Distinguishing between Fake News and Satire with Transformers

Jwen Fai Low[a], Benjamin C. M. Fung[a,*], Farkhund Iqbal[b], Shih-Chia Huang[c,*]

[a] *School of Information Studies, McGill University, Montreal, QC, Canada, H3A 1X1*
[b] *College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates*
[c] *Department of Electronic Engineering, National Taipei University of Technology, Taiwan*

The official version of this article is published in Elsevier ESWA in January 2022.

*Corresponding authors
Email addresses: jwen.low@mail.mcgill.ca (Jwen Fai Low), ben.fung@mcgill.ca (Benjamin C. M. Fung), farkhund.iqbal@zu.ac.ae (Farkhund Iqbal), schuang@ntut.edu.tw (Shih-Chia Huang)

Abstract

Indiscriminate elimination of harmful fake news risks destroying satirical news, which can be benign or even beneficial, because both types of news share highly similar textual cues. In this work we applied a recent development in neural network architecture, transformers, to the task of separating satirical news from fake news. Transformers have hitherto not been applied to this specific problem. Our evaluation results on a publicly available and carefully curated dataset show that the performance from a classifier framework built around a DistilBERT architecture performed better than existing machine-learning approaches. Additional improvement over baseline DistilBERT was achieved through the use of non-standard tokenization schemes as well as varying the pre-training and text pre-processing strategies. The improvement over existing approaches stands at 0.0429 (5.2%) in F1 and 0.0522 (6.4%) in accuracy. Further evaluation on two additional datasets shows our framework's ability to generalize across datasets without diminished performance.

Keywords: fake news, satire, sarcasm, deep learning, transformers, BERT, DistilBERT, classification

1. Introduction

Fake news spreads quickly online, more so than verified news (Vosoughi et al., 2018). The uncontrolled spread of fake news can have damaging consequences, as in the case of telecommunication towers being burned due to a conspiracy theory linking COVID-19 with 5G technology (Ahmed et al., 2020). A growing awareness of the danger posed by fake news has been accompanied by calls to keep its spread in check (Xie et al., 2020). However, when combating fake news, satirical news risks being lumped in with fake news as they exhibit highly similar textual cues, especially when contrasted against real news (Horne and Adali, 2017). So great is their similarity that researchers have had success in classifying fake and legitimate news using a classifier trained on a dataset where fake news has been substituted with satirical news (Jeronimo et al., 2019).

We must avoid grouping satirical news together with fake news as they serve very different purposes. *Fake news* is "news articles that are intentionally and verifiably false, and could mislead readers" (Allcott and Gentzkow, 2017). *Satirical news* is "factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behavior that is shameful, corrupt, or otherwise 'bad' " (Golbeck et al., 2018).

Satire serves as a part of the Fifth Estate along with news parody and non-mainstream media sources like columnists and bloggers in critiquing both the people in power and the mainstream news media. This form of criticism wrapped in humor indirectly educates people on political issues and can motivate their participation in the political process, which is crucial in the functioning of a healthy civil society.

Developing the capability to distinguish between fake and satirical news can aid social media companies in crafting more nuanced policies on managing fake news. Such policies are not only helpful in preserving the Fifth Estate but are also more likely to be amenable to government officials concerned over corporate overreach, e.g., European leaders' displeasure over Twitter's banning of Trump (Jennen and Nussbaum, 2021). To contribute to tackling the problem of separating fakes and satires, we present a transformer-based classifier framework. Comprehensive benchmarking on a meticulously constructed dataset of real-world fake and satirical news (Golbeck et al., 2018) demonstrates the viability of our approach and provides information on optimal model configuration. Further benchmarking on two more datasets demonstrates that the performance advantage held by

our approach is generalizable to different datasets and data availability conditions.

The contribution we present in this paper can be summarized as follows:

- Development of a novel framework for classifying satire and fake news that achieves state-of-the-art classification results. Our framework is based on DistilBERT (Sanh et al., 2020), which also makes it the first ever application of a transformer architecture of any kind to the problem of distinguishing between fake and satirical news.

- Pre-training transformer. Differing from many previous works that focus on the fine-tuning stage of transformers, we investigate how different pre-training choices, in terms of the sample sizes of the pre-training process and the dataset used — target dataset, target dataset with sorted sentences, and in-domain external dataset — can ultimately impact the classifier performance.

- Information aggregation method. We examined how two non-standard tokenization schemes using multiple aggregator tokens, one based on article section segmentation and the other based on sentence segmentation, fare against the use of a single aggregator token that is standard to vanilla DistilBERT, vanilla BERT, and BERT variants.

- Text truncation strategy. A text truncation strategy where sentences are sorted according to their informativeness is compared against the naive baseline of using the text as is to address the issue of the entire news article text not fitting within transformers' token limits.

- Extensive empirical evaluation. We conducted extensive testing of our framework on a carefully constructed dataset consisting of real-world fake news articles from a balanced and varied number of sources. The influence that each framework component has on final classification quality is investigated.

- Generalizability evaluation. Additional testing of our framework on two other datasets featuring fake and satirical news was conducted to observe if our approach can be applied to new datasets without degrading its state-of-the-art performance, especially when encountering non-ideal conditions, such as a dataset with more than two classes and a highly imbalanced class distribution.

2. Preliminaries and related work

2.1. Defining satire and fake news

As defined earlier in the introduction, this paper considers *fake news* to be "news articles that are intentionally and verifiably false, and could mislead readers" (Allcott and Gentzkow, 2017). As for *satirical news*, it is news that is "factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behavior that is shameful, corrupt, or otherwise 'bad' " (Golbeck et al., 2018).

Establishing what constitutes fake and satirical news is crucial as there is a lack of consensus on what each of these terms mean even among academic researchers (Golbeck et al., 2018; Tandoc et al., 2018). In this section, we will discuss the importance of distinguishing satire from fake news, the criteria that can be used to distinguish them, and the reasons for adopting these criteria.

Satire needs to be considered as distinct from fake news because satire is a form of social good. There are those who disagree. First Draft News, a non-profit addressing "challenges relating to trust and truth in the digital age" that counts Google News Initiative as its founding member (News, 2020) and has received funding from the Facebook Journalism Project and Twitter, considers satire a form of misinformation that does not intend to cause harm but has the potential to mislead (Wardle, 2017). First Draft News's opposition to satire is countered by much existing evidence on the benefits brought about by satire. Beyond their entertainment value, researchers have found satirical news stories to serve critical functions that include raising awareness among its audience (Becker and Bode, 2018), building greater understanding and sharpening opinions on political issues among its audience (Brewer et al., 2013), galvanizing its audience into political action (Chen et al., 2017), and telling truth to power (Jones, 2017; Mukurunge and Rapitse, 2019). We believe these benefits outweigh whatever potential harm satirical news may do.

We are not alone in our desire to treat satirical news and fake news as two separate categories. For instance, Golbeck et al. (2018) created a fake and satirical news dataset mostly because they wanted to determine if linguistic and thematic features are sufficient to differentiate satire and fake news. Tandoc et al. (2018) personally disagree with satire being lumped into the category of fake news even though their literature survey found that there are many researchers who consider satire as fake news.

If one accepts our premise that satire should be considered as distinct from fake news, the question now turns to finding the criteria that will allow us to label articles as either satire or fake news. These criteria are not only for separating satire and fake news but also for determining whether an article is fit for inclusion in either category. Two major criteria — factuality and intent — that have been used in existing attempts at providing a typology for fake news (Tandoc et al., 2018; Zannettou et al., 2019) serve equally well as criteria for identifying satire and fake news.

Factuality is an important criterion in determining if a news article counts as fake news. In the absence of this criterion, fake news as a term risks being co-opted by "some circles as an attack on legitimate, factually correct stories when people in power simply dislike what they have to say" (Golbeck et al., 2018). Some researchers consider propaganda as a form of fake news and define propaganda simply as news stories created to influence public perceptions for political aims (Tandoc et al., 2018; Zannettou et al., 2019), without a stringent requirement for the presence of factual inaccuracies within the news stories. Our requirement would exclude factually accurate propaganda from being considered as fake news. The requirement for the presence of factual inaccuracy extends to satirical news as well; factually accurate satire cannot exist as it would simply be legitimate news.

Intent is an important criterion for separating satire and fake news. Without intent as a criterion, satire and fake news would be indistinguishable from each other as both types of news share the commonality of being factually inaccurate. With intent as a criterion, satire stands apart from fake news because satire is an attempt to mock bad behavior while fake news intends to mislead its readers. For some researchers such as Zannettou et al. (2019), satirical news articles being oft-disseminated on social media to audiences unaware of their provenance is sufficient basis to condemn them to the category of fake news. But since we consider intent when categorizing news in our paper, satirical news does not count as fake news as people mistaking it for real news is an *unintended* consequence. Similarly, legitimate articles with factual inaccuracies that appear to be products of honest mistakes or unreliable information sources are not considered fake news.

2.2. Automatically classifying satire and fake news

Distinguishing between satire and fake news can be difficult for algorithms as fake news bears a lot of textual cues that make it resemble satirical news more than real news (Rubin et al., 2016; Horne and Adali, 2017), and often satirical news is reprinted in trustworthy outlets as if the reporting

were true (Rubin et al., 2016). Researchers have even found success in using satirical news to stand in for fake news when training a classifier for legitimate and fake news (Jeronimo et al., 2019).

Much research on automatic classification has been done in areas that are adjacent but not directly relevant to ours. For instance, there have been numerous attempts to identify fake news, such as (Rashkin et al., 2017; Ghosh and Shah, 2018) and as evinced by various literature reviews, e.g., (Volkova et al., 2017; Zannettou et al., 2019). There is research on detecting news occupying the gray area between real and fake — news of unknown veracity, i.e. rumors (Alkhodair et al., 2020a). There is also some research interest in telling satirical news apart from real news, e.g., (Burfoot and Baldwin, 2009; Rubin et al., 2016; Ahmed et al., 2018; De Sarkar et al., 2018). Work also has been done on classifying real, fake, and satirical news at the same time, e.g., (de Morais et al., 2019; Vaibhav et al., 2019). However, we are aware of only three pioneering works that address the more challenging task of distinguishing *purely* between satire and fake news: (Volkova et al., 2017; Horne and Adali, 2017; Golbeck et al., 2018).

Volkova et al. (2017) explored two problems, classifying verified and suspicious news and classifying different types of suspicious news. Their attempt at classifying suspicious news into four types of suspicious news categories — propaganda, hoaxes, satire, and clickbait — is the only portion of their work that is relevant to ours. Classifying different types of suspicious news is not strictly the same as separating fake news from satire; their pernicious inclusion of propaganda under the banner of fake news is a major methodological difference from our work of separating fake news and satire. To illustrate why we believe so, we will use the New York Times as an example. In Figure 1 of their paper, the New York Times (NYTimes) is classified as a "verified" news source. But according to Noam Chomsky, who along with Edward Herman put forward the well-cited propaganda model in *Manufacturing Consent* (Herman and Chomsky, 2010), the New York Times falls under the category of pure propaganda (Chomsky, 2015). Another methodological difference with the dataset used in (Volkova et al., 2017) is the highly imbalanced class distribution, with the number of propaganda posts being an order of magnitude greater than other categories of fake news. While Volkova et al. (2017) stated that the source labels were manually verified for quality assurance, they did not state that they manually verified the labels for each and every news story. This is tantamount to judging the contents of a book (all news stories) by its cover (the source).

Horne and Adali (2017) sought to identify which news stories are real, fake, or satire. Of the

three datasets included in their study, only one dataset is relevant to our work, which is their political news data set. The dataset consists of real, fake, and satire categories and has a collection of 75 stories for each category. While the class distribution is balanced for this dataset, the sources are not, as the fake news stories are sourced from only eight outlets and, for satirical news stories, only six outlets. Like Volkova et al. (2017), they also did not hand code the stories individually and instead labeled them according to the trustworthiness of their respective sources.

In the work by Golbeck et al. (2018), fake news and satire are given definitions that align closely with the definitions used in this paper. Additionally, a well-considered set of data collection guidelines are used to not only clearly delineate the boundaries between fake and satirical news but also to ensure diversity and reliability in the collected data (Section 4.1 details those guidelines). As Golbeck et al. (2018) are the only ones we are aware of thus far to have released such a rigorously vetted fake and satirical news dataset, later works dealing with the automatic identification of fake and satirical news such as (Shabani and Sokhn, 2018; Das and Clark, 2019) have relied solely on the dataset shared by Golbeck et al. (2018) and did not use the dataset from either Horne and Adali (2017) or Volkova et al. (2017).

Based on the quality of the datasets and methodological concerns, we had intended to rely solely on the Golbeck et al. (2018) dataset. After all, claims of generalizability made in two other papers (Shabani and Sokhn, 2018; Das and Clark, 2019) rested entirely upon results obtained from just the single high quality Golbeck et al. (2018) dataset. However, to avoid accusations that we are cherry-picking datasets and to assuage concerns over our approach basing its generalizability claims on evaluations from a single dataset, we reported our framework's performance on Horne and Adali (2017) and Volkova et al. (2017) datasets as well, even though we are opposed to the inclusion of these datasets in our paper on methodological grounds.

2.3. Transformers

Transformers are a class of neural network architecture that has proven itself superior in many natural language processing (NLP) tasks, such as news rumor detection (Alkhodair et al., 2020b), question answering, and language inference, compared to recurrent neural networks as well as traditional machine learning (Vaswani et al., 2017). The success of transformers is due to the attention mechanism being able to better model dependencies between tokens even when the tokens

are far apart from each other compared to models that rely on recurrence. Some of the more successful and widely used transformer variants are BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), OpenAI's GPT (Generative Pre-trained Transformer) (Radford et al., 2018), and XLNet (Yang et al., 2019).

2.3.1. BERT and DistilBERT

Of the existing transformer architectures, we chose to build our framework around DistilBERT (Sanh et al., 2020), a BERT derivative. BERT is a language model that incorporates only the encoder section of the original transformer model out of the decoder and encoder sections (Vaswani et al., 2017), allowing each token in BERT to attend to context from both directions instead of only from the left as in the original transformer. We chose DistilBERT because BERT is considered a baseline architecture and DistilBERT is simply a more efficient derivative of BERT.

BERT's status as a baseline architecture can be seen by the many fake news detection studies that uses it unmodified (see Section 2.3.2). The large number of BERT-derived architectures, such as RoBERTa (Liu et al., 2019b) and FlauBERT (Le et al., 2020), is further evidence of BERT being a baseline architecture.

DistilBERT's development was spurred by BERT's computationally demanding nature as even BERT-Base, which requires significantly less memory than BERT-Large to run, can prove too taxing for many systems. Through the use of the compression technique known as knowledge distillation, DistilBERT's makers, `huggingface`, claim that despite DistilBERT being 40% of BERT's size and 60% faster, it retains up to 99% of BERT's language understanding capabilities.

2.3.2. Identifying fake news with transformers

The classification of fake news using transformers have almost always involved leveraging vanilla BERT, e.g., (Jwa et al., 2019; Liu et al., 2019a; Aggarwal et al., 2020; Kaliyar et al., 2021), instead of other transformer architectures, with few exceptions, e.g., (Hassan and Lee, 2021). Of those who used BERT, all of them are concerned with fine-tuning the transformer except for Jwa et al. (2019) who did pay attention to further pre-training the transformer, albeit only with an *external* in-domain dataset and not with the *target* dataset itself (Section 4.1 of our paper discusses the distinction between *external* and *target* datasets).

The core of the BERT architecture is usually left un-tampered with. Where researchers tend to focus their creativity is on modifying the inputs received by BERT and finding new ways to process the outputs from BERT. The two-stage model from Liu et al. (2019a) feeds a coarse-grained label obtained from the first-stage BERT classifier as input to the second-stage BERT classifier. FakeBERT from Kaliyar et al. (2021) uses layers of convolutional neural networks to pool the outputs from BERT instead of more conventional linear layers.

There are researchers who do not even attempt to make significant alterations to the inputs or outputs, which can be taken as a testament to the robustness of the BERT architecture. The BAKE model by Jwa et al. (2019), in the authors' own words, is "essentially a BERT model, [and they] call it BAKE because [they] first applied it to the task of fake news detection in [their] case". The exBAKE variant does feature a significant change as it uses weights altered by additional pre-training on an external in-domain dataset. Aggarwal et al. (2020) also used the BERT model as is, with the only notable change being their dataset cleaning step. This consists of the removal of outlier data, which are texts that are either too short or too long, and the removal of what they termed "noise", consisting of characters such as punctuation marks, numbers, and newlines.

The way in which Hassan and Lee (2021) relied on transformers was not to fine-tune one to a chosen target task but to obtain sentence-level embeddings without further fine-tuning. They used the Universal Sentence Encoder (USE), specifically the `universal-sentence-encoder-large` model that is trained with a Transformer (Vaswani et al., 2017) encoder[1].

For the validation of their classifiers, each paper relied on just one dataset. Liu et al. (2019a) used the LIAR dataset, which is a collection of short statements from Politifact, Aggarwal et al. (2020) used the NewFN dataset, Kaliyar et al. (2021) used a Kaggle dataset of fake and real news covering the 2016 US presidential election, and Jwa et al. (2019); Hassan and Lee (2021) used the Fake News Challenge FNC-1 dataset. For pre-training, the `External` dataset that Jwa et al. (2019) used was a corpus of CNN and Daily Mail articles originally used to test text summarization algorithms.

Jwa et al. (2019) fed BERT with both the headline and body text of an article and no indication was given on how text sequences that exceeded BERT's token length limit were handled. The

[1] `https://aihub.cloud.google.com/p/products%2F42c1bfd4-8104-450c-a348-29b047d3691c`

work by Liu et al. (2019a) dealt with short statements and their dataset features 17.9 tokens on average, which is well under the 768-token length limit of the BERT-Base variant they used. In the work by Aggarwal et al. (2020) the text sequences in their dataset has a median length of 597 tokens. Aggarwal et al. (2020) opted for a naive strategy of limiting sequence length to just 256 tokens, half that of the 512-token length limit of their particular variant of BERT-Base. Going well below the limit was done in the interest of reducing runtime and computing resource usage. Kaliyar et al. (2021) also used naive truncation with the token length limit of 512 matching that of their BERT-Base model. The USE found in the classifier built by Hassan and Lee (2021) does not have a hard length limit.

Vanilla BERT has also made appearances in multimodal fake news classifiers. Multimodal classification problems are where different modalities (text, images, etc.) are considered before arriving at a labeling decision. In the circulation of fake news, doctored or misleading images can often be found alongside untruthful text articles. Two teams, one composed of researchers affiliated with institutions from India and Japan (Singhal et al., 2019) and another featuring members hailing from Spain and China (Giachanou et al., 2020) arrived at very similar approaches for making full use of available image and text data in detecting fake news. The graphical data found in a news article is handled by convolutional neural networks, with Singhal et al. (2019) using VGG-19 and Giachanou et al. (2020) using VGG-16. The original BERT architecture with BERT-Base weights is leveraged in both works to deal with the textual component of news articles. One of the contributions made by the later work of Giachanou et al. (2020) is the use of text-image similarity as a classification feature. This similarity is computed between word embeddings of image tags extracted by VGG and the title text's word embeddings. Singhal et al. (2019) used Twitter and Weibo data while Giachanou et al. (2020) used the GossipCop subset of the FakeNewsNet collection. BERT is capable of accommodating lengthier texts, but both teams elected to limit the length of text sequences processed by BERT. In (Singhal et al., 2019), BERT has access to up to 23 words in Twitter tweets and 200 characters of Weibo posts. In (Giachanou et al., 2020), BERT is fed the text sequences derived only from an article's text and each sequence is length-limited to 64 tokens.

Although the vast improvements brought about by transformers has led to their widespread application across a variety of domains, we are only aware of transformers being used for the detection of fake news among real news, and we have not seen transformers applied in any capacity

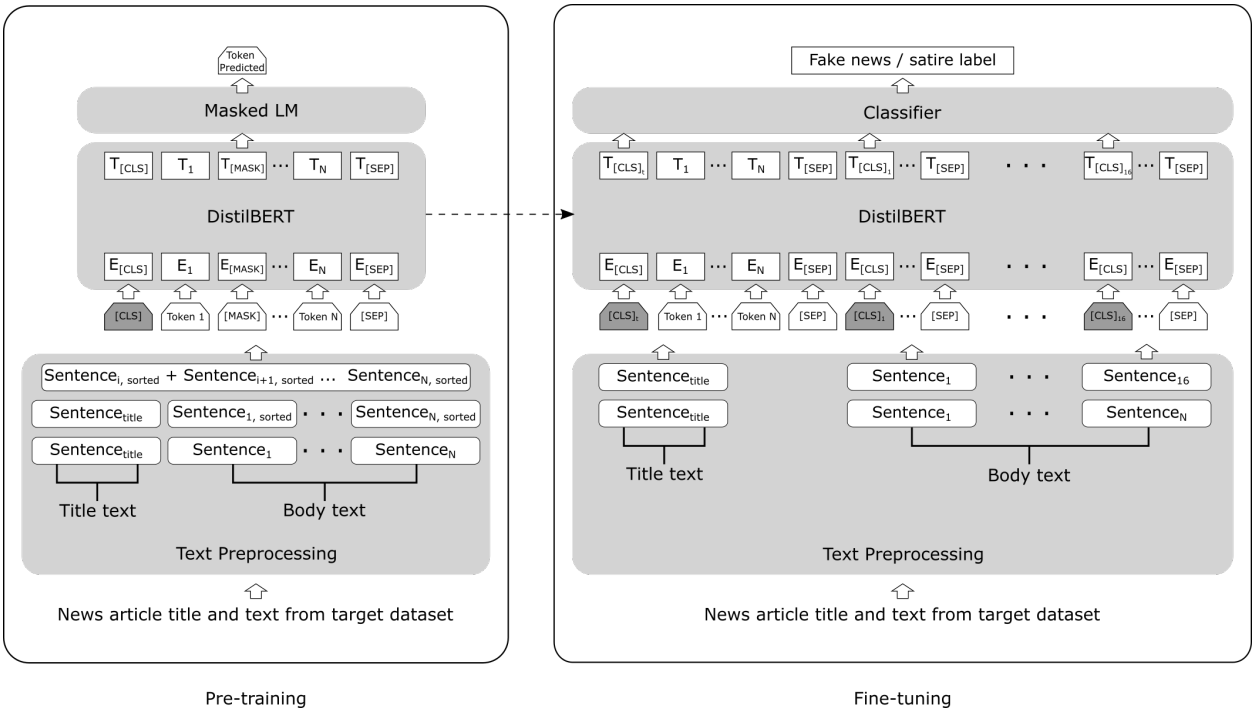

Fig. 1. Fake news and satire classification framework. The particular configuration presented in this figure, `MultiCLS-SimSents-Title+Body`, is pre-trained using the masked language modeling task with the same dataset as the dataset used for the target/downstream fake news and satire classification task. During pre-training, sentences from an article's body are first re-ordered by similarity to its title (`SimSents`). Then, a random sentence i is picked as the starting point, concatenated with sentences following it, and tokenized. During fine-tuning, DistilBERT parameters are initialized using values obtained at the end of pre-training. 17 special aggregator tokens, `[CLS]`, are used during fine-tuning (`MultiCLS`), with 1 for the article title and the remaining 16 for the first 16 sentences from the article's body with their original order preserved (`Title+Body`).

to the specific task of telling apart fake news and satire.

3. Identifying satire *and* fake news with transformers

This section presents the three major components of our proposed fake and satirical news classifier framework: the information aggregator token design, transformer pre-training choice, and the text preprocessing method for both the pre-training and fine-tuning stages. Each unique combination will be referred to as a model. DistilBERT coupled with pre-trained `distilbert-base-uncased` weights serves as the starting point for all our experiments.

Figure 1 illustrates one model from our framework and shows how the components interact with each other.

Scheme	Tokens															
SingleCLS	[CLS]	animal	noises	.	cats	me	##ow	.	dogs	bark	.	[SEP]	[PAD]	[PAD]	[PAD]	[PAD]
TBCLS	[CLS]	animal	noises	.	[SEP]	[CLS]	cats	me	##ow	.	dogs	bark	.	[SEP]	[PAD]	[PAD]
MultiCLS	[CLS]	animal	noises	.	[SEP]	[CLS]	cats	me	##ow	.	[SEP]	[CLS]	dogs	bark	.	[SEP]

Table 1. Tokenization under single-CLS, TB-CLS, and multi-CLS schemes given an article titled "Animal noises." with the body text "Cats meow. Dogs bark.", assuming a 16-token limit.

3.1. Single-CLS, TB-CLS and multi-CLS for information aggregation

BERT and DistilBERT take sequences of tokenized texts as input. A sequence can be any arbitrary span of text, be it a single linguistic sentence or multiple sentences. Tokenization relies on the WordPiece Model (WPM) vocabulary, built on the basis of statistical NLP to address the issue of keeping fixed vocabulary sizes while still being able to process rare words (Wu et al., 2016). In BERT and DistilBERT, additional tokens that perform special functions are inserted among the tokenized texts. One such special token is the classifier token, `[CLS]`, used to aggregate information for an entire sequence of tokens, with that aggregated information typically being used for classification tasks as the name implies. The separator token, `[SEP]`, demarcates the end of a sequence. CLS tokens are placed at the start of sequences and SEP at the end.

In the original formulation of BERT and DistilBERT, only one CLS token is used per sequence of tokens. Recently, researchers found that state-of-the-art results for abstractive and extractive text summarization can be achieved (Liu and Lapata, 2019) when using multiple CLS tokens instead of one, in addition to other modifications to standard BERT architecture. Their results prompted us to investigate if the use of multiple CLS tokens would also help in our fake news and satire classification task. We experimented with two new schemes: (1) TB-CLS, short for Title-Body-CLS, where the article title is represented by one CLS token, and the entire body text is represented by another CLS token; and (2) multi-CLS, where every linguistic sentence gets one CLS token that serves to represent information from that sentence only. In essence, TB-CLS uses article section segmentation while multi-CLS uses sentence segmentation when placing CLS tokens. The differences in the tokenization schemes can be observed in Table 1.

We limited ourselves to having only a maximum of 16 sentences and 1 title sentence for a total of 17. In cases where we have less than 17 sentences and consequently less than 17 CLS tokens, the values for the missing CLS tokens are substituted with zeros.

While transformers output contextual vectors for all tokens, we used only CLS token vectors

during classification. Accordingly, the number of neurons for each of the two hidden feedforward neural layers that together output the fake/satire class probabilities corresponds to the number of CLS tokens multiplied by the dimensionality of the token embedding, with the single-CLS, TB-CLS, and multi-CLS models having 768, 1,536, and 13,056 units per layer, respectively. The final classification layer applies logarithmic Softmax to the outputs, while the preceding layer uses ReLU (rectified linear units).

3.2. Unsupervised pre-training

Although the default `distilbert-base-uncased` weights for DistilBERT has already undergone extensive unsupervised pre-training on massive general text datasets, namely the English Wikipedia and the Toronto Book Corpus, its performance can still be improved upon with further unsupervised pre-training on datasets that are closer to the problem domain of interest (Sun et al., 2019). This can be accomplished by either using the same dataset as the task of interest (denoted as `Target`) or a separate dataset from the same domain as the task of interest (denoted as `External`). We selected a significantly larger within-domain external dataset (NELA-GT is 2GB, while the data by Golbeck et al. (2018) used for our target task is only 1.44MB) to observe if the performance gain typically associated with using a larger dataset for pre-training (Liu et al., 2019b) can overcome the fact that the pre-training dataset differs from the target task dataset, even if the external dataset is within-domain.

The original BERT uses MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) as pre-training objectives. NSP and MLM are unsupervised objectives, allowing a transformer's language modeling capability to be improved with unlabeled text data, which is vastly more common than labeled data. In MLM, the transformer is tasked with predicting tokens that have been randomly masked. In NSP, the transformer is tasked with determining if a sentence B is actually the next sentence that follows a sentence A. As RoBERTa (a Robustly Optimized BERT Pre-training Approach) (Liu et al., 2019b) demonstrated that downstream task performance is retained or even improved when excluding the NSP objective during pre-training in certain settings, we chose to pre-train with MLM as the only objective. DistilBERT developers themselves pre-trained with MLM as the only objective as well.

To generate the samples for pre-training, we follow the set of guidelines outlined in RoBERTa. A

random document is first selected. Then a random linguistic sentence is selected as the starting point within the document. This sentence and the sentences following it are tokenized, then truncated to the first 512 tokens (DistilBERT's token length limit). 15% of the tokens in the sequence are randomly selected for possible replacement. 80% of these randomly selected tokens are replaced with the `[MASK]` token, 10% are replaced with random tokens in the vocabulary, and 10% are left unchanged. Only dynamic masking was used, i.e., the masking pattern changes every time we retrieve a sequence. Unlike RoBERTa, if there is space for additional tokens (e.g., when a very short sentence at the end of a document was selected), we filled the additional spaces with padding tokens, `[PAD]`; in RoBERTa, a special separator token is inserted at the end of the first document, and the additional space is filled by another randomly masked sequence of sentence(s) from another randomly selected document. RoBERTa performed slightly better when sequences were restricted to come from a single document instead of multiple documents, but its developers ultimately chose to use sentences from multiple documents to avoid variable batch sizes. There are no explicit guidelines for handling documents with distinct title and body sections in RoBERTa, so we simply treated the title and body texts for each document as one contiguous sequence of text.

In addition to using the datasets as is, we also investigated the effect of training on text with sentences sorted by informativeness, reasoning that models that rely on texts sorted using the same method during the fine-tuning stage would benefit from it. Specifically, the sentences in each article's body text are sorted in an ascending order based on TF-IDF (term frequency-inverse document frequency) cosine distance from the article's title (see Section 3.3 for details). Since the fine-tuning stage is performed only on the target dataset, we do not believe that pre-training with sorted sentences for the external dataset would be beneficial. Therefore, we tested the effect of text sorting during pre-training only on the target dataset and excluded the external dataset. Weights obtained via pre-training on articles with sorted sentences are referred to as `SimSents`.

Pre-training is usually a time-consuming process. In the interest of making our transformer-based approach more attractive than existing options, we also investigated whether light pre-training with a small number of samples could give the same benefits as more extensive pre-training using a large number of samples. We ended up with six different sets of pre-trained weights for our comparisons (Table 2).

Although there have been BERT-based architectures that attempted to adapt to a domain by

Name	Dataset	# steps	Batch size	Total samples
`SimSents(7.5M)`	Golbeck et al. (2018) with TF-IDF sorted body text	750K	10	7.5M
`SimSents(30K)`	Golbeck et al. (2018) with TF-IDF sorted body text	15K	2	30K
`Target(7.5M)`	Golbeck et al. (2018)	750K	10	7.5M
`Target(30K)`	Golbeck et al. (2018)	15K	2	30K
`External(7.5M)`	NELA-GT	750K	10	7.5M
`External(30K)`	NELA-GT	15K	2	30K

Table 2. Pre-trained weights.

pre-training from scratch, such as SciBERT (pre-trained on scientific papers (Beltagy et al., 2019)), the pre-training in this work was not begun from scratch with randomly initialized weights but was instead started from `distilbert-base-uncased` weights. This decision was made out of concern for DistilBERT's language understanding capabilities, which might not be fully developed if pre-training from scratch using an extremely small dataset (1.44MB Golbeck et al. (2018) dataset), especially when the total number of training samples is low. To put the sample size into perspective, our most extensively trained model at 7.5M samples is only 0.9% of the lightest RoBERTa model, which was pre-trained over 800M samples (100K steps with a batch size of 8K). DistilBERT's developers mentioned they used "up to 4K examples per batch" but did not mention the number of steps.

3.3. Text preprocessing

During pre-training and fine-tuning we experimented with two text preprocessing strategies. The naive strategy uses the provided text as is and cuts off tokens that exceed DistilBERT's 512-token limit, even though we might potentially be excluding text that could help us discriminate between fake and satirical news. The other strategy we looked into is to sort sentences in an article's body text based on their informativeness, reasoning that there are sentences that we can safely exclude when they exceed the token limit, as they are filled with extraneous details that will not be conducive in helping us determine whether an article is fake or satire. This second strategy is essentially text summarization.

Rather than immediately going for resource-intensive state-of-the-art unsupervised text summarization methods, we opted to start with a very simple baseline method. That method is to pick sentences from the body text that most closely resemble the title of the news article because the words in a title often capture the gist of a story. Our implementation involves first using

the pre-trained Punkt English sentence tokenizer from the NLTK Python package to get a list of sentences for each article, then creating a TF-IDF values table for all the words of all the sentences found in an article, and finally sorting the sentences so that those with the smallest TF-IDF cosine distances from the title are at the top. During the pre-training phase, the above method was used to sort each article's body's sentences in the target dataset so that the dataset could be used for MLM pre-training to obtain the `SimSent` set of weights.

The tokenization of a text sequence is achieved through splitting by punctuation and WordPiece (Wu et al., 2016) prior to truncating them to fit DistilBERT's 512-token limit. If a model calls for sentence sorting, then sentence tokenization occurs before this word tokenization step.

In the fine-tuning stage for multi-CLS models, regardless of whether TF-IDF sorted or unsorted sentences are used, a limit N must be imposed on the number of body text sentences selected for consideration, as the classifier cannot account for a variable number of CLS token vectors. We let $N = 16$ for all multi-CLS models. Both TB-CLS and single-CLS models do not have a limit on the number of sentences but are limited by the space remaining on the per-sequence 512-token limit after accounting for the title text tokens and the special CLS and SEP tokens. As TB-CLS and single-CLS models do not need a CLS token inserted before every sentence, the sentences are concatenated together with one whitespace acting as a separator between them prior to word tokenization.

In multi-CLS models, each sentence (including the title) is limited to a maximum of 62 tokens, not including the special CLS and SEP tokens. In the hypothetical case where we did not use the title text and have 16 sentences with ≥ 62 tokens prior to prepending and appending the 2 special tokens, the transformer would only be able to process the first 8 sentences because $8 \times 64 = 512$. No per-sentence length limits are imposed on single-CLS and TB-CLS models.

When the body of a news article is used unaltered, we denote it as `Body` in the model name. When sentences from the body bearing the highest TF-IDF cosine similarities with the title are used before other sentences, we denote it as `SimSents`. If the title text is used, it will be indicated by `Title`. In single-CLS models, `Title` is always prepended to either `Body` or `SimSents` using a period combined with a whitespace as the separator. For all models, `Title` is never used by itself, as our initial exploratory tests showed that the `Title` alone does not give satisfactory results.

Parameters

For the Python packages `random`, `numpy`, and `pytorch`, we used a random seed of 42. The DistilBERT implementation we used is from the `transformers` package released by `huggingface`. Source code for our classifier framework has been made available publicly[2].

For pre-training and fine-tuning, AdamW is used as the optimizer with a learning rate of 2e−5 and a weight decay of 1e−2 for the core DistilBERT model. The same values for learning rate and weight decay are used for MLM during pre-training and for the classifier layers during fine-tuning. These values are within the generally recommended range for pre-training and fine-tuning. The number of steps and batch sizes for pre-training are detailed earlier in Section 3.2. During fine-tuning, the models are trained for 10 epochs with a batch size of 8 for each cross validation split.

4. Evaluation and discussion

4.1. Datasets

We used four datasets in our experiments. The first dataset is used for pre-training and for evaluating the performance of our classifier. This dataset is known as the *target* dataset. The second dataset is used exclusively for pre-training only. This dataset is known as the *external* dataset. The third and fourth datasets are used to allay concerns over our classifier's generalizability. Table 3 contains the descriptive statistics, separated by class and text feature, for each of the three datasets that were used for evaluations, i.e. the target and generalizability datasets.

Fake and Satirical News (Target)[3]. The dataset provided by Golbeck et al. (2018) is used to benchmark our DistilBERT-based classifier framework on the target/downstream task — distinguishing between fake news and satire — hence its designation as the *target* dataset. This dataset is also used for pre-training DistilBERT. It contains a total of 283 fake and 203 satirical news articles. One sample from each category from the dataset can be found in Table 4.

A carefully thought-out set of guidelines was used in building this dataset. News defined as fake must not include opinion pieces, satire, and legitimate news with factual inaccuracy. To ensure articles are similar so as to diminish the chances of an article's topic influencing its fake/satire classification, (1) the stories are all exclusively centered around American politics and (2) the articles

[2] `https://github.com/jwenfai/fakenews-clean`
[3] `https://github.com/jgolbeck/fakenews.git`

Dataset	Class	Text	# Characters				# Words				# Sentences				# DistilBERT Tokens			
			Min	Max	Med	90%	Min	Max	Med	90%	Min	Max	Med	90%	Min	Max	Med	90%
Golbeck et al. (2018)	Fake	Title	14	182	73	99	2	34	13	19	1	3	1	1	3	46	15	21
		Body	86	25439	1897	5663	19	4155	350	1074	1	191	13	35	19	5304	392	1174
	Satire	Title	31	147	69	97	5	27	11	19	1	2	1	1	5	33	13	21
		Body	404	30490	1722	3423	80	6023	336	713	2	297	12	30	91	6678	366	747
Horne and Adali (2017)	Fake	Title	46	146	75	94	6	27	14	18	1	3	1	2	6	31	16	20
		Body	590	10052	2241	5934	114	1985	432	1134	4	61	17	37	122	2192	474	1228
	Satire	Title	23	112	61	86	4	20	11	14	1	2	1	1	4	20	12	17
		Body	190	8368	1355	3637	37	1653	270	728	1	127	7	31	39	1755	298	779
Volkova et al. (2017)	Propaganda	Body	7	152	122	140	2	48	21	28	1	12	1	2	4	120	39	57
	Clickbait	Body	17	148	120	140	3	33	21	28	1	5	1	2	6	103	41	53
	Satire	Body	9	148	117	140	2	46	20	27	1	5	1	2	4	63	37	50
	Hoax	Body	19	148	118	140	3	39	21	28	1	6	1	3	6	71	36	50

Table 3. Summary statistics for text features found in target and generalizability datasets. Med stands for median and 90% stands for the 90th-percentile.

have to be posted after January 2016 (this resulted in dates ranging from January 2016 to October 2017). There must be no borderline cases; fake news stories are incontrovertibly "factually incorrect and deceptive", and satirical stories are "obviously satirical". Another requirement is that only five articles/samples at most from each website/source are included in the dataset — a balanced number of samples reduces the risk of training a classifier that overfits on some specific writing style/tone/markers/indicators that might be commonly found within all articles from a particular source. The authors hand coded each individual story instead of relying on a proxy indicator like the reputations of source websites. Crucially, the dataset authors recognize that there are fake news websites using disclaimers of being satirical as a way to deflect from accusations of publishing fake news, so the authors have ruled out grey area cases from their dataset. For fake news stories, the dataset authors have also included a link rebutting the claims made by the stories.

There are three features for each news article: title, URL, and body. We did not use the URL because we aim to distinguish satire from fake news purely through processing natural language. Many of the URLs in the dataset also no longer function.

NELA-GT (External)[4]**.** The NELA-GT dataset, provided by Nørregaard et al. (2019), is *never used for the evaluation* of our classifier framework's effectiveness and is used solely to pre-train DistilBERT, hence its designation as the *external* dataset. The motivating factor in our use of

[4] `https://doi.org/10.7910/DVN/ULHLCB`

Fake News		
Title	URL	Body
Trump Has Fired Muslim Sharia Judge Arrested And Charged	https://worldpoliticsnow.com/2017/08/breaking-trump-fired-muslim-sharia-judge-arrested-charged/	As it turns out, in Donald Trump's America, you can't just go declaring religious laws from pagan countries that hate us to be legal if you sit on the federal judiciary. 22nd Circuit Court of Appeals Judge Hassan al-Hallamallala-Smith — an Obama nominee in 2009 — was taken into federal custody by the FBI for breaking the oath he took to uphold the Constitution. Under title 18 US Code Subsection 1209.3 ...
Satirical News		
Title	URL	Body
President Trump Fires All 14 Muslim Federal Judges	http://asamericanasapplepie.org/2017/08/16/of-course-he-cant-do-that/	Barack Obama managed to pack the federal judiciary with bleeding heart liberals, professional lobbyists for climate change and other science-fiction scams and 15 Moslems, none of whom were born in this country. President Trump had no choice but to fire one of them last month when he ruled that Sharia Law was OK if the defendant is Moslem and reasonably expected to follow his faith or face eternity in hell. You can force a Christian to pay for birth control ...

Table 4. Examples from the "Fake News vs Satire" dataset released by Golbeck et al. (2018)

this dataset is the desire to observe if pre-training with a dataset from the same domain as the dataset used in the downstream task (news stories) can improve performance and if a larger and more diverse dataset is better for pre-training.

NELA-GT is a corpus of 713,534 English news articles pulled from a wide variety of popular and less well-known news sources (totaling 194) such as Al Jazeera, Infowars, Shadow Proof, Spiegel, and The Michelle Malkin Blog. As the dataset was built for the study of misinformation, the veracity of the sources is as diverse as the sources themselves (e.g., the dataset included articles from satirical news sites such as The Beaverton and The Onion). The dates of the articles in NELA-GT range from February 1st, 2018, to November 30, 2018, which do not overlap with the target dataset's range of dates.

Horne and Adali (2017) (Generalizability dataset 1)[5]. This dataset contains three categories of political news — real, fake, and satire — with 75 items for each category. As our work is concerned only with distinguishing between fakes and satires, we only used the 75 fake and 75 satirical articles. Each article is composed of two sections of text, title and body. The categorization of news articles in this dataset is wholly dependent upon the articles' respective sources. An article from a source deemed to be satirical is considered satirical irrespective of its content. The same

[5] https://github.com/rpitrust/fakenewsdata1

applies to fake news.

There are a total of eight sources used for fake articles and they are derived from a list of fake and misleading news websites from Zimdars (2016a), an assistant professor of communication and media at Merrimack College at the time the list was first published (Zimdars, 2016b). To be included, the fake news sources must have had at least one story proven as false on a fact-checking website. There are six sources for satirical news and the criterion for inclusion is that they must explicitly state their satirical nature on their website's front page.

Volkova et al. (2017) (Generalizability dataset 2)[6]. This dataset consists of 30,894 propaganda, 836 clickbait, 2,083 satire, and 1,341 hoax tweets. The number of tweets we managed to retrieve from the "rehydration" process [7] is significantly lower than the number of tweets found in the original paper[8] (56,271 propaganda, 1,366 clickbait, 3,156 satire, and 4,549 hoaxes) likely due to many tweets being deleted over the intervening time period. The class distribution of the dataset remains highly imbalanced in favor of propaganda tweets, although hoax tweets now constitutes a smaller portion of the dataset than in the original.

Unlike the other three datasets used in our work, the news items are tweets, meaning that they are severely limited in length compared to regular news articles due to the character limit imposed by Twitter. They do not have clearly delineated title and body sections. Many tweets are simply the titles and/or subtitles of complete articles followed by a hyperlink to said articles.

The categorization of the tweets are wholly dependent upon the reputation of the tweets' respective sources (the Twitter account where the tweet originated from and that account's corresponding website if it has one). The sources themselves are classified based on information from "several public resources that annotate suspicious Twitter accounts" and these resources were not identified by Volkova et al. (2017) except for PropOrNot[9] and Fake News Watch[10]. A total of 174 suspicious

[6] `https://github.com/jwenfai/fakenews-clean/blob/master/volkova/volkova_rehydrated_id.csv`

[7] In accordance to Twitter data sharing policy, many researchers only share tweet IDs publicly and other researchers intending to access data associated with the IDs such as the text itself must reconstitute the tweet using Twitter APIs in a process known as rehydration, i.e. request the full Tweet, user, or Direct Message content. Twitter's content redistribution policy can be found at `https://developer.twitter.com/en/developer-terms/agreement-and-policy`.

[8] `http://www.cs.jhu.edu/~svitlana/TwitterList`

[9] `http://www.propornot.com/p/the-list.html`

[10] The original website at `http://www.fakenewswatch.com/` is inaccessible, US Library of Congress archival snapshot of the site on March 2017 can be found at `https://webarchive.loc.gov/all/20170301201107/http://www.fakenewswatch.com/`.

news accounts were identified, although we do not know how many of these accounts remain among the tweets we succeeded in rehydrating. All tweets were collected from the suspicious accounts one week before and after the Brussels bombing on March 22, 2016. Retweets from other users that mention the suspicious accounts were also collected and given the same label as the originals'.

4.2. Metrics and comparison basis

We adhere to the standard of reporting means of scores from 10-fold cross validation used by Golbeck et al. (2018) as well as other publications (Shabani and Sokhn, 2018; Das and Clark, 2019) that tested their classifier models against the Golbeck et al. (2018) dataset. The stratified k-fold implementation from `scikit-learn` was used for this purpose.

We compare models from our framework against other approaches that are purely machine-based, as in methods that have no human involvement in determining if a news story is fake or satire. In our comparison against existing state-of-the-art approaches, we report F1 and accuracy scores. F1 scores were reported in the original paper by Golbeck et al. (2018) and a subsequent paper by Das and Clark (2019). Accuracy is reported only because it was the sole metric reported in the Shabani and Sokhn (2018) paper. Class imbalance biasing accuracy scores is not an issue with the Golbeck et al. (2018) dataset as both classes have fairly equal numbers of instances (283 fake and 203 satirical articles).

For comparisons between different model configurations (pre-training, CLS type, text preprocessing) within our framework, we report the support[11]-weighted F1.

4.3. Metrics and comparison basis for generalizability datasets

We reserved the taxing comprehensive evaluations of our satire versus fake news classifier to be performed on just the Golbeck et al. (2018) *target* dataset; the testing of our framework against the Horne and Adali (2017) and Volkova et al. (2017) *generalizability* datasets are limited but rigorous. The tests are rigorous in that they employed the same test conditions and metrics as those used by the dataset creators and they are limited in that not every model configuration within our framework was tested, partly due to limitations imposed by the nature of the dataset themselves and the aforementioned test conditions.

[11]Number of true instances for each label.

While both the Volkova et al. (2017) and Horne and Adali (2017) papers have received large numbers of citations, our literature review failed to reveal any subsequent work that tested classifiers against the datasets found in these papers. Therefore, we report only the metrics used by the original authors of the datasets to enable comparison. For the Horne and Adali (2017) dataset, the original metric is accuracy. For the Volkova et al. (2017) dataset, the original metrics are micro-averaged F1 and macro-averaged F1. Horne and Adali (2017) tested their classifier using 5-fold cross validation while Volkova et al. (2017) used 10-fold cross validation; we tested on the same numbers of folds for the respective datasets.

The reported scores for these datasets should not be directly compared against those obtained from the Golbeck et al. (2018) dataset due to different evaluation designs. The main difference is that our evaluations with the Golbeck et al. (2018) dataset relied on using title and body texts together, which could not be done on either generalizability datasets for different reasons.

In their evaluations, Horne and Adali (2017) did not use features generated from both the title and body texts simultaneously, instead opting to classify an article either solely with features from an article's title or solely with features from an article's body. A fair comparison required us to limit our classifier to using either title or body text. This excludes evaluating the `TBCLS` aggregator token design. We also excluded the `SimSents` text summarization technique from evaluations as it summarizes an article's body text based on the article's title. When using title text as the feature, sentence tokenization was performed on the title as otherwise the `MultiCLS` models would be nigh indistinguishable from `SingleCLS` models.

The Volkova et al. (2017) dataset consists of tweets without distinct title and body sections. In the absence of separate title and body sections, `SimSents` and `TBCLS` were excluded from evaluations. Rehydration has managed to recover only about half of all the tweets used in the original evaluations performed by Volkova et al. (2017) in addition to changing the class distribution slightly, so our scores are only good proxy indicators of how our framework would perform if the complete Twitter data were made available to us.

For the evaluations involving both generalizability datasets, we elected not to pre-train the classifier on the respective target texts as pre-training can be time-consuming, choosing instead to use the set of weights pre-trained extensively on the NELA-GT corpus of English news articles (`External`). This still allows us to evaluate if using weights that were further pre-trained to fit the

Model	F1	Acc.
`TBCLS-SimSents(7.5M)-Title+Body`	**.8679**	**.8686**
`MultiCLS-Target(7.5M)-Title+Body`	.8671	.8685
`MultiCLS-SimSents(7.5M)-Title+Body`	.8630	.8644
Das and Clark (2019)	.825	-
Shabani and Sokhn (2018) (machine-only)	-	.8164
Golbeck et al. (2018)	.791	-

Table 5. Performance comparison of models from our framework against existing state-of-the-art on the Golbeck et al. (2018) dataset. Accuracy is the sole metric reported in Shabani and Sokhn (2018).

news domain can boost performance over using weights that have received no further pre-training.

4.4. Results discussion

Models from our framework succeeded in attaining state-of-the-art performance. Overall, the ideal configuration for our framework is to first pre-train the transformer on millions of masked examples of sorted sentences from the texts of the target task and then using that set of weights when fine-tuning with multiple CLS aggregator tokens, one for each sentence, on unsorted sentences.

Performance comparison of existing approaches for classifying fake news and satire against the three best models built within our framework can be found in Table 5. A breakdown on how choices in pre-training sample sizes, pre-training dataset, aggregator token architecture, and text preprocessing affect the performance of a model built within our framework can be found in Table 6.

To fully reap the rewards of pre-training, it needs to be extensive (e.g., 7.5M samples) and not light (e.g., 30K samples). This is evinced by models using extensively pre-trained weights topping the performance table. Light pre-training not only results in performance that is at best mildly better than no pre-training but also in some cases can be even worse than no pre-training. The most illustrative example is the group of `SingleCLS` models using `Title+Body` text, which all performed worse when using lightly pre-trained weights than models with no pre-training, regardless of the pre-training dataset used. Many lightly pre-trained models also failed to outperform existing approaches, just like their counterparts without pre-training.

The size and diversity of the dataset used during pre-training do not appear to affect the efficacy of the pre-training process, at least not as much as the resemblance of the pre-training dataset to the downstream/target task dataset does. Across all CLS schemes, text preprocessing methods, and pre-training sample sizes, models using weights pre-trained on the target dataset or sorted sentences from the target dataset (`SimSents`) always perform better than those using weights pre-trained

CLS type	Text	# pre-train samples	Pre-training dataset: SimSents F1	Target F1	External F1	NoPretrain F1	Pre-trained (mean) F1
MultiCLS	Title+SimSents	7.5M	.8613	.8478	.8248	.7800	.8446
		30K	.8233	.8187	.7938		.8119
	Title+Body	7.5M	.8630	.8671	.8371	.7913	.8557
		30K	.8077	.8251	.8037		.8122
TBCLS	Title+SimSents	7.5M	.8483	.8564	.8311	.8091	.8453
		30K	.8256	.8295	.7977		.8176
	Title+Body	7.5M	**.8679**	.8568	.8251	.7872	.8499
		30K	.7983	.8242	.7875		.8033
SingleCLS	Title+SimSents	7.5M	.8627	.8415	.8275	.8005	.8439
		30K	.8247	.8239	.8160		.8215
	Title+Body	7.5M	.8542	.8606	.8471	.8294	.8540
		30K	.8150	.8283	.8071		.8168
Title+SimSents (mean)		7.5M	.8574	.8486	.8278	.7965	.8446
		30K	.8245	.8240	.8025		.8170
Title+Body (mean)		7.5M	.8617	.8615	.8364	.8026	.8532
		30K	.8070	.8259	.7995		.8108

Table 6. Distinguishing between fake and satirical news with DistilBERT-based models. Scores under the Pre-trained (mean) column are row-wise means of scores that used pre-trained weights, SimSents, Target, and External. The Title+SimSents (mean) and Title+Body (mean) rows are column-wise means.

on an external dataset. Still, using weights from extensive pre-training with an external dataset is beneficial, as doing so always results in better performance than comparable models with no pre-training. A number of models from our framework that did not use pre-trained weights fared worse than existing state-of-the-art approaches but managed to outperform the existing approaches when using pre-trained weights extensively pre-trained on the external dataset.

As we expected, fine-tuning with sorted sentences (Title+SimSents) performs better when paired with weights pre-trained on sorted sentences (SimSents) than when paired with unsorted sentences (Target), especially when the pre-training process is extensive. While we also expected that this would translate into SimSents-Title+SimSents models performing better than SimSents-Title+Body, the reverse actually holds true. Furthermore, models using SimSents weights that place themselves among the best performing ones (see Table 5) did so when paired with unsorted sentences (Title+Body).

In general, fine-tuning with sorted sentences does not confer the same benefit as pre-training with sorted sentences. Based on the averaged scores, we can observe that models using sorted sentences performed similarly and are always worse than models using unsorted sentences when

extensively pre-trained. This is also borne out in a comparison of the models on an individual basis. Among extensively pre-trained models, models that fine-tune with unsorted sentences almost invariably perform better than equivalent ones using sorted sentences, with two exceptions being `SingleCLS-SimSents` and `TBCLS-External`.

Our observations led us to conclude that for the task of distinguishing fake news and satire, the best approach is to pre-train extensively on sorted sentences and use unsorted sentences when fine-tuning.

Among all extensively pre-trained models that fine-tune on unsorted sentences, the best models on average are those that use multi-CLS, followed by single-CLS, and finally TB-CLS. Coupled with the fact that two multi-CLS models occupy the top three spots in terms of F1 and AUC-ROC, we believe the multi-CLS class of models to be best suited for the task for classifying fake news and satire, even though the top scoring model uses TB-CLS.

Pre-training is the most important factor in realizing the full potential of transformers as the top models all use extensively pre-trained weights. Without pre-training, the majority of models do not outperform existing state-of-the-art approaches, with `SingleCLS-NoPretrain-Title+Body` being the sole exception. As the pre-training step is in essence masked language modelling, this implies that developing a language model that is better fitted to the texts found in the target task results in better classification results. As for models using more than one aggregator CLS token occupying the top spots in performance, being allowed to spread out information aggregation over multiple CLS tokens instead of squeezing all the aggregated information into one token may have allowed for more nuanced classification decisions to be made.

4.5. Results discussion for generalizability datasets

Our approach generalizes well beyond the Golbeck et al. (2018) dataset as it outperforms by a considerable margin the best classifiers that can be found in (Horne and Adali, 2017) and (Volkova et al., 2017), which are the only other papers featuring publicly available datasets that distinguish between satire and fake news that we are aware of. Results from the Volkova et al. (2017) dataset also shows that our approach is not just applicable to a multiclass scenario but is also capable of correctly classifying minority classes even when a severe imbalance exists in class sizes.

Table 7 shows our DistilBERT-based classifier framework's performance when tested against the

Model/ CLS type	Text	# pre-train samples	Pre-training-dataset: External Acc.	Pre-training-dataset: NoPretrain Acc.
MultiCLS	Title	7.5M	.8133	.8000
	Body	7.5M	**.8733**	.8533
SingleCLS	Title	7.5M	.8400	.8267
	Body	7.5M	**.8733**	.8467
Horne and Adali (2017)	Title	-	-	.55
	Body	-	-	.67

Table 7. Distinguishing between fake and satirical news with DistilBERT-based models on the Horne and Adali (2017) dataset. Accuracy is the sole metric reported in Horne and Adali (2017).

Model/ CLS type	Text	# pre-train samples	Pre-training-dataset: External F1 Micro	External F1 Macro	NoPretrain F1 Micro	NoPretrain F1 Macro
MultiCLS	Body	7.5M	**.9984**	**.9938**	.9981	.9923
SingleCLS	Body	7.5M	.9981	.9923	.9979	.9918
Volkova et al. (2017)	Body	-	-	-	.92	.71

Table 8. Distinguishing between propaganda, hoaxes, clickbait, and satire with DistilBERT-based models on the Volkova et al. (2017) dataset. Micro-F1 and macro-F1 are the only metrics reported for multi-class classification in Volkova et al. (2017).

Horne and Adali (2017) dataset. Results from testing against the Volkova et al. (2017) dataset are shown in Table 8.

Findings from our evaluations using the Golbeck et al. (2018) dataset are mostly equally valid for both generalizability datasets. Models using weights that have been further pre-trained on an in-domain dataset definitively perform better than those using weights that were not pre-trained further. The case for using multiple CLS aggregator tokens is slightly weaker here as `SingleCLS` models outperform `MultiCLS` models in the Horne and Adali (2017) dataset when classifying an article based on its `Title` text. This can perhaps be attributed to many article titles being one sentence long, preventing the advantages of using multiple CLS tokens from being fully expressed. Both single- and multi-CLS configurations performed equally well in all other scenarios for both datasets. Regardless of the number of CLS tokens used, our models still obtained better scores than previous approaches.

Using the body text instead of the title text when classifying an article in the Horne and Adali (2017) dataset always improves classification performance. This concurs with our experience of

poor performance when classifying articles in the Golbeck et al. (2018) dataset using title text alone during early trials that ultimately led us to exclude this framework configuration choice from in-depth evaluations. Unlike the classifier used in Horne and Adali (2017), our approach exhibited a less severe performance gap between using title text alone and using body text alone, with a maximum difference of 0.06 in accuracy score instead of 0.12.

The Volkova et al. (2017) dataset is dominated by the propaganda class (30,894 instances), which is several times the size of the satire, hoax, and clickbait classes (4,260 instances combined). Recall that a macro-F1 score is obtained by calculating the F1 on a per class basis and taking the unweighted mean of those scores, while a micro-F1 score is calculated by counting the number of true positives, false negatives, and false positives globally. A macro-F1 score significantly lower than a micro-F1 score in a class-imbalanced scenario indicates that a classifier is worse at correctly labeling the minority classes than it is at labeling the majority class, which is the case with the best classifier presented in the original paper by Volkova et al. (2017). Since our classifier achieved near parity on both micro- and macro-F1 metrics, our classifier is equally adept at correctly labeling the minority classes as it is at labeling the majority propaganda class.

4.6. Potential improvements

There are a number of unexplored opportunities for improving upon the existing framework.

4.6.1. Different transformer architectures

In this work, the particular transformer variant we are using, DistilBERT, can be fairly described as lightweight. There are many more complex and computationally taxing transformer variants that are not BERT-derived that we have yet to test, e.g., CTRL, GPT2, and XLNet, that have higher token limits or other ways to treat long text sequences that exceed their token limits.

4.6.2. Different text summarization techniques

Similarly, there are better text summarization techniques in the literature that we have not explored, such as hierarchical attention networks, or even transformers-based text summarization, e.g., (Liu and Lapata, 2019). Summarization techniques that do not rely on using pre-existing title text can also be explored. An example is one that sorts sentences based on the sum of their TF-IDF values. Another example is the use of Code Quality Principle (CQP) to automatically identify

sentences highly relevant to a text's topic, which is also automatically identified and is composed of words with high Term Frequencies (TF) (Lloret et al., 2013).

4.6.3. News in other languages

So far, American political news written in English has received the lion's share of research attention when it comes to the problem of distinguishing between real, fake, and satirical news, as reflected by the datasets found in wide usage in the literature. Applying transformers to datasets in other languages, such as this French fake news and satire dataset (Liu et al., 2019c) or these Brazilian Portuguese datasets that contain legitimate, fake, and satirical news (de Morais et al., 2019; Jeronimo et al., 2019), would allow us to ascertain if transformers can hold their performance advantage across different languages.

4.6.4. Outperforming classifier models that involve humans

Crowdsourcing the task of discerning fake news from satire achieved an accuracy of 0.84 (Shabani and Sokhn, 2018), a result bested by almost all of our framework's models that have been extensively pre-trained on the target dataset with either sorted or unsorted sentences. Unfortunately, even our best-scoring model attained only 0.8686 in accuracy, which, while close, is still lower than the 0.872 achieved by their hybrid machine-crowd ensembling approach (Shabani and Sokhn, 2018).

Developing a machine-only approach that does better than methods involving humans offers several advantages. In a media environment where a high volume of news is being pumped out at every waking hour, human labor can be a bottleneck for these labor-intensive methods. Methods that depend on humans can also be finicky to tune to achieve the best result. The aforementioned hybrid machine-crowd ensembling approach has to find the correct threshold to determine when the classification label by the ensemble of machine-learning models is insufficient and a news story needs to be evaluated by crowd workers; a threshold value that works well on one dataset often does not work well in another.

4.6.5. Improved interpretability

Feature engineering and feature importance are discussions that are very prominent in fake news detection works that rely on traditional machine learning, e.g., (Rubin et al., 2016; Shabani and Sokhn, 2018; Das and Clark, 2019), but are noticeably absent in works that rely on transformers,

e.g., (Jwa et al., 2019; Liu et al., 2019a), because neural network language models are notoriously difficult to dissect. There is space for future work that focuses solely on the mechanisms by which transformers distinguish between fake and satirical news, perhaps drawing upon existing work on visualizing the innards of transformers, e.g., (Coenen et al., 2019).

5. Conclusion

In this paper, we presented a novel DistilBERT-based framework for separating fake news and satire. We demonstrated that with sufficiently extensive MLM pre-training, be it using the target dataset or an external in-domain dataset, a number of our models using different text preprocessing strategies in the fine-tuning stage can achieve results better than existing state-of-the-art methods, with the best models using multiple CLS tokens instead of the standard singular CLS token. Our work shows that when using weights that have previously been pre-trained on large text corpora (Wikipedia etc.), further pre-training with the downstream task's dataset can drastically improve results even when the downstream dataset is relatively minuscule. Successful validation of our approach over several datasets demonstrates that the observed performance improvement over existing approaches is unlikely to be a localized phenomenon but is instead general. Additionally, our approach for telling apart fakes and satires offers an advantage over existing approaches by using open source code throughout the entire process instead of generating features using proprietary code such as IBM tone analyzer in the case of (Das and Clark, 2019) and Linguistic Inquiry and Word Count (LIWC) in the case of (Shabani and Sokhn, 2018). We hope that our work will contribute towards keeping humor (and, to a lesser extent but of no less importance, political awareness and political engagement) from being unintentionally suppressed in our relentless pursuit to eradicate fake news.

Acknowledgements

The second author is supported by the Discovery Grants (RGPIN-2018-03872) from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chairs Program (950-230623). The third author is supported by Cluster project (#R16083) and Provost Research Fellowship grant (#R20093) from Zayed University, United Arab Emirates.

References

Aggarwal, A., Chauhan, A., Kumar, D., Mittal, M., Verma, S., 2020. Classification of Fake News by Fine-tuning Deep Bidirectional Transformers based Language Model. EAI Endorsed Transactions on Scalable Information Systems 7. URL: https://eudl.eu/doi/10.4108/eai.13-7-2018.163973.

Ahmed, H., Traore, I., Saad, S., 2018. Detecting opinion spams and fake news using text classification. Security and Privacy 1, e9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9, doi:10.1002/spy2.9.

Ahmed, W., Vidal-Alaball, J., Downing, J., Seguí, F.L., 2020. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. Journal of Medical Internet Research 22, e19458. URL: https://www.jmir.org/2020/5/e19458/, doi:10.2196/19458.

Alkhodair, S.A., Ding, S.H.H., Fung, B.C.M., Liu, J., 2020a. Detecting breaking news rumors of emerging topics in social media. Information Processing and Management (IP&M) 57, 1–13. doi:10.1016/j.ipm.2019.02.016.

Alkhodair, S.A., Fung, B.C.M., Ding, S.H.H., Cheung, W.K., Huang, S.C., 2020b. Detecting high-engaging breaking news rumors in social media. ACM Transactions on Management Information Systems (TMIS) 12, 8.1–8.16. doi:10.1145/3416703.

Allcott, H., Gentzkow, M., 2017. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives 31, 211–236. URL: http://www.aeaweb.org/articles?id=10.1257/jep.31.2.211, doi:10.1257/jep.31.2.211.

Becker, A.B., Bode, L., 2018. Satire as a source for learning? The differential impact of news versus satire exposure on net neutrality knowledge gain. Information, Communication & Society 21, 612–625. URL: https://doi.org/10.1080/1369118X.2017.1301517, doi:10.1080/1369118X.2017.1301517.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 3615–3620. URL: https://www.aclweb.org/anthology/D19-1371, doi:10.18653/v1/D19-1371.

Brewer, P.R., Young, D.G., Morreale, M., 2013. The Impact of Real News about "Fake News": Intertextual Processes and Political Satire. International Journal of Public Opinion Research 25, 323–343. URL: https://academic-oup-com.proxy3.library.mcgill.ca/ijpor/article/25/3/323/786961, doi:10.1093/ijpor/edt015.

Burfoot, C., Baldwin, T., 2009. Automatic satire detection: Are you having a laugh?, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09, Association for Computational Linguistics, Suntec, Singapore. p. 161. URL: http://portal.acm.org/citation.cfm?doid=1667583.1667633, doi:10.3115/1667583.1667633.

Chen, H.T., Gan, C., Sun, P., 2017. How Does Political Satire Influence Political Participation? Examining the Role of Counter- and Pro-Attitudinal Exposure, Anger, and Personal Issue Importance. International Journal of Communication 11, 19. URL: https://ijoc.org/index.php/ijoc/article/view/6158.

Chomsky, N., 2015. Noam Chomsky: The New York Times is pure propaganda. URL: https://www.salon.com/2015/05/25/noam_chomsky_the_new_york_times_is_pure_proganda_partner/.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., Wattenberg, M., 2019. Visualizing and Measuring the Geometry of BERT. arXiv:1906.02715 [cs, stat] URL: http://arxiv.org/abs/1906.02715, arXiv:1906.02715.

Das, D., Clark, A.J., 2019. Satire vs Fake News: You Can Tell by the Way They Say It, in: 2019 First International Conference on Transdisciplinary AI (TransAI), pp. 22–26. doi:10.1109/TransAI46475.2019.00012.

de Morais, J.I., Abonizio, H.Q., Tavares, G.M., da Fonseca, A.A., Barbon, S., 2019. Deciding among Fake, Satirical, Objective and Legitimate news: A multi-label classification system, in: Proceedings of the XV Brazilian Symposium on Information Systems, Association for Computing

Machinery, New York, NY, USA. pp. 1–8. URL: https://doi.org/10.1145/3330204.3330231, doi:10.1145/3330204.3330231.

De Sarkar, S., Yang, F., Mukherjee, A., 2018. Attending Sentences to detect Satirical Fake News, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA. pp. 3371–3380. URL: https://www.aclweb.org/anthology/C18-1285.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Ghosh, S., Shah, C., 2018. Towards automatic fake news classification. Proceedings of the Association for Information Science and Technology 55, 805–807. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2018.14505501125, doi:10.1002/pra2.2018.14505501125.

Giachanou, A., Zhang, G., Rosso, P., 2020. Multimodal Multi-image Fake News Detection, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 647–654. doi:10.1109/DSAA49011.2020.00091.

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K.H., Bonk, C., Bouzaghrane, M.A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J.B., Falak, W., Gieringer, C., Graney, J., Hoffman, K.M., Huth, L., Ma, Z., Jha, M., Khan, M., Kori, V., Lewis, E., Mirano, G., Mohn IV, W.T., Mussenden, S., Nelson, T.M., Mcwillie, S., Pant, A., Shetye, P., Shrestha, R., Steinheimer, A., Subramanian, A., Visnansky, G., 2018. Fake News vs Satire: A Dataset and Analysis, in: Proceedings of the 10th ACM Conference on Web Science, Association for Computing Machinery, Amsterdam, Netherlands. pp. 17–21. URL: https://doi.org/10.1145/3201064.3201100, doi:10.1145/3201064.3201100.

Hassan, F.M., Lee, M., 2021. Multi-stage News-Stance Classification Based on Lexical and Neural Features, in: Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., Corchado, E. (Eds.), 13th International Conference on Computational Intelligence in Security for Information Systems

(CISIS 2020), Springer International Publishing, Cham. pp. 218–228. doi:10.1007/978-3-030-57805-3_21.

Herman, E.S., Chomsky, N., 2010. Manufacturing Consent: The Political Economy of the Mass Media. Random House.

Horne, B.D., Adali, S., 2017. This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News, in: Eleventh International AAAI Conference on Web and Social Media, pp. 759–766. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15772.

Jennen, B., Nussbaum, A., 2021. Germany and France Oppose Trump's Twitter Exile. Bloomberg.com URL: https://www.bloomberg.com/news/articles/2021-01-11/merkel-sees-closing-trump-s-social-media-accounts-problematic.

Jeronimo, C.L.M., Marinho, L.B., Campelo, C.E.C., Veloso, A., da Costa Melo, A.S., 2019. Fake News Classification Based on Subjective Language, in: Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, Association for Computing Machinery, New York, NY, USA. pp. 15–24. URL: https://doi.org/10.1145/3366030.3366039, doi:10.1145/3366030.3366039.

Jones, M.O., 2017. Satire, social media and revolutionary cultural production in the Bahrain uprising: From utopian fiction to political satire. Communication and the Public 2, 136–153. URL: https://doi.org/10.1177/2057047317706372, doi:10.1177/2057047317706372.

Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H., 2019. exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). Applied Sciences 9, 4062. URL: https://www.mdpi.com/2076-3417/9/19/4062, doi:10.3390/app9194062.

Kaliyar, R.K., Goswami, A., Narang, P., 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications URL: https://doi.org/10.1007/s11042-020-10183-2, doi:10.1007/s11042-020-10183-2.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B.,

Besacier, L., Schwab, D., 2020. FlauBERT: Unsupervised Language Model Pre-training for French. arXiv:1912.05372 [cs] URL: http://arxiv.org/abs/1912.05372, arXiv:1912.05372.

Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., Lu, X., 2019a. A Two-Stage Model Based on BERT for Short Fake News Detection, in: Douligeris, C., Karagiannis, D., Apostolou, D. (Eds.), Knowledge Science, Engineering and Management, Springer International Publishing, Cham. pp. 172–183. doi:10.1007/978-3-030-29563-9_17.

Liu, Y., Lapata, M., 2019. Text Summarization with Pretrained Encoders, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 3730–3740. URL: https://www.aclweb.org/anthology/D19-1387, doi:10.18653/v1/D19-1387.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs] URL: http://arxiv.org/abs/1907.11692, arXiv:1907.11692.

Liu, Z., Shabani, S., Balet, N.G., Sokhn, M., 2019c. Detection of Satiric News on Social Media: Analysis of the Phenomenon with a French Dataset, in: 2019 28th International Conference on Computer Communication and Networks (ICCCN), pp. 1–6. doi:10.1109/ICCCN.2019.8847041.

Lloret, E., Romá-Ferri, M.T., Palomar, M., 2013. COMPENDIUM: A text summarization system for generating abstracts of research papers. Data & Knowledge Engineering 88, 164–175. URL: https://www.sciencedirect.com/science/article/pii/S0169023X13000815, doi:10.1016/j.datak.2013.08.005.

Mukurunge, T., Rapitse, S., 2019. Comic Relief, Telling Unwanted Truth to Power and Ethical Journalism: Satire Writing in Print Media in Lesotho. Journal of Psychology Research 9. URL: http://www.davidpublisher.org/index.php/Home/Article/index?id=41968.html, doi:10.17265/2159-5542/2019.09.004.

News, F.D., 2020. About. URL: https://firstdraftnews.org:443/about/.

Nørregaard, J., Horne, B.D., Adalı, S., 2019. NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. Proceedings of the International AAAI Conference on Web and Social Media 13, 630–638. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/3261.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training. Technical Report. OpenAI. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y., 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark. pp. 2931–2937. URL: https://www.aclweb.org/anthology/D17-1317, doi:10.18653/v1/D17-1317.

Rubin, V., Conroy, N., Chen, Y., Cornwell, S., 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News, in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection, Association for Computational Linguistics, San Diego, California. pp. 7–17. URL: http://aclweb.org/anthology/W16-0802, doi:10.18653/v1/W16-0802.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs] URL: http://arxiv.org/abs/1910.01108, arXiv:1910.01108.

Shabani, S., Sokhn, M., 2018. Hybrid Machine-Crowd Approach for Fake News Detection, in: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), pp. 299–306. doi:10.1109/CIC.2018.00048.

Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S., 2019. SpotFake: A Multi-modal Framework for Fake News Detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp. 39–47. doi:10.1109/BigMM.2019.00-44.

Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. How to Fine-Tune BERT for Text Classification?, in:

Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (Eds.), Chinese Computational Linguistics, Springer International Publishing, Cham. pp. 194–206. doi:10.1007/978-3-030-32381-3_16.

Tandoc, E.C., Lim, Z.W., Ling, R., 2018. Defining "Fake News": A typology of scholarly definitions. Digital Journalism 6, 137–153. URL: https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1360143, doi:10.1080/21670811.2017.1360143.

Vaibhav, V., Mandyam, R., Hovy, E., 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification, in: Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), Association for Computational Linguistics, Hong Kong. pp. 134–139. URL: https://www.aclweb.org/anthology/D19-5316, doi:10.18653/v1/D19-5316.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N., 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada. pp. 647–653. URL: http://aclweb.org/anthology/P17-2102, doi:10.18653/v1/P17-2102.

Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science 359, 1146–1151. URL: https://science.sciencemag.org/content/359/6380/1146, doi:10.1126/science.aap9559.

Wardle, C., 2017. Fake news. It's complicated. URL: https://firstdraftnews.org:443/latest/fake-news-complicated/.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato,

Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs] URL: http://arxiv.org/abs/1609.08144, arXiv:1609.08144.

Xie, B., He, D., Mercer, T., Wang, Y., Wu, D., Fleischmann, K.R., Zhang, Y., Yoder, L.H., Stephens, K.K., Mackert, M., Lee, M.K., 2020. Global health crises are also information crises: A call to action. Journal of the Association for Information Science and Technology n/a, 1–5. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24357, doi:10.1002/asi.24357.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. XLNet: Generalized autoregressive pretraining for language understanding, in: Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Zannettou, S., Sirivianos, M., Blackburn, J., Kourtellis, N., 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. Journal of Data and Information Quality 11, 10:1–10:37. URL: http://doi.org/10.1145/3309699.

Zimdars, M., 2016a. False, Misleading, Clickbait-y, and Satirical "News" Sources. URL: https://docs.google.com/document/d/10eA5-mCZLSS4MQY5QGb5ewC3VAL6pLkT53V_81ZyitM/preview?usp=embed_facebook.

Zimdars, M., 2016b. My 'fake news list' went viral. But made-up stories are only part of the problem. Washington Post URL: https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem/.