Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2702–2713
Marseille, 20-25 June 2022

Applying Automatic Text Summarization for Fake News Detection

Philipp Hartl, Udo Kruschwitz
University of Regensburg,
Universitätsstraße 31, 93053 Regensburg, Germany
philipp1.hartl@stud.uni-regensburg.de, udo.kruschwitz@ur.de

Abstract

The distribution of *fake news* is not a new but a rapidly growing problem. The shift to news consumption via social media has been one of the drivers for the spread of misleading and deliberately wrong information, as in addition to its ease of use there is rarely any veracity monitoring. Due to the harmful effects of such *fake news* on society, the detection of these has become increasingly important. We present an approach to the problem that combines the power of transformer-based language models while simultaneously addressing one of their inherent problems. Our framework, CMTR-BERT, combines multiple text representations, with the goal of circumventing sequential limits and related loss of information the underlying transformer architecture typically suffers from. Additionally, it enables the incorporation of contextual information. Extensive experiments on two very different, publicly available datasets demonstrate that our approach is able to set new state-of-the-art performance benchmarks. Apart from the benefit of using automatic text summarization techniques we also find that the incorporation of contextual information contributes to performance gains.

Keywords: Fake News Detection, Text Summarization, BERT, Ensemble

1. Introduction

With the rise of the Internet as the most influential information medium, the consumption of news and information has changed substantially. More recently, social media has become the primary source of information, changing this yet again (Shearer and Mitchell, 2021). Unfortunately, there are typically little to no checks on what information is posted and its veracity, thereby enabling the wide spread of *fake news* – intentionally and verifiably false information with the purpose of deceiving its reader (Allcott and Gentzkow, 2017). The scale of the problem is such that it has become an urgent social and political issue (Nakov et al., 2021a). The current global pandemic, for example, has even demonstrated that false information can be life-threatening (Marco-Franco et al., 2021). Fittingly, the World Health Organization (WHO) is talking about fighting not only a pandemic, but also an infodemic (Hua and Shaw, 2020), a flood of information, including false and misleading information, which causes confusion amongst the public. Unfortunately, it is typically not a trivial task for humans to judge whether a piece of information is false or not (Rubin, 2010).

In order to push forward the state of the art (SOTA) in fake news detection we present an end-to-end deep learning approach based on the Transformer architecture (Vaswani et al., 2017) at the core of which we incorporate different methods to transform the original text into some condensed form. Due to architectural restrictions, transformer-based models like BERT are limited to specific input sequence lengths (Devlin et al., 2019), which are shorter than many news articles (Souma et al., 2019). To better capture the missing information, we therefore propose **CMTR-BERT** (**C**ontextual **M**ulti-**T**ext **R**epresentations for fake news detection with **BERT**) which is an ensemble of BERT models. CMTR-BERT is particularly aimed at longer sequences and additional contextual information. The proposed model incorporates three different ways to deal with long sequences, namely a simplified *hierarchical transformer* representation adopted from Pappagari et al. (2019), *extractive* as well as *abstractive* text summarization. Also, the model enables contextual data to be incorporated for fake news detection via additional BERT embeddings. Furthermore, the high-level architecture is language-agnostic, thereby offering plenty of future directions to reproduce our experiments in other languages.

To the best of our knowledge, this is the first attempt at utilizing automatic text summarization to reduce text complexity for fake news classification.

The main contributions of this work are as follows:

- We propose an end-to-end deep learning framework to integrate different news and social context features for fake news detection.
- We use automatic text summarization techniques to circumvent information loss on long sequences.
- We combine different textual representations for classification.
- Using different benchmark datasets, we empirically investigate the influence of social context and automatic text summarization on fake news detection performance.
- To foster reproducibility, we make our code and models available to the community.[1]

[1] https://github.com/phHartl/lrec_2022

2. Related Work

2.1. Fake News Detection

Fake News detection systems typically adopt one of three general approaches or a combination of them (Sharma et al., 2019). The most commonly used way is based on the *content*, which can be either linguistic, auditory (e.g., attached voice recordings) or visual (e.g., images or videos). These approaches are often either knowledge- or style-based. The former uses methods of information retrieval to extract concrete statements, which are then automatically checked against knowledge graphs (Pan et al., 2018) or documents retrieved from the web on the fly (Magdy and Wanas, 2010). While these approaches are the most straightforward, wrong information and missing knowledge about recent topics limit its applicability. The latter typically produce either certain interpretable cues (Vrij, 2005; Lesce, 1990; Pennebaker et al., 2001) or apply a more general linguistic analysis and focus on deception or objectivity detection (Feng et al., 2012; Rubin et al., 2015). Linguistic approaches are however often outclassed by deep learning concepts such as LSTM architectures (Bahad et al., 2019) or attention-based networks (Bahad et al., 2019). Approaches which only focus on the content might miss valuable context information. Hence, *context-based* solutions target secondary information such as user engagements (Shu et al., 2019) and dissemination networks (Shu et al., 2020b) on social platforms. These methods are based on the assumption, that there is a difference in interaction between fake and real news. This can either be done by using hand-engineered features (Ding et al., 2020b), propagation (Shu et al., 2020b), temporal (Ferrara, 2020) or stance analysis (Sobhani et al., 2016). While contextual information can be useful when available, it is often not or only partially available. *Intervention-based* methods try to dynamically interpret real-time dissemination data. These are arguably the least common approaches used at the moment, because of their difficult way to evaluate (Sharma et al., 2019). When used though, they try to intervene the process of fake news spreading through e.g., injecting of true news into social networks (Farajtabar et al., 2017) or user intervention (Papanastasiou, 2020; Kim et al., 2018).

2.2. Attention-based Systems

Recently, transformer-based approaches have led to a paradigmatic shift in NLP dominating various leaderboards ranging from question-answering[2] to fake news detection (Ding et al., 2020a). Their huge advantage comes from models like *BERT* (Devlin et al., 2019). One of the main drawbacks of those, however, is the maximum sequence length each model is able to process, which comes at a maximum of 512 tokens (word pieces) for BERT. Unfortunately, fake news articles often are a longer than this value (Souma et al., 2019). By default, BERT-based models simply truncate the text to the desired input length. This leads to the loss of potentially important information in the later parts of the input text. To tackle this problem, Dai et al. (2019) developed Transformer-XL, which splits the input sequence in smaller chunks and injects the self-attention of the previous part into the next one as additional context with relative positional encodings. While this solves the problem of longer input sequences, it removes the bidirectional property of BERT, which is one of its major advantages and takes a lot of computing time to train. Pappagari et al. (2019) introduced hierarchical transformer representations, which is a conceptually similar approach, but built on top of BERT. Another possible solution is to again separate the document into different segments, but now focus on the *[CLS]* token for each segment instead. This token is designed to provide an embedding for the entire sequence (e.g a sentence). Mulyar et al. (2020) use this strategy to classify clinical documents with ClinicalBERT (Alsentzer et al., 2019).

2.3. Automatic Text Summarization

One of the most common and effective ways for humans to learn are summaries (Dunlosky et al., 2013). Their objective is to produce a representation, which includes the main ideas of the input (Radev et al., 2002), while being also shorter than it (Radev et al., 2002; Tas and Kiyani, 2017) and which additionally should avoid repetitions (Moratanch and Chitrakala, 2017). Similar to the human concept of attention, which previously has been used successfully in language understanding in e.g. Transformers, summarizations might be able to reduce the textual scope for e.g. BERT models while also incorporating thoughts which normally would have been lost. Usually, summarizations are either extractive or abstractive. *Extractive* summarization focuses on extracting key phrases from the input document. These snippets are then concatenated into a summarization of desired length. The goal of *abstractive* summaries is to generate a new text by paraphrasing the main concepts of the input sequence in fewer and clearer words (Moratanch and Chitrakala, 2016). This is a more human approach to text summarization, but also requires the ability to actually generate new text, which in itself is difficult (El-Kassas et al., 2021).

2.4. Datasets

Fake News detection has often been limited by data quality and availability. Most datasets adapt labels directly assigned by journalists (Wang, 2017), but there also exist approaches which condense the labelling into fewer labels (Shu et al., 2020a) or calculate them via a scoring system (Zhou et al., 2020b). The domain varies, but tends to be of political nature (Silverman et al., 2016; Wang, 2017) or interest (Li et al., 2020; Zhou et al., 2020b). There are datasets which only

[2] https://rajpurkar.github.io/SQuAD-explorer/

contain short statements (Hanselowski et al., 2018) or long texts (Shu et al., 2020a), based on social media text data (Mitra and Gilbert, 2015; Ma et al., 2017) or actual news articles (Nørregaard et al., 2019). Especially earlier datasets, only contained textual information and lacked any additional contextual information (Silverman et al., 2016; Wang, 2017). However, due to recent research in context-based fake news detection, datasets emerged which additionally got visual, social-context and spatio-temporal information (Shu et al., 2020a; Zhou et al., 2020b). Most of the datasets are in English (Guo et al., 2022), but there are also some in other languages (Vogel and Jiang, 2019) and even multilingual ones (Li et al., 2020).

2.5. Concluding Remarks

Our exploration of related work suggests there is a gap in applying automatic text summarization to tackle the problem of sequence limits for fake news detection. We suspect a positive influence of summarizations on classification performance due to the additional information present in the text, which normally would have been lost when using Transformers.

3. Methodology

We will now introduce CMTR-BERT (**C**ontextual **M**ulti-**T**ext **R**epresentations for fake news detection with **BERT**), a BERT-based ensemble model which uses a combination of different text representations and additional context information.

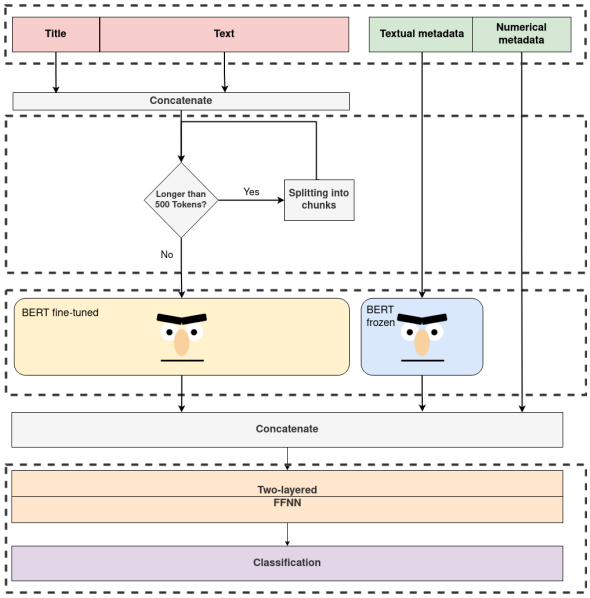

Figure 1: Model architecture (best viewed in colour)

This is a hybrid approach of content- & context-based fake news detection. It is based on the assumption of stylistic differences between fake and real information, combined with the different social reactions towards them. The basic model architecture is inspired by Ostendorff et al. (2019), while the idea to use automatic summarizations as additional text representations has been inspired by Li and Zhou (2020). CMTR-BERT consists of four major components for each text representation (see Figure 1 from top to bottom):

1. Contextual feature extraction
2. Hierarchical input representation
3. A fine-tuned BERT & frozen BERT model
4. Classification

The following sections will explain each component in more detail.

3.1. Contextual Feature Extraction

To achieve the best classification performance, it is important to represent incoming information in a comprehensive way. In any text classification task, there is an input sequence of interest present. In the case of fake news, this is typically an article written by one or more agents. We consider this part to be the classification *content* (marked red in Figure 1), which for news articles consists of the *title* and its *text*. Additional *context* (marked green in Figure 1) can also be supplied when available, which can be of different modalities such as text, visual or auditorial nature. To be understood by machines though, content & context information needs to be transformed into numerical values. For textual inputs, this can easily be achieved with pre-trained BERT embeddings. For other modalities, CMTR-BERT can either consider them a numerical input, e.g. when there already has been a transformation into a matrix representation or alternatively as textual input. Models like image2sentence (Vinyals et al., 2016) provide a textual representation of inputs of other modalities. It is important to map the associations between *content* & *context*, to get contextual content representation. We decided to represent this connection with a concatenation of *content* and textual & numerical *context*.

3.2. Hierarchical Input Representation

To circumvent the problem of long sequences, we used a modified version of the hierarchical input representation (HIR) proposed by Pappagari et al. (2019). While in the original work the authors use an additional Transformer or LSTM on top of the BERT embeddings of the text chunks, we opted not to do this. Instead, we went for a concatenation of the embeddings afterwards. We did this because Mulyar et al. (2020) showed that the concatenation of sequence embeddings is either on-par or better than with an additional LSTM or Transformer on top. This furthermore conceptually and computationally facilitates the model, which is a desired characteristic. We kept the splitting of the input sequence into different parts with overlap, though.

3.3. BERT

This component consists of two instances of BERT (Devlin et al., 2019) as seen in Figure 1. To differentiate learning between content and context information, we decided to use two independent BERT instances. Originally we considered fine-tuning them both, but it was not feasible with our resources. Our graphical units (GPU) simply had not enough memory for both. We decided to fine-tune with the content information and freeze the other BERT model during training.

3.4. Classification

The output of the aforementioned BERT models is concatenated with additional numerical context information to get the complete contextual feature representation of both content and context. The resulting representation is able to illustrate the relationship between different input sequences in a sophisticated way. To obtain a class label, this sequence is passed into a FFNN. This neural network is optimized during training and learns the differences in the aforementioned representation which separates fake from true news.

3.5. Summarization

To further circumvent the loss of potentially important information, we use abstractive and extractive summarizations. Unfortunately, producing summarizations is extremely elaborate and to our knowledge there does not exist a dataset where summarizations are manually made for fake news or claim detection. Therefore, we produced such summaries automatically.

CMTR-BERT combines all three different text representations in a majority voting ensemble. Individual models (as displayed in Figure 1) are trained for the original text, abstractive and extractive summaries. Each of those then classifies the unknown input sequences, and the ensemble now decides via majority voting which class is finally assigned.

4. Implementation

In the following section, we provide an overview of how the model described in the previous section has been implemented. We explain the datasets used, the experimental setup & the classification problem(s). Everything has been implemented in Python using PyTorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2019). For more details, please refer to the project GitHub page.

4.1. Datasets

We chose two datasets. FakeNewsNet (Shu et al., 2020a) is a common reference collection. The other one is CT-FAN 21 (Shahi et al., 2021b), published for the 2021 CLEF CheckThatLab! Fake News Detection challenge 3a (Nakov et al., 2021b).

FakeNewsNet: In contrast to other datasets in the domain like LIAR (Wang, 2017) or FEVER (Thorne et al., 2018), FakeNewsNet not only provides the actual news text but also additional social context (e.g. Twitter interactions) and spatio-temporal information (e.g. Twitter user locations). The authors use two fact-checking websites (*Politifact* & *GossipCop*) to get relevant fake and real news. Both cover different domains of false information. The dataset is not available publicly due to legal reasons. Therefore, we used the official GitHub repository[3] to obtain our own, slightly different (e.g. due to removed articles), copy of the dataset.

CT-FAN 21: This dataset got four different classes to predict as defined in Shahi et al. (2021a). The distribution of each class in the provided training and test data can be seen in Table 1.

Dataset	False	Partially False	True	Other
Training	486	235	153	76
Test	113	141	69	41

Table 1: CT FAN 21 statistics

Additionally, through a data sharing agreement, it is forbidden to redistribute the dataset, identify individuals and the original entries on the fact-checking websites. Therefore, we refrained from finding this information, although it would have been useful for classification purposes as demonstrated on a similar task (Yuan et al., 2020).

4.2. Data Preparation

Before we can use the data, there is still need for some limited pre-processing. Due to the differences in the available information in the two datasets, the preparation is not exactly the same but conceptually similar.

Domain	Fake	Real
Politifact	375	449
GossipCop	4761	14954

Table 2: FakeNewsNet after preprocessing

FakeNewsNet: Before doing anything else, we removed all data points which did not contain any news article information or had no file available after running the data generation script (see Table 2). Afterwards, we converted the labels to numerical values. Before starting the training, we split the dataset into a training and a validation set using the common 80/20 split, which has been used before with this dataset (Cui et al., 2019; Zhou et al., 2020c). For both domains, the labels are not equally distributed, with true news being more prevalent than fake news. To circumvent this problem, we randomly oversampled the minority class during training with the imbalanced-learn package (Lemaître et al., 2017). Additionally, we generated *abstractive* and *extractive* summaries for each text once and saved them, as this took a considerable time and reduces the

[3] https://github.com/KaiDMML/FakeNewsNet

impact of the non-deterministic algorithms used. Before sending the text and the metadata into the model, we also tokenized and normalized the input data with the BertTokenizer.

CT-FAN 21: As the data of CT-FAN 21 is directly available, there was no need to remove any entries. Again, we started with converting all labels to numerical values. As the four classes are not equally distributed, we applied *random oversampling* as well. As there is a separate test set provided by the task organizers, we chose not to split the training data and train with the whole dataset. The generation of both summarizations, the tokenization and normalization is done the same way as for FakeNewsNet.

4.3. Model Implementation

4.3.1. Content and Context Extraction

For the content information, we used the news article's *title* and *text*. For FakeNewsNet, we gathered context information in the form of the *author* as well as the *source URL*. For the latter, we extracted the original URL when the Wayback Machine was used to get the article. This circumvents a potential bias of deleted/removed articles being primarily fake and easily identifiable via the URL. Furthermore, we tried to incorporate different aspects of the social context features provided by FakeNewsNet. We gathered all *tweet authors*, *tweet texts* and the *number of retweets* to have information about the interacting users, the posts and the response behaviour. We specifically did not go for any network representation as it is beyond the scope of this work.

4.3.2. Hierarchical Input Representation

In our hierarchical transformer variant, we split the text into *overlapping parts of 500 tokens* with a *stride length of 50*, as a qualitative examination resulted in these values performing the best.

4.3.3. BERT

Both instances of BERT are implemented using the *bert-base-uncased* model, with 12 encoder layers and hidden dimensions of 768, which are the default values. Due to limited computational resources, we could not use a more sophisticated BERT model like RoBERTa (Liu et al., 2019). One model learned the differences in the textual content (*title* & *text*) of the news articles, while the other provided the embeddings learned by BERT for additional textual information (*author*, *source URL*, *tweet authors* & *tweet texts*). For each of those, we concatenated the information with a view to the news article.

4.3.4. Classification

After passing the BERT models, the representations are getting concatenated into a 5377-dimensional representation. This includes four parts of textual news embeddings, each of dimension 768 as a result of the hierarchical input representation, one embedding of 768 dimension of all other textual information and a single dimension to represent the number of retweets. We artificially limited the news title & text representation to four BERT embedding blocks, as only a very small percentage ($< 5\%$) of the input texts is longer than that. However, a longer representation is possible if desired. On top of that concatenation, we use a fully connected FFNN with two layers with an ReLU activation function. To calculate the loss during back propagation, we use cross entropy.

4.3.5. Summarizations

For *extractive* summarization, we use the system implemented by Miller (2019), which already has been used before and ensures comparability (Li and Zhou, 2020). This method first embeds the sentences using BERT, clusters them afterwards, and then finds the sentences closest to the cluster's centroids. To better resolve appearing incoherences, we furthermore use the neuralcoref library[4]. We set a summarization ratio of 0.40 empirically, which is in line with recommendation for summary length (Radev et al., 2002).

We implemented an *abstractive* technique based on BART (Lewis et al., 2020). This model is specifically well suited for text generation, outperforming similar ones on question-answering tasks like SQuAD (Lewis et al., 2020). Because of the repetitive nature of greedy and beam search (Vijayakumar et al., 2016; Shao et al., 2017), we used *Top-K* (Fan et al., 2018) and *Top-p* sampling (Holtzman et al., 2019) for our summaries. The exact model we used is *sshleifer/distilbart-cnn-12-6*[5], which is a smaller BART model trained on a news summarization dataset by Hermann et al. (2015). In our final configuration, we used the 100 (Top-K) most likely words and a probability (Top-p) of 95%. However, like BERT, BART has also a maximum sequence limit of 1024 tokens. To circumvent this problem we used the technique described in Section 3.2, however, with a length of 1000 tokens. This ensures, that all text parts are taken into consideration when producing a summarization. We also tried to get a summarization ratio of roughly 40% for better comparability to the extractive approach. However, as both approaches are not deterministic, this cannot always be guaranteed.

4.4. Experimental Setup

For training, we represented each news content as a string concatenation of *[CLS] + title + [SEP] + text*, where *text* is either the original text or one of the two summaries produced and [CLS] is a classification token and [SEP] is a token to indicate a separation between two sentences. For FakeNewsNet, when applicable, we also gathered the additional metadata as described before. All tweet authors are concatenated together into a

[4] https://github.com/huggingface/neuralcoref

[5] https://huggingface.co/sshleifer/distilbart-cnn-12-6

Datasets	Metric	SAF	SENTI	RST	LIWC	HPFN	SAFE	dEFEND	BERT-baseline
Politifact	Accuracy	0.691	0.760	0.796	0.830	0.843	0.874	**0.904**	0.823
	Precision	-	0.810	0.821	0.855	0.835	0.889	**0.902**	0.805
	Recall	-	0.760	0.752	0.792	0.851	0.903	**0.956**	0.807
	F1	0.706	0.784	0.785	0.822	0.843	0.896	**0.928**	0.805
Gossipcop	Accuracy	0.689	0.740	0.600	0.725	**0.861**	0.838	0.808	0.790
	Precision	-	0.760	0.623	0.773	0.854	**0.857**	0.729	0.553
	Recall	-	0.740	0.596	0.637	0.869	**0.937**	0.782	0.693
	F1	0.717	0.750	0.614	0.698	0.791	**0.895**	0.755	0.614

Table 3: Baselines (best values in each row **marked**). Unless specified, values are taken from the original papers.

long string representation with a delimiter in between each. The *tweet texts* are concatenated together into one string representation. Additionally, we removed all duplicates. Tweets are separated by a *[SEP]* token.
For training, we use an 80/20 training/validation split for FakeNewsNet and trained with all training data for CT-FAN 21. We used the same initial random state and split for all configurations to provide comparability. We used a batch size of 8, an initial learning rate of 5e-5, a weight decay of 0.01 and three training epochs with an AdamW (Loshchilov and Hutter, 2018) optimizer. Everything was trained on four RTX 2080 Ti with 11 GB VRAM. This process was then repeated ten times and the average values are reported. For FakeNewsNet we report $Accuracy$, $Precision$, $Recall$ and F_1, while for CT-FAN 21 we report the corresponding macro values. These metrics are typically used to measure classification performance (Chen et al., 2018; Cui et al., 2019; Zhou et al., 2020a).
Additionally, these trials are used as a bootstrapping method for statistical analysis. We use paired non-parametric test measures for inferential analysis, as parametric variants are not applicable here (Jurafsky and Martin, 2021). Hence, we use the Wilcoxon signed-rank test, Friedman test (Friedman, 1937) and Nemenyi test for post-hoc analysis (Nemenyi, 1963). All statistical analysis has been implemented using scipy (Virtanen et al., 2020), pandas (McKinney, 2010) and scikit-posthocs [6]. We use a threshold of $p < 0.05$ to determine whether a significant difference is present or not.

5. Results

We now look at the results we obtained. This includes a performance contextualization with various comparable systems found in the literature (for more detailed results of an ablation study to testify which model component contributes most to classification performance, see the GitHub project page).

5.1. Baseline Systems

We compare our work against strong baselines reported in the literature. We will report results for CT-FAN 21, but will primarily focus on FakeNewsNet as that dataset contains contextual information.

- Two baselines use linguistic information. *RST* is text-only based and extracts rhetorical features and transforms it into a tree structure. *LIWC* is a linguistic cue set designed to identify psycholinguistic differences in texts (Pennebaker et al., 2001). Both values for RST and LIWC reported here are adopted from (Shu et al., 2020b).

- *SAF* (**S**ocial **A**rticle **F**usion) represents news content features with an encoder-decoder architecture and captures temporal social interactions with an LSTM. The results reported here are adopted from Shu et al. (2020a).

- *SENTI* - Ding et al. (2020b) trained a Naive Bayes classifier, a decision tree and a bidirectional LSTM. Their best-performing model is based on decision trees and abbreviated as *SENTI* by us.

- *HPFN* (**H**ierarchical **P**ropagation **N**etwork **F**eature) combines macro- and micro-level propagation networks with additional linguistic information. Here classification happens with a range of linguistic, social and temporal features (Shu et al., 2020b).

- *SAFE* (**S**imilarity-**A**ware **F**ak**E** news detection) uses multi-modal (textual and visual) information (Zhou et al., 2020c). They specifically investigate the similarity between textual and visual information for fake news detection. They use a modified version of TEXT-CNN (Kim, 2014) for both textual and visual information.

- *dEFEND* (**E**xplainable **F**ak**E** **N**ews **D**etection) focuses on explainable fake news detection. It consists of a news content encoder, a user comment encoder, sentence-comment co-attention (Shu et al., 2019).

- *BERT-baseline* is a *bert-base-uncased* model with default parameters and a classification layer on top. During training, only the last layer is optimized while BERT itself is frozen, apart from that the setup is identical to Section 4.4.

Detailed baseline results are shown in Table 3.

[6] https://scikit-posthocs.readthedocs.io/

Dataset	Metric	BERT	SAFE	dEFEND	CMTR-BERT O	CMTR-BERT A	CMTR-BERT E	CMTR-BERT C	CMTR-BERT
Politifact	Accuracy	0.924	0.874	0.904	0.950	0.948	0.950	0.912	**0.956**
	Precision	0.934	0.889	0.902	0.953	0.945	0.941	**0.977**	0.958
	Recall	0.897	0.903	**0.956**	0.936	0.943	0.952	0.827	0.947
	F1	0.914	0.896	0.928	0.944	0.943	0.946	0.895	**0.952**
Gossipcop	Accuracy	0.863	0.838	0.808	0.960	0.956	0.958	0.957	**0.963**
	Precision	0.741	0.854	0.729	0.926	0.917	0.924	0.859	**0.936**
	Recall	0.666	0.937	0.782	0.908	0.898	0.899	**0.985**	0.910
	F1	0.701	0.895	0.755	0.917	0.907	0.911	0.918	**0.923**

Table 4: Performance of CMTR-BERT compared to SOTA. The best values in each row are **marked**.

5.2. CMTR-BERT

For a fair comparison, we built several variants of CMTR-BERT, which we compare with either content-only systems or more complex systems presented here (for detailed configurations and corresponding results please consult the GitHub repository). We include the aforementioned BERT-baseline and a trained BERT model here, together with several CMTR-BERT variants. Additionally, there are variants of CMTR-BERT with only one text representation (**O**riginal text, **A**bstractive summaries or **E**xtractive summaries) and the complete ensemble models, either with content data (**CMTR-BERT**) or without (**CMTR-BERT C**).

5.3. Performance Analysis

An overview of results can be seen in Table 4 and Table 5 (FakeNewsNet) as well as Table 6 (CT-FAN 21). For FakeNewsNet we have the following observations:

- CMTR-BERT beats state-of-the-art systems on F1 for both datasets by large margins. This indicates the effectiveness of the outlined framework on a common fake news detection dataset (Table 4).
- Interestingly, we even get the highest F1 score for GossipCop when we remove *all* context features from our model, as seen in Table 5.
- Our selection of context features seems to capture the most important aspects quite well, as CMTR-BERT C performs extraordinary well for both datasets. Being on par with SAFE or better and outclassing dEFEND by a huge margin for GossipCop.
- Contextual information is more important for GossipCop than it is for Politifact. Without contextual information, CMTR-BERT performs worse for GossipCop than it does for Politifact, despite being on par otherwise (see Table 5).
- BERT alone performs very well for Politifact ($F_1 = 0.914$) but has problems with GossipCop ($F_1 = 0.701$). This performance amplifies the previous assumption, that contextual information is important for GossipCop.
- There seems to be little to no difference between original texts and abstractive or extractive summarizations, despite the severe reduction in textual scope.

Dataset	Metric	LIWC	CMTR-BERT O w/o context	CMTR-BERT A w/o context	CMTR-BERT E w/o context	CMTR-BERT w/o context
Politifact	Accuracy	0.830	0.930	0.918	0.923	**0.933**
	Precision	0.855	**0.937**	0.910	0.927	0.929
	Recall	0.792	0.907	0.912	0.904	**0.924**
	F1	0.822	0.921	0.910	0.915	**0.926**
Gossipcop	Accuracy	0.725	0.867	0.858	0.860	**0.869**
	Precision	0.773	**0.759**	0.737	0.744	0.772
	Recall	0.637	**0.657**	0.642	0.644	0.650
	F1	0.698	0.705	0.685	0.690	**0.706**

Table 5: Performance without additional context information. The best values in each row are **marked**.

- As expected, our ensemble model outperforms a single text representation in both datasets for most metrics.

As CT-FAN 21 does not provide any contextual information, we are limited to compare our different text representations here. Due to the fact, that this task is a multi-class problem, we did not use our majority ensemble here as voting draws might occur:

- Our proposed model performs best here using extractive summarizations and would have been ranked in sixth place in the official runs (Nakov et al., 2021b), indicating also a competitive performance for a multi-class problem.[7]
- The difference between a basic BERT model and a fine-tuned variant is a lot more pronounced here. We suspect, that fine-tuning is more crucial for this problem due to the difference in data used compared to, e.g. Politifact for FakeNewsNet.
- For this dataset, abstractive summarization seems to perform poorly, resulting in worse performance than a normal BERT model.

5.4. Component Analysis

First, we investigated whether our model architecture improves the classification measurably. A one-sided Wilcoxon test between BERT and CMTR-BERT w/o context, with significant results for both Politifact ($Z = 46, p < .05$) and GossipCop ($Z = 45, p < .05$), indicates a performance gain. It remains, however unclear, which parts of the model contribute to this difference the most, so we conducted an additional ablation study.

[7]The best performing system *NoFake* (Kumari, 2021) used additional context and training. If we ignore system runs which used additional information in this specific task, our system ranks third, with the best one being from Martinez-Rico et al. (2021), which also uses Transformers.

Dataset	Metric	NoFake	NLP & IR@UNED	BERT-baseline	BERT	CMTR-BERT O w/o context	CMTR-BERT A w/o context	CMTR-BERT E w/o context
CT-FAN 21	Accuracy	**0.853**	0.528	0.316	0.453	0.461	0.441	0.480
	Precision	-	-	0.128	0.424	0.446	0.422	**0.467**
	Recall	-	-	0.251	0.402	0.414	0.384	**0.435**
	F1-macro	**0.838**	0.468	0.124	0.395	0.406	0.373	0.428

Table 6: Performance of CMTR-BERT compared to submissions of 2021's CheckThatLab. The best values in each row are **marked**.

The first part we investigated is the hierarchical input representation. We compared *BERT & CMTR-BERT O w/o context*, which resulted in a significant result for CT-FAN 21 ($Z = 46, p < .05$), but not for FakeNewsNet's Politifact ($Z = 35, p = 0.25$) and GossipCop ($Z = 38, p = 0.16$) domain.

Furthermore, we are interested in measurable differences between our textual representations, as there seems to be little difference for FakeNewsNet but considerable divergences for CT-FAN 21. For all datasets, we applied a three factored Friedman test, which resulted in significant results for GossipCop ($\chi^2 = 8.60, p < 0.05$) & CT-FAN 21 ($\chi^2 = 12.60, p < 0.01$), but not for Politifact ($\chi^2 = 2.92, p = 0.23$). Post-hoc tests, along with the data seen in Table 5 and Table 6, suggest that models trained on abstractive summarization perform worse. However, for CT-FAN 21 prior extractive summarization improves performance.

Another major part of the model is its ensemble structure. We investigate this by comparing all aforementioned textual representations with the ensemble within a four factored Friedman test. While not significant for Politifact ($\chi^2 = 5.42, p = 0.14$), the p-value is lower than before, which might suggest that with more trials an effect is measurable here. The same test on GossipCop is highly significant ($\chi^2 = 1836, p < 0.01$). After running post-hoc tests, we can confirm the significant difference between CMTR-BERT O and CMTR-BERT A ($p < 0.01$) found in the paragraph above. Additionally, there are significant effects between CMTR-BERT - CMTR-BERT A ($p < 0.01$) and CMTR-BERT - CMTR-BERT E ($p < 0.05$). Contextualizing these results with Table 5 we can deduce that CMTR-BERT performs significantly better than with abstractive or extractive summarizations alone.

To analyse the influence of contextual data, we compare *CMTR-BERT O*, *CMTR-BERT A*, *CMTR-BERT E & CMTR-BERT* with their *w/o content* counterparts. The results are both highly significant for Politifact ($Z = 11, p < 0.01$) and GossipCop ($Z = 0, p < 0.01$). After further investigating the results using one-sided Wilcoxon tests for Politifact ($Z = 769, p < 0.01$) and GossipCop ($Z = 820, p < 0.01$), it becomes apparent that the model performs significantly better with context data than without.

Additionally, we also investigated which contextual dimension has the most impact on classification performance for FakeNewsNet. For both datasets, the *Source URL* seems to be the most important factor, with the *Tweets* being the second most valuable.

Lastly, we investigated whether it is feasible to train on one domain of FakeNewsNet and use the classifier on the other one. Here, however, fine-tuning is actually hurting the performance and context-based systems perform better. You can find the corresponding additional tables and calculations on our GitHub repository.

6. Conclusion

We have presented an ensemble approach for fake news detection that is based on the powerful paradigm of transformer-based embeddings and utilizes text summarization as the main text transformation step before classifying a document.

CMTR-BERT is able to achieve state-of-the-art results for a common fake news benchmark collection and provides competitive results for a second one. Our results indicate a measurable advantage of our architecture in comparison to a standard BERT model.

Furthermore, our results emphasise the importance of context information for fake news detection once more. Not only do all context-aware systems perform substantially better, it also seems feasible to not use content information at all. While each text representation individually considered does not consistently bring advantages, the combination of all three seems to be the key. It remains unclear to what extent our input transformation influences the performance, as the results here are not decisive.

Overall, our results suggest that this is a worthwhile direction of work, and we plan to explore this further. Specifically, we are interested in using human summarizations, as arguably automatic summarization techniques are not on par with them yet and might negatively influence the system. Ideally, there would be an additional dataset with context information as well as aforementioned corresponding summarizations. This might also be done with a small subset of, e.g. FakeNewsNet which gets manually annotated.

We would also like to see our approach used with other datasets and different context information to get a deeper understanding into which type of information is key for effective fake news detection. Future work will also include the utilization of more recent transformer models and the exploration of other text classification tasks.

Acknowledgements

This work was supported by the project *COURAGE: A Social Media Companion Safeguarding and Educating Students* funded by the Volkswagen Foundation, grant number 95564.

7. Bibliographical References

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

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Bahad, P., Saxena, P., and Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science*, 165:74–82. Publisher: Elsevier.

Chen, T., Li, X., Yin, H., and Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 40–52. Springer.

Cui, L., Wang, S., and Lee, D. (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, pages 41–48, New York, NY, USA, August. Association for Computing Machinery.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Ding, J., Hu, Y., and Chang, H. (2020a). BERT-Based Mental Model, a Better Fake News Detector. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, pages 396–400. Association for Computing Machinery, New York, NY, USA, April.

Ding, L., Ding, L., and Sinnott, R. O. (2020b). Fake News Classification of Social Media Through Sentiment Analysis. In Surya Nepal, et al., editors, *Big Data – BigData 2020*, Lecture Notes in Computer Science, pages 52–67, Cham. Springer International Publishing.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychol Sci Public Interest*, 14(1):4–58, January. Publisher: SAGE Publications Inc.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679. Publisher: Elsevier.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., and Zha, H. (2017). Fake news mitigation via point process based intervention. In *International Conference on Machine Learning*, pages 1097–1106. PMLR.

Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175.

Ferrara, E. (2020). What Types of COVID-19 Conspiracies are Populated by Twitter Bots? *CoRR*, abs/2004.09531.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701, December. Publisher: Taylor & Francis.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 02.

Hanselowski, A., Avinesh, P. V. S., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., and Gurevych, I. (2018). A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Hua, J. and Shaw, R. (2020). Corona virus (Covid-19)"infodemic" and emerging issues through a data lens: The case of china. *International journal of en-*

vironmental research and public health, 17(7):2309. Publisher: Multidisciplinary Digital Publishing Institute.

Jurafsky, D. and Martin, J. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd (draft), 29th December 2021 edition. https://web.stanford.edu/~jurafsky/slp3/.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., and Gomez-Rodriguez, M. (2018). Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 324–332, New York, NY, USA, February. Association for Computing Machinery.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Kumari, S. (2021). NoFake at CheckThat! 2021: fake news detection using BERT. *arXiv preprint arXiv:2108.05419*.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563. Publisher: JMLR. org.

Lesce, T. (1990). Scan: Deception detection by scientific content analysis. *Law and Order*, 38(8):3–6.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Li, Q. and Zhou, W. (2020). Connecting the Dots Between Fact Verification and Fake News Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1820–1825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Li, Y., Jiang, B., Shu, K., and Liu, H. (2020). MM-COVID: A Multilingual and Multidimensional Data Repository for CombatingCOVID-19 Fake News. *arXiv:2011.04088 [cs]*, November. arXiv: 2011.04088 version: 1.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2018). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ma, J., Gao, W., and Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada, July. Association for Computational Linguistics.

Magdy, A. and Wanas, N. (2010). Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 103–110.

Marco-Franco, J. E., Pita-Barros, P., Vivas-Orts, D., González-de Julián, S., and Vivas-Consuelo, D. (2021). COVID-19, Fake News, and Vaccines: Should Regulation Be Implemented? *International Journal of Environmental Research and Public Health*, 18(2):744. Publisher: Multidisciplinary Digital Publishing Institute.

Martinez-Rico, J. R., Martínez-Romo, J., and Araujo, L. (2021). NLP&IR@ UNED at CheckThat! 2021: check-worthiness estimation and fake news detection using transformer models. *2021 Working Notes of CLEF - Conference and Labs of the Evaluation Forum*.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt et al., editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Miller, D. (2019). Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Mitra, T. and Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth international AAAI conference on web and social media*.

Moratanch, N. and Chitrakala, S. (2016). A survey on abstractive text summarization. In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)*, pages 1–7. IEEE.

Moratanch, N. and Chitrakala, S. (2017). A survey on extractive text summarization. In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–6, January.

Mulyar, A., Schumacher, E., Rouhizadeh, M., and Dredze, M. (2020). Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models. *arXiv:1910.13664 [cs]*, September. arXiv: 1910.13664.

Nakov, P., Corney, D. P. A., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., and Martino, G. D. S. (2021a). Automated fact-checking for assisting human fact-checkers. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence,*

IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, pages 4551–4558. ijcai.org.

Nakov, P., Martino, G. D. S., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., Nikolov, A., Shahi, G. K., Struß, J. M., and Mandl, T. (2021b). The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In Djoerd Hiemstra, et al., editors, *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 639–649. Springer.

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons.* Princeton University.

Nørregaard, J., Horne, B. D., and Adalı, S. (2019). Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 630–638.

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., and Gipp, B. (2019). Enriching BERT with Knowledge Graph Embeddings for Document Classification. *arXiv:1909.08402 [cs]*, September. arXiv: 1909.08402.

Pan, J. Z., Pavlova, S., Li, C., Li, N., Li, Y., and Liu, J. (2018). Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Papanastasiou, Y. (2020). Fake News Propagation and Detection: A Sequential Model. *Management Science*, 66(5):1826–1846, January. Publisher: INFORMS.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Pennebaker, J., Francis, M., and Booth, R. (2001). *Linguistic inquiry and word count (LIWC)*. Lawerence Erlbaum Associates.

Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408. Publisher: MIT Press.

Rubin, V. L., Conroy, N. J., and Chen, Y. (2015). Towards news verification: Deception detection methods for news discourse. In *Hawaii International Conference on System Sciences*, pages 5–8.

Rubin, V. L. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10. Publisher: Wiley Online Library.

Shahi, G. K., Dirkson, A., and Majchrzak, T. A. (2021a). An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22:100104.

Shahi, G. K., Struß, J. M., and Mandl, T. (2021b). Overview of the CLEF-2021 CheckThat! lab task 3 on fake news detection. *Working Notes of CLEF*.

Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., and Kurzweil, R. (2017). Generating High-Quality and Informative Conversation Responses with Sequence-to-Sequence Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42. Publisher: ACM New York, NY, USA.

Shearer, E. and Mitchell, A. (2021). News use across social media platforms in 2020. Publisher: Pew Research Center.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, Anchorage AK USA, July. ACM.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020a). 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, June. Publisher: Mary Ann Liebert, Inc., publishers.

Shu, K., Mahudeswaran, D., Wang, S., and Liu, H. (2020b). Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. *ICWSM*, 14:626–637, May.

Silverman, C., Strapagiel, L., Shaban, H., Hall, E., and Singer-Vine, J. (2016). Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate. https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis.

Sobhani, P., Mohammad, S., and Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 159–169.

Souma, W., Vodenska, I., and Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46. Publisher: Springer.

Tas, O. and Kiyani, F. (2017). A survey auto-

matic text summarization. *PressAcademia Procedia*, 5(1):205–213.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (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, June. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, USA. Curran Associates Inc.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663. Publisher: IEEE.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Vogel, I. and Jiang, P. (2019). Fake news detection with the new German dataset "GermanFakeNC". In *International Conference on Theory and Practice of Digital Libraries*, pages 288–295. Springer.

Vrij, A. (2005). Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law*, 11(1):3. Publisher: American Psychological Association.

Wang, W. Y. (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, July. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., and Funtowicz, M. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuan, C., Ma, Q., Zhou, W., Han, J., and Hu, S. (2020). Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. (2020a). Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice*, 1(2):1–25, July.

Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020b). ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212. Association for Computing Machinery, New York, NY, USA, October.

Zhou, X., Wu, J., and Zafarani, R. (2020c). SAFE: Similarity-Aware Multi-modal Fake News Detection. In Hady W. Lauw, et al., editors, *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 354–367, Cham. Springer International Publishing.