Automatic Fake News Detection with Pre-trained Transformer Models

Mina Schütz[1,2](✉), Alexander Schindler[2], Melanie Siegel[1], and Kawa Nazemi[1]

[1] Darmstadt University for Applied Sciences, 64295 Darmstadt, Germany
{melanie.siegel,kawa.nazemi}@h-da.de
[2] Austrian Institute of Technology GmbH, 1210 Vienna, Austria
{mina.schuetz,alexander.schindler}@ait.ac.at
http://www.h-da.de
http://www.ait.ac.at

Abstract. The automatic detection of disinformation and misinformation has gained attention during the last years, since fake news has a critical impact on democracy, society, and journalism and digital literacy. In this paper, we present a binary content-based classification approach for detecting fake news automatically, with several recently published pre-trained language models based on the Transformer architecture. The experiments were conducted on the FakeNewsNet dataset with XLNet, BERT, RoBERTa, DistilBERT, and ALBERT and various combinations of hyperparameters. Different preprocessing steps were carried out with only using the body text, the titles and a concatenation of both. It is concluded that Transformers are a promising approach to detect fake news, since they achieve notable results, even without using a large dataset. Our main contribution is the enhancement of fake news' detection accuracy through different models and parametrizations with a reproducible result examination through the conducted experiments. The evaluation shows that already short texts are enough to attain 85% accuracy on the test set. Using the body text and a concatenation of both reach up to 87% accuracy. Lastly, we show that various preprocessing steps, such as removing outliers, do not have a significant impact on the models prediction output.

Keywords: Fake news · Fake news detection · Transformer · BERT · Pre-trained language model

1 Introduction

The increased usage of social media and news consumption over the internet has helped in spreading fake news. Therefore, fake news has already had effects on political processes [17]. Even though a clear definition of the term fake news is not yet decided, automatic fake news detection with machine learning techniques can help users to identify signs of deception easier [22]. On the contrary, expert-based fact-checking needs many resources and is time-consuming, therefore it is an important goal to develop automatic machine learning algorithms

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12667, pp. 627–641, 2021.
https://doi.org/10.1007/978-3-030-68787-8_45

[11]. For content-based fake news detection, the Transformer models seem to be a promising approach, which were introduced by Vaswani et al. [40]. Research, using transfer learning, has already outperformed methods, based on state-of-the-art results, in numerous NLP downstream tasks [8,18,21,42]. Due to insufficient comparative results, the goal of this work is to show to which extent pre-trained language models are useful for content-based fake news detection and whether they gain promising results in predicting the classification of body texts and titles of news articles.

The paper is structured as follows. In Sect. 2, we give a brief overview of the definition of fake news. Afterwards, in Sect. 3, we discuss the previous work and state-of-the-art language models, followed by the related work in content-based fake news detection via Transformers. Section 4 describes the methodology, data and preprocessing steps. We illustrate the conducted experiments, results and evaluations in Sect. 5 and 6. We conclude our paper with a summary of the main contributions and give suggestions for future work.

2 Fake News

Usually scientific publications differ in definitions for the term fake news [43]. The intention to create such false news pieces has various reasons. On the one hand, there is a financial motive, where people and companies gain revenue through spreading false articles and generating clicks [15]. Intentions can also be malicious, if the news article is only created to hurt one or more individuals, manipulate public opinion, or spread an ideology [33]. Rubin et al. [29] state that fake articles "[...] may be misleading or even harmful, especially when they are disconnected from their original sources and context." However, Mahid et al. [22] defined it narrower: "Fake news is a news articles that is intentionally and verifiable false." This definition is used by several other publications [7,32]. Some studies have broader definitions of fake news, as Sharma et al. [33]: "A news article or message published and propagated through media, carrying false information regardless the means and motives behind it." This definition integrates fabricated as well as misleading content. Depending on intention and factuality there are many similar concepts of news that fall under the fake news definition: Misinformation (unintentional) [3], disinformation (intentional) [5], satire [17], fabrications [15], clickbait [5], hoaxes [29], rumors [24], propaganda [5]. In this work we define fake news as the following: Fake news is an article which propagates a distorted view of the real world regardless of the intention behind it.

3 State-of-the-Art

There are many promising approaches to detect fake news during the last years. Accordingly, the methods vary from simple (e.g. Naïve Bayes) to more complex methods (e.g. CNN, RNN, and LSTM) resulting in a wide range of prediction outcomes. Several surveys have been published, that give an overview over methods, such as social-context based, content-based and knowledge-based as well as hybrid detection approaches [24,26,33,43]. However, when focusing on content-based classification, Transformer-based models were recently introduced, having results exceeding or outperforming in a wide range of research tasks [39]. The pre-trained models can be fine-tuned with a dataset of a specific NLP task, where the available corpora are often small [39]. Additionally, word embeddings are a significant improvement for language modeling [16]. Embeddings create a numeric representation of the input with additional positional embeddings to represent the position of tokens in a sentence [12]. The standard Transformer architecture consists of an encoder and decoder with self-attention, to capture the context of a word in a sentence [39].

3.1 Transformer and Language Models

There have been several language models already been made publicly available. **ELMo** (Embeddings from Language Models) is bilateral and a deep contextualized word representation, developed to improve word embeddings [25] and to predict the next word in a sentence [10]. Also, ELMo uses both encoder and decoder of the Transformer architecture [13]. However, **ULMFiT** (Universal Language Model Fine-Tuning) uses a multi-layered BiLSTM without the attention-mechanism [10]. Howard and Ruder [14] pre-trained ULMFiT on general data and fine-tuned it on a downstream task, which works well with limited labeled data in multiple languages [14]. **GPT** (Generative Pre-Training Transformer) on the other hand is a multi-layered Transformer decoder [10], which is an extension of the architecture of ELMo and ULMFiT without the LSTM model [27]. However, the second GPT model (GPT-2) has more parameters than the original (over 1.5 billion), which was only released with a smaller version of parameters to the public [12]. Recently the third version (GPT-3) was released [4]. **GROVER** is a semi-supervised left-to-right decoder, which is trained on human-written text.

However, **BERT** is one of the latest innovations in machine learning techniques for NLP and was developed by Google in 2019 [8]. The Transformer "[...] is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" [8]. For pre-training, Devlin et al. [8] constructed a dataset that has over 800 million words. BERT only uses the encoder of the Transformer structure [16] and the WordPiece embedding model, which has around 30,000 tokens in its vocabulary [8]. The embedding is a combination of multiple tokens, so that fewer vocabulary errors occur [28]. Devlin et al. [8] used two pre-training models. The first one is called Masked Language Modeling (MLM). This means that during training 15% of a sentence is not represented by the original tokens and instead replaced with a "[MASK]" token, so that the model can learn the whole context of the sequence [13]. MLM is used because the masked word would see itself during pre-training, due to the bidirectionality of the model [8]. The second pre-training model is called Next Sentence Prediction (NSP), where the model takes a sentence pair as an input [13].

Liu et al. [21] stated that the BERT model is undertrained and therefore created **RoBERTa** (A Robustly Optimized BERT). Their model has been trained on additional data with a longer period of time and dynamic pre-training during MLM. They gained state-of-the-art results in GLUE, RACE, and SQuAD and improved the results of the original BERT [21]. After the release of RoBERTa, the authors of study [18] published "A Lite BERT" (**ALBERT**) version of BERT. They criticized, that the original BERT has limitations regarding the GPU and TPU memory. The training time for the original model is quite long and therefore they set their goal to reduce parameters in BERT. ALBERT gained state-of-the-art results in the following natural language processing tasks: GLUE, RACE and SQuAD. Their results were even better than the before mentioned RoBERTa, despite having less parameters than the original BERT version [18]. The distilled version of BERT (**DistilBERT**) is another newly developed model with a reduction of the original model size by 40%. The model is 60% faster than the original BERT, which makes it cheaper, while still gaining similar results as BERT [30]. However, **XLNet** uses autoregressive language modeling and outperforms BERT on 20 NLP tasks, such as question answering, natural language inference, and sentiment analysis. Yang et al. [42] stated that BERT has problems with the masking in pre-training and fine-tuning and therefore used a different approach to gain better results. They also used two streams for the attention instead of only one [42].

3.2 Related Work

A few studies have rather applied stance detection than classification of fakeness in an article to provide new information about false articles. Jwa et al. [16] focused on the stance between headlines and texts of articles with the FakeNewsChallenge (FNC-1)[1] dataset. Stance detection describes, whether the text is in favor or against a given object. Jwa et al. [16] tested two approaches with BERT. For the first model they only changed the loss function during fine-tuning, whereas for the second model additional news data was gathered for pre-training. Also, Dulhanty et al. [9] used the FNC-1 dataset, but tested it with RoBERTa. Slovikovskaya [37] used the same dataset but added additional data for stance detection. The author used BERT, XLNet and RoBERTa, whereas the latter gained the best result. Similarly to Jwa et al. [16], Soleimani et al. [38] created two BERT models for evidence retrieval and claim verification based on the data of the FEVER[2] challenge. Another approach on the relation between two titles of fake news was proposed by Yang et al. [41]. They used the data by the WSDM 2019 Classification Challenge on Kaggle[3] with titles in Mandarin.

[1] https://github.com/FakeNewsChallenge.

[2] https://fever.ai/.

[3] https://www.kaggle.com/c/fake-news-pair-classification-challenge.

Regarding binary classification, Mao and Liu [23] presented an approach on the 2019 FACT challenge[4] with Spanish data. The data was labeled in *fact* and *counterfact*. The authors said, that their model was overfitting, hence they only had an accuracy of 0.622 as a result. Levi et al. [19] studied the differences between titles and body text of fake news and satire with BERT as a model. Rodriguez and Iglesias [28] compared BERT to two other neural networks with a binary fake news classification. They used the Getting Real About Fake News[5] dataset with additional real news articles. However, Aggarwal et al. [1] tested XGboost, CNN, and BERT with the NewsFN dataset, which is very well balanced into fake and real articles. Their best result was 97.021% Accuracy with the BERT-base-uncased version.

Liu et al. [20] did a multi-classification on short statements with BERT and had an accuracy of 41.58% with additional metadata and 34.51% with statements alone. Antoun et al. [2] used XLNet, RoBERTa and BERT with a dataset from the QICC competition[6] for a binary classification of fake news. Their best model (XLNet) gained an F1-score of 98% accuracy. The second task was a news domain detection, split into six classes: Politics, Business, Sports, Entertainment, Technology, and Education. For this task they used several more models than only the Transformers. RoBERTa gained 94% accuracy, whereas a Bi-LSTM with attention had the same result but an overall better performance. The model was based on word embeddings of ELMo. It has to be mentioned though, that the used dataset only contained 432 articles in total. However, Cruz et al. [6] created a dataset for binary fake news classification for the Filipino language. Additionally, they looked into generalizability across different domains, the influence of pretraining on language models and the effect of attention heads on the prediction output. They used ULMFiT, BERT, and GPT-2 for their experiment, whereas GPT-2 gained the best results with multi-tasking attention heads (96.28% accuracy). The study by Schwarz et al. [31] explored embeddings of multi-lingual Transformers as a framework to detect fake news.

4 Methodology

For this work we used the FakeNewsNet [34–36] dataset, which provides news articles that have a binary classification (*fake* or *real*) and is automatically updated. Since this work presents a content-based approach, only the body text and titles from the dataset were used. As a ground truth Shu et al. [34] used the fact-checking websites PolitiFact and GossipCop. In this work the following Transformer models were used for the experiments: BERT, RoBERTa, ALBERT, DistilBERT, and XLNet.

[4] https://www.fing.edu.uy/inco/grupos/pln/fact/.

[5] https://www.kaggle.com/mrisdal/fake-news.

[6] https://www.hbku.edu.qa/en/qicc.

4.1 Data Distribution

At the time of downloading the data, the set contained 21,658 news articles. Since in this work the title and body text are needed, all rows, where one of those features was missing, were deleted. After this process the dataset contained 5,053 fake and 15,998 real articles, which are in total 21,041. The mean length of body text was 3408,728 characters, whereas the titles had a length of 59,106. The longest body text in general contained 100,000 and the title 200 characters. The shortest ones 14 (text) and 2 (title). When comparing fake and real body texts it could be observed that the real body texts mean value is about 300 characters longer, whereas the fake titles are about 7 characters longer than real ones. The cleaned dataset was used for the following preprocessing steps and creation of the different files for the experiments.

4.2 Preprocessing

There were different types of preprocessing steps carried out to test, whether the models have different prediction outcomes based on the article length and other factors. The first step was to delete all titles, which were shorter than 20 and longer than 120 characters. Most of the short titles were rather the website names, the articles were published on. Also, the longer titles were often error messages, which the model should not learn the difference of fake and real articles on. This was discovered by going through a sub-sample of titles manually. The same process was used for the body texts, since many short texts were extracted error messages instead of actual content. Therefore, all body texts with more than 10,000 and less than 1,000 characters were deleted. After going through the dataset manually, it stood out that many of the articles that have been labeled as real were transcripts. Transcripts are conversations or interviews, often from politicians and contain mostly spoken word. Since the dataset contains more real articles than fake ones, it could be a problem for the model to distinguish spoken language and written articles. Based on this examination the second preprocessing step was to remove all articles with more than nineteen colons. The transcripts usually started around 20 colons per body text. All articles contained HTML strings, because the dataset was retrieved by a crawler. It stood out that many fake articles contained *[edit]*, which was the only string that was deleted from the dataset, since there are fewer fake articles and the models should not learn the differences between fake and real based only on this. The last preprocessing step included deleting all non-ASCII signs and digits, to see if this makes any difference when evaluating the experiments. Additionally, the newline tags were deleted for all preprocessed files.

In Table 1 all files, with the various preprocessing steps, are shown. They were split in text only, titles only and the concatenation of titles and text. Depending on the preprocessing steps the smallest dataset has a more balanced distribution than the original data: 3,358 fake and 8,586 real articles, which are in total 11,944. The longest text would be 9,919 and shortest 926 characters long. The titles from 20 up to 120 characters.

Table 1. Preprocessed files.

File no.	Type	Length	Transcript	Edit	ASCII/Digits	Dataset size
1	Text	Yes	Yes	Yes	Yes	11,944
2	Text	Yes	Yes	Yes	No	11,944
3	Text	Yes	Yes	No	No	11,944
4	Text	Yes	No	No	No	12,172
5	Text	No	No	No	No	21,041
6	Title	No	No	No	No	21,041
7	Title	Yes	No	No	No	12,172
8	Title	Yes	No	No	Yes	12,172
9	Both	No	No	No	No	21,041
10	Both	Yes	No	No	No	15,355
11	Both	Yes	Yes	No	No	15,103
12	Both	Yes	Yes	Yes	Yes	15,103

The dataset was split in training set (80%) and test set (20%), which was carried out with a stratified split to balance the classes in both sets. During the implementation of the models, the training set was additionally split into training and validation (10% from training). Depending on the file size and preprocessing steps the classes are more or less balanced (less for the largest dataset). Other standard preprocessing methods, such as removing stop words, punctuation, lemmatization and stemming were not carried out, because the Transformer models need all tokens to understand the context of the sentence. Therefore, valuable information goes missing if the words are cut, deleted or the sentence structure is altered.

5 Experiments

The experiments in this work were carried out five different Transformer models with the PyTorch version of the HuggingFace Transformers library[7] on a GeForce

Table 2. Used Transformer models for the experiments.

Model	Layers	Hidden States	Attention Heads	Parameter
BERT-BASE-CASED	12	768	12	110 Million
ROBERTA-BASE	12	768	12	125 Million
ALBERT-BASE-V2	12	768	12	11 Million
XLNET-BASE	12	768	12	110 Million
DISTILBERT-BASE-CASED	6	768	12	65 Million

[7] https://github.com/huggingface/transformers.

GTX TITAN X as GPU. The used models are also shown in Table 2. They all have the same count of layers except DistilBERT, which is the distilled version of the original BERT model and therefore only has 6 layers instead of 12.

The first experiments were conducted with file no. 1, which is completely preprocessed and only contains body text. This was also used to figure out valuable hyperparameters for the following experiments. The batch size and maximum sequence length was used as recommended by Devlin et al. [8]. After testing different batch sizes, learning rates, warm-up steps, epochs and sequence lengths, the best hyperparameters were used for the other experiments. In this work, we tested the experiments with more than the usual maximum of 5 epochs to gain insight, whether the loss curves change with more epochs and influence the prediction outcomes. First, the different preprocessed body text files were run through the BERT-base-cased model, then the files containing only titles and then the combination of titles and body text. After looking at the results of the BERT-model, the same hyperparameters were used for other Transformer models.

6 Results

As mentioned before, the experiments were split in only body text, only titles and a concatenation of titles and text of the articles. Also, the cased models were used with no lower-casing during tokenization. For only body text, documented in Table 3, the highest accuracy gained was 0.87. For each experiment the best model is highlighted in bold. The first experiment however shows that the models do not work well with a high learning rate, when predicting the labels on this dataset. The best results are gained with RoBERTa, however accuracy values with XLNet are similar. The results show that all models have a good prediction with different hyperparemeters.

For comparison reasons, the maximum sequence lengths have not been changed over 512 tokens, even when the model had a higher sequence length available. Additionally, the results (Table 3) show that the different preprocessing steps have no major impact on the prediction. Although file 5, which is not preprocessed at all, gains the best results with all models, the accuracy and loss are not significantly apart from other experiments. This shows that deleting transcripts, which could be a learned bias during training, has no further impact on the outcome of the models.

However, the results of the titles (Table 4) have a lower accuracy result and higher loss than using the body texts. The highest accuracy was 0.85. Again, RoBERTa and XLNet gained the best results, respectively and show the same behavior as with the body texts and preprocessing.

Lastly, in Table 5 the results of the concatenation of titles and body texts are shown. Again, the highest accuracy value is 0.87, but for this type of experiments the best models were DistilBERT and XLNet. The results are only slightly different for each of the models. Also, the preprocessing did not change the predictions significantly. It is notably though, that the experiments gain the overall best results out of the three different types.

Table 3. Body text only - experiment results.

Model	File	Epoch	Batch	LR	Warm-Up	Max Seq	Val Acc	Test Acc	Loss
BERT	1	5	6	5e−5	0	512	0.82	0.83	0.48
ROBERTA	1	5	6	5e−5	0	512	0.76	0.76	0.56
ALBERT	1	5	6	5e−5	0	512	0.75	0.72	0.60
XLNET	1	5	6	5e−5	0	512	0.77	0.81	0.54
DISTILBERT	1	5	6	5e−5	0	512	**0.85**	**0.86**	**0.19**
BERT	1	15	16	2e−5	100	256	0.85	0.86	0.01
ROBERTA	1	15	16	2e−5	100	256	0.86	0.87	0.02
ALBERT	1	15	16	2e−5	100	256	0.83	0.83	0.03
XLNET	1	15	16	2e−5	100	256	**0.86**	**0.87**	**0.01**
DISTILBERT	1	15	16	2e−5	100	256	0.85	0.86	0.01
BERT	1	10	6	2e−5	100	512	0.85	0.86	0.01
ROBERTA	1	10	6	2e−5	100	512	**0.86**	**0.87**	**0.05**
ALBERT	1	10	6	2e−5	100	512	0.82	0.83	0.17
XLNET	1	10	6	2e−5	100	512	0.86	0.85	0.04
DISTILBERT	1	10	6	2e−5	100	512	0.85	0.85	0.02
BERT	1	10	16	2e−5	0	256	0.85	0.85	0.02
ROBERTA	1	10	16	2e−5	0	256	**0.86**	**0.87**	**0.03**
ALBERT	1	10	16	2e−5	0	256	0.82	0.83	0.07
XLNET	1	10	16	2e−5	0	256	0.86	0.85	0.04
DISTILBERT	1	10	16	2e−5	0	256	0.85	0.87	0.02
BERT	2	10	16	2e−5	0	256	0.84	0.85	0.02
ROBERTA	2	10	16	2e−5	0	256	**0.86**	**0.87**	**0.03**
ALBERT	2	10	16	2e−5	0	256	0.82	0.82	0.14
XLNET	2	10	16	2e−5	0	256	0.86	0.86	0.03
DISTILBERT	2	10	16	2e−5	0	256	0.86	0.85	0.02
BERT	2	10	6	2e−5	0	512	0.85	0.87	0.02
ROBERTA	2	10	6	2e−5	0	512	**0.86**	**0.87**	**0.05**
ALBERT	2	10	6	1e−5	0	512	0.83	0.84	0.08
XLNET	2	10	6	2e−5	0	512	0.83	0.85	0.08
DISTILBERT	2	10	6	2e−5	0	512	0.86	0.85	0.02
BERT	4	10	16	2e−5	0	256	0.84	0.86	0.02
ROBERTA	4	10	16	2e−5	0	256	0.87	0.87	0.04
ALBERT	4	10	16	2e−5	0	256	0.83	0.83	0.08
XLNET	4	10	16	2e−5	0	256	**0.87**	**0.87**	**0.03**
DISTILBERT	4	10	16	2e−5	0	256	0.86	0.85	0.02
BERT	5	10	16	2e−5	0	256	0.87	0.86	0.08
ROBERTA	5	10	16	2e−5	0	256	**0.88**	**0.87**	**0.10**
ALBERT	5	10	16	2e−5	0	256	0.85	0.84	0.17
XLNET	5	10	16	2e−5	0	256	0.87	0.86	0.09
DISTILBERT	5	10	16	2e−5	0	256	0.86	0.86	0.09
BERT	1	10	16	2e−5	0	256	0.87	0.86	0.08
ROBERTA	1	10	16	2e−5	50	256	**0.86**	**0.87**	**0.04**
ALBERT	1	10	16	2e−5	50	256	0.82	0.83	0.05
XLNET	1	10	16	2e−5	50	256	0.86	0.86	0.02
DISTILBERT	1	10	16	2e−5	50	256	0.84	0.85	0.02

Table 4. Title only - experiment results.

Model	File	Epoch	Batch	LR	Warm-up	Max Seq	Val Acc	Test Acc	Loss
BERT	6	5	32	2e−5	0	128	0.84	0.83	0.15
ROBERTA	6	5	32	2e−5	0	128	0.84	0.85	0.26
ALBERT	6	5	32	2e−5	0	128	0.82	0.82	0.01
XLNET	6	5	32	2e−5	0	128	**0.85**	**0.85**	**0.02**
DISTILBERT	6	5	32	2e−5	0	128	0.84	0.84	0.17
BERT	7	5	32	2e−5	0	128	0.84	0.84	0.11
ROBERTA	7	5	32	2e−5	0	128	**0.85**	**0.85**	**0.25**
ALBERT	7	5	32	2e−5	0	128	0.81	0.81	0.20
XLNET	7	5	32	2e−5	0	128	0.84	0.83	0.26
DISTILBERT	7	5	32	2e−5	0	128	0.83	0.85	0.16
BERT	8	5	32	2e−5	0	128	0.83	0.83	0.10
ROBERTA	8	5	32	2e−5	0	128	**0.86**	**0.85**	**0.26**
ALBERT	8	5	32	2e−5	0	128	0.82	0.81	0.18
XLNET	8	5	32	2e−5	0	128	0.83	0.84	0.23
DISTILBERT	8	5	32	2e−5	0	128	0.83	0.83	0.16
BERT	6	10	32	2e−5	0	128	0.84	0.84	0.08
ROBERTA	6	10	32	2e−5	0	128	**0.84**	**0.85**	**0.16**
ALBERT	6	10	32	2e−5	0	128	0.81	0.81	0.08
XLNET	6	10	32	2e−5	0	128	0.85	0.84	0.23
DISTILBERT	6	10	32	2e−5	0	128	0.84	0.84	0.09
BERT	6	30	32	2e−5	0	128	0.84	0.83	0.07
ROBERTA	6	30	32	2e−5	0	128	0.85	0.85	0.08
ALBERT	6	30	32	2e−5	0	128	0.82	0.81	0.06
XLNET	6	30	32	2e−5	0	128	**0.85**	**0.85**	**0.07**
DISTILBERT	6	30	32	2e−5	0	128	0.84	0.85	0.06

To compare these results, we applied some of the methods, which were used in the original paper. The authors of the dataset [34] split the data in PolitiFact and GossipCop articles separately. The best result was 0.723 accuracy with a CNN for GossipCop articles and 0.642 accuracy for PolitiFact with Logistic Regression. For our evaluation we used Gaussian Naive Bayes, Support Vector Machine and Logistic Regression with One Hot Encoding and the default parameters of ScikitLearn, as the original paper has done. We used our former preprocessed files for both types, because we also had to apply standard preprocessing, such as: stemming, lemmatization, removing stop-words and punctuation.

As shown in Table 6, the standard supervised methods seem to have problems with either false positive (FP) or false negative (FN) classification results. The only model that has results closely to the Transformer models are the SVM and LR, but seem to train only on one class. On the contrary it can be seen that

Table 5. Title and body text - experiment results.

Model	File	Epoch	Batch	LR	Warm-up	Max Seq	Val Acc	Test Acc	Loss
BERT	9	10	16	2e−5	0	256	0.87	0.87	0.07
ROBERTA	9	10	16	2e−5	0	256	0.88	0.87	0.09
ALBERT	9	10	16	2e−5	0	256	0.86	0.85	0.10
XLNET	9	10	16	2e−5	0	256	0.87	0.86	0.09
DISTILBERT	9	10	16	2e−5	0	256	**0.87**	**0.87**	**0.08**
BERT	10	10	16	2e−5	0	256	0.87	0.86	0.08
ROBERTA	10	10	16	2e−5	0	256	0.87	0.87	0.09
ALBERT	10	10	16	2e−5	0	256	0.84	0.85	0.09
XLNET	10	10	16	2e−5	0	256	**0.87**	**0.87**	**0.08**
DISTILBERT	10	10	16	2e−5	0	256	0.86	0.86	0.08
BERT	11	10	16	2e−5	0	256	0.87	0.86	0.09
ROBERTA	11	10	16	2e−5	0	256	0.86	0.87	0.09
ALBERT	11	10	16	2e−5	0	256	0.83	0.82	0.13
XLNET	11	10	16	2e−5	0	256	**0.87**	**0.87**	**0.09**
DISTILBERT	11	10	16	2e−5	0	256	0.86	0.86	0.08
BERT	13	10	16	2e−5	0	256	0.87	0.86	0.08
ROBERTA	13	10	16	2e−5	0	256	0.87	0.87	0.10
ALBERT	13	10	16	2e−5	0	256	0.85	0.84	0.09
XLNET	13	10	16	2e−5	0	256	0.87	0.86	0.09
DISTILBERT	13	10	16	2e−5	0	256	**0.87**	**0.86**	**0.08**

Table 6. Comparison of transformer models against a baseline.

Model	File	Type	Accuracy	TN	FP	FN	TP
NB	1	Text	0.299	89	1628	45	627
SVM	1	Text	0.713	1688	29	656	16
LR	1	Text	0.718	1716	1	671	1
XLNET	1	Text	0.86	**1572**	**145**	**201**	**471**
NB	8	Title	0.297	77	1678	32	648
SVM	8	Title	0.707	1711	44	669	11
LR	8	Title	0.720	1753	2	679	1
ROBERTA	8	Title	0.85	**1598**	**157**	**222**	**458**

two experiments[8] with the Transformer models have a more balanced confusion matrix, even though one class has more articles in the dataset. This shows, that those models gain better results overall.

[8] XLNet - epochs: 5, batch: 32, LR: 2e−5, warm-up: 0, max seq: 128/RoBERTa: epochs: 10, batch: 6, LR: 2e−5, warm-up: 0, max seq: 512.

Table 7. Sensitivity specific metrics for all models.

Model	File	Type	Precision	Recall	F1
BERT	1	Text	0.84	0.81	0.81
RoBERTa	1	Text	**0.84**	**0.82**	**0.82**
ALBERT	1	Text	0.79	0.78	0.77
DISTILBERT	1	Text	0.84	0.81	0.81
XLNET	1	Text	0.83	0.80	0.80
NB	1	Text	0.47	0.49	0.42
SVM	1	Text	0.53	0.50	0.04
LR	1	Text	0.60	0.50	0.002

Lastly, in Table 7 sensitivity metrics are compared with one experiment on the body text[9]. For each metric the macro-average was chosen, regarding the imbalance of the classes, which shows that Transformers are a better solution for this dataset.

7 Conclusion and Future Work

The results of this work show that a content-based approach can gain promising results for detecting fake news, even without setting hand-engineered features and only titles. Although, literature has shown that some approaches still have better results than the Transformer models. The results of this work are comparable with the current state-of-the-art fake news detection approaches, especially in the field of the newly invented Transformer architectures. Almost all experiments, after the fine-tuning of the hyperparameters, had results over 80% accuracy in the validation and test set without overfitting the data. Therefore, this work shows that Transformer models can also detect fake news based on short statements as well as complete articles. Fake news detection is still underrepresented in the research process. Especially automatic detection, without human intervention, is an open research issue. An important factor for further research is to explore methods of explainable artificial intelligence, to help understanding the difference in fake news concepts and to gain insights into the models and which words have to highest impact to predict the fake and real classes as well as the high accuracy for short titles of news articles and the influence of removing spoken language.

Acknowledgments. This is a revised and enhanced work based on a masterthesis conducted at the Darmstadt University of Applied Sciences in collaboration with the Austrian Institute of Technology (AIT). The Darmstadt University supported this work with the research in Information Science (https://sis.h-da.de/) and the Research Group on Human-Computer Interaction and Visual Analytics (https://vis.h-da.de/).

[9] Hyperparameters - epochs: 10, batch size: 16, LR: 2e−5, warm-up: 0, max. seq: 256.

References

1. Aggarwal, A., Chauhan, A., Kumar, D., Mittal, M., Verma, S.: Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Trans. Scalable Inf. Syst. Online First (2020). https://doi.org/10.4108/eai.13-7-2018.163973
2. Antoun, W., Baly, F., Achour, R., Hussein, A., Hajj, H.: State of the art models for fake news detection tasks. In: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), pp. 519–524 (2020)
3. Bara, G., Backfried, G., Thomas-Aniola, D.: Fake or fact? Theoretical and practical aspects of fake news. In: Bossé, É., Rogova, G.L. (eds.) Information Quality in Information Fusion and Decision Making. IFDS, pp. 181–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03643-0_9
4. Brown, T.B., et al.: Language models are few-shot learners (2020)
5. Campan, A., Cuzzocrea, A., Truta, T.M.: Fighting fake news spread in online social networks: actual trends and future research directions. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 4453–4457 (2017)
6. Cruz, J.C.B., Tan, J.A., Cheng, C.: Localization of fake news detection via multitask transfer learning. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 2596–2604. European Language Resources Association, Marseille, France, May 2020. https://www.aclweb.org/anthology/2020.lrec-1.316
7. Della Vedova, M.L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., de Alfaro, L.: Automatic online fake news detection combining content and social signals. In: 2018 22nd Conference of Open Innovations Association (FRUCT), pp. 272–279 (2018)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
9. Dulhanty, C., Deglint, J.L., Daya, I.B., Wong, A.: Taking a stance on fake news: towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection (2019)
10. Ghelanie, S.: From word embeddings to pretrained language models - a new age in NLP - part 2 (2019). https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-2-e9af9a0bdcd9?gi=bf1f5e22e8e4. Accessed 03 Mar 2020
11. Graves, L.: Understanding the promise and limits of automated fact-checking, February 2018
12. Géron, A.: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edn. O'Reilly Media Inc. (2019)
13. Horev, R.: Bert explained: state of the art language model for NLP, November 2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270. Accessed 05 Nov 2019
14. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 328–339. Association for Computational Linguistics, Melbourne, Australia, July 2018. https://doi.org/10.18653/v1/P18-1031. https://www.aclweb.org/anthology/P18-1031

15. Tandoc Jr., E.C., Lim, Z.W., Ling, R.: Defining “fake news”. Digit. J. **6**(2), 137–153 (2017). https://doi.org/10.1080/21670811.2017.1360143
16. Jwa, H., Oh, D., Park, K., Kang, J., Lim, H.: exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). Appl. Sci. **9**(19), 4062 (2019). https://doi.org/10.3390/app9194062
17. Khan, S.A., Alkawaz, M.H., Zangana, H.M.: The use and abuse of social media for spreading fake news. In: 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), pp. 145–148 (2019)
18. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations (2019)
19. Levi, O., Hosseini, P., Diab, M., Broniatowski, D.: Identifying nuances in fake news vs. satire: using semantic and linguistic cues. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda (2019). https://doi.org/10.18653/v1/d19-5004
20. Liu, C., et al.: A two-stage model based on BERT for short fake news detection. In: Douligeris, C., Karagiannis, D., Apostolou, D. (eds.) KSEM 2019. LNCS (LNAI), vol. 11776, pp. 172–183. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29563-9_17
21. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach (2019)
22. Mahid, Z.I., Manickam, S., Karuppayah, S.: Fake news on social media: brief review on detection techniques. In: 2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA), pp. 1–5 (2018)
23. Mao, J., Liu, W.: Factuality classification using the pre-trained language representation model BERT. In: IberLEF@SEPLN (2019)
24. Oshikawa, R., Qian, J., Wang, W.Y.: A survey on natural language processing for fake news detection (2018)
25. Peters, M.E., et al.: Deep contextualized word representations (2018)
26. Rana, D.P., Agarwal, I., More, A.: A review of techniques to combat the peril of fake news. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1–7 (2018)
27. Rizvi, M.S.Z.: Demystifying BERT: a comprehensive guide to the groundbreaking NLP framework, September 2019. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/. Accessed 05 Nov 2019
28. Rodríguez, À.I., Iglesias, L.L.: Fake news detection using deep learning (2019)
29. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, ASIST 2015. American Society for Information Science, USA (2015). https://doi.org/10.5555/2857070.2857153
30. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter (2019)
31. Schwarz, S., Theóphilo, A., Rocha, A.: EMET: embeddings from multilingual-encoder transformer for fake news detection. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2020, pp. 2777–2781 (2020)
32. Shabani, S., Sokhn, M.: Hybrid machine-crowd approach for fake news detection. In: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), pp. 299–306 (2018)
33. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: a survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol. **37**(4) (2019)

34. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: a data repository with news content, social context and spatialtemporal information for studying fake news on social media (2018)
35. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
36. Shu, K., Wang, S., Liu, H.: Exploiting tri-relationship for fake news detection. arXiv preprint arXiv:1712.07709 (2017)
37. Slovikovskaya, V.: Transfer learning from transformers to fake news challenge stance detection (FNC-1) task (2019)
38. Soleimani, A., Monz, C., Worring, M.: Bert for evidence retrieval and claim verification (2019)
39. Uszkoreit, J.: Transformer: a novel neural network architecture for language understanding, August 2017. https://ai.google-blog.com/2017/08/transformer-novel-neural-network.html. Accessed 01 Dec 2019
40. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
41. Yang, K.C., Niven, T., Kao, H.Y.: Fake news detection as natural language inference (2019)
42. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNET: generalized autoregressive pretraining for language understanding (2019)
43. Zhou, X., Zafarani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surve. (2020). https://doi.org/10.1145/3395046