on Theory and Practice of Digital Libraries*, pages 288–295. Springer.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Christian M Wellbrock. 2011. Die journalistische qualität deutscher tageszeitungen–ein ranking. *Medienwirtschaft*, 8(2):22–31.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020a. Recovery: A multimodal repository for COVID-19 news credibility research. *CoRR*, abs/2006.05557.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020b. SAFE: similarity-aware multi-modal fake news detection. *CoRR*, abs/2003.04981.

Xinyi Zhou and R. Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *ArXiv*, abs/1812.00315.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *SIGKDD Explor. Newsl.*, 21(2):48–60.

A Appendix

A.1 Preprocessing:

The following preprocessing steps were taken before using our texts as input to our models (in order to avoid models relying on trivial features):

- removal of phrase 'Bleiben Sie aufmerksam!', very prevalent in PravdaTV's articles
- removal of advertisement paragraph for book "Illuminatenblut: Die okkulten Rituale der Elite" by PravdaTV" ('... Wenn Sie mehr über die heimlichen Machenschaften der Elite erfahren wollen,...')
- removal of paragraphs that contain (capitalized) 'COMPACT' and 'Am besten gleich hier bestellen'. Compact-Online spreads a lot of advertisement for their magazines in their articles, mentioning COMPACT-Sonderausgabe, COMPACT-Magazin, etc.; since their phrases were almost always similar, but not exactly the same, it was hard to find a systematic approach to remove these advertisements. Thus, we removed the paragraphs from these articles as a whole in order to make sure there's no bias in our dataset
- removal of text brackets with text () at the end of each article. Many publishers appended (dpa) or similar to their articles
- removal of 'von [Name]' at end of first paragraph. Some publishers used this to reference the author in their articles
- removal of pattern 'rt/dpa', 'ns/sna/tm' at the end of news text

Table 6: Data on the average user engagement for each article

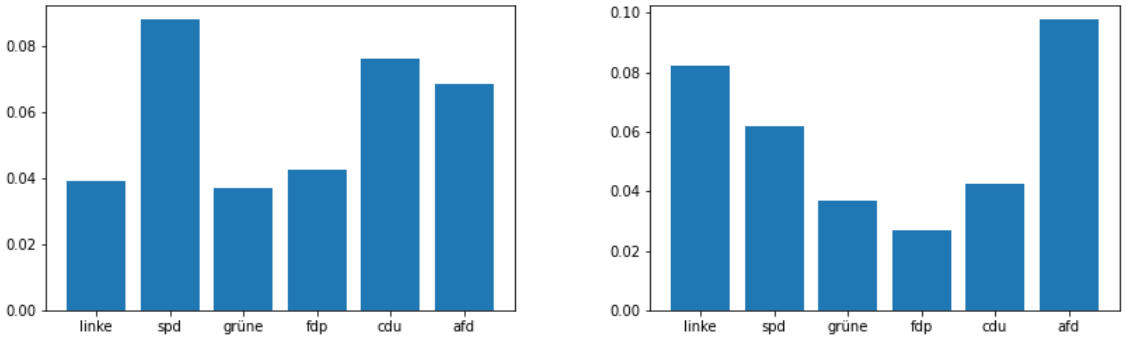
Engagement Feature	Real	Fake
Feature per tweet:		
Number of likes	8.90	3.48
Number of retweets	2.72	1.85
Number of replies	1.40	0.50
Number of quotes	0.43	0.19
Feature per article:		
Number of tweets	8.4	9.7
Number of likes	74.6	33.9
Number of retweets	22.8	18.0
Number of replies	11.7	4.9
Number of quotes	3.6	1.9

Table 7: Data on the average user profile for each tweet sharing an article

Profile Feature	Real	Fake
Number of followers	129,714	2,618
Number of friends	1,304	1,791
Overall number of tweets	67,732	55,093
Overall number of liked tweets	23,565	23,197

Table 8: Data on the syntactic complexity of news articles

Syntactic Feature	Real	Fake
Mean Word Length (in characters)	5.78	5.90
Clauses per Sentence	1.74	1.90
Coordinate Phrases per Clause	0.44	0.57
Mean length of Clause (in tokens)	9.25	9.99
Mean length of Sentence (in tokens)	16.0	19.2



(a) Reliable publishers' mentions

(b) Unreliable publishers' mentions

Figure 4: Average number of mentions of political parties in the German parliament per news article by publisher