Using Natural Language to Predict Bias and Factuality in Media with a Study on Rationalization

by

Kunal Tangri

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2021

Author . . . .
Department of Electrical Engineering and Computer Science
February 2021

Certified by . . . .
James Glass
Senior Research Scientist
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Using Natural Language to Predict Bias and Factuality in Media with a Study on Rationalization

by

Kunal Tangri

Submitted to the Department of Electrical Engineering and Computer Science
on February 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Fake news is a widespread problem due to the ease of information spread online, and its ability to deceive large populations with intentionally false information. The damage it causes is exacerbated by its political links and loaded language, which make it polarizing in nature, and preys on peoples' psychological biases to make it more believable and viral. In order to dampen the influence of fake news, organizations have begun to manually tag, or develop systems to automatically tag, false and biased information. However, manual efforts struggle to keep up with the rate at which content is published, and automated methods provide very little explanation to convince people of their validity. In an effort to address these issues, we present a system to classify media sources' political bias and factuality levels by analyzing the language that gives fake news its contagious and damaging power. Additionally, we survey potential approaches for increasing the transparency of black-box fake news detection methods.

Thesis Supervisor: James Glass
Title: Senior Research Scientist

Acknowledgments

The past year of research in the Spoken Language Systems group has been an incredibly collaborative and educational experience, and I am grateful to all the members within the group for their advice, discussions, and exemplary work ethics. This environment is a result of the hard work my advisor, Jim Glass, puts into developing and maintaining our research community, and I feel very lucky to have been a part of it. His openness and encouragement of new ideas, commitment to fostering collaboration, and unyielding support have been an invaluable part of my experience.

I am also thankful for the postdoctoral advisors and other peers who were always willing to share their thoughts and advice during my work - Ramy Baly, Tianxing He, Mitra Mohtarami, and Moin Nadeem. They helped shape my research thought process, and offered many interesting conversations.

Additionally, I have been fortunate that my interest in combating fake news is also shared by the Defense Science and Technology Agency of Singapore, and for their willingness to financially support me as a Research Assistant to work on this problem.

Finally, I am extremely blessed for my loving parents, and their endless encouragement and intellectual curiosity. I strive to live by the examples and standards they have set for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern technology has enabled information to spread faster and reach a much broader audience than ever before. Through social media, people can share news stories with the click of a button and online resources have made it easier for people to create websites that host information. This increased availability of information is a huge benefit to society, as the general population is more informed, and it is easier for people to collaborate and stay in touch. However, social media and online resources have lowered the barrier to entry for publishing and spreading information to large populations, which has enabled malicious actors to spread fake news. As the prevalence of fake news has grown, the field of fake news detection has spawned to combat it.

In an article discussing its impact on the 2016 U.S. Presidential election, Allcott and Gentzkow offer a concise definition of fake news: *potentially misleading information that is intentionally and verifiably false* (Allcott and Gentzkow, 2017). As people have been shown to be poor judges of false information (Kumar et al., 2016), factuality prediction has become an essential component in fake news detection systems. However, we believe that fact-checkers, alone, are not sufficient for stifling fake news, and, in fact, they have been shown to potentially entrench people further in their factually incorrect viewpoints (Nyhan and Reifler, 2010).

This problem is related to, if not exacerbated by, the politicization of news. A study by Iyengar and Hahn shows that people prefer to consume media from sources that agree with their own political ideologies (Iyengar and Hahn, 2009). Furthermore, due to confirmation bias, people are more receptive to information that confirms their own views (Nickerson, 1998). These psychological tendencies, when coupled together, result in people being more likely to believe in news that aligns with their political ideologies, inherently linking the problem of fake news to political bias. Ucinski, Klofstad, and Atkinson confirmed this link exists by showing that partisan attachment is correlated with whether people believe information being presented to them (Uscinski et al., 2016). This is extremely concerning, given that the context of Allcott and Gentzkow's article was the impact of fake news on a Presidential election, and that other surveys of the field show that fake news is typically used for political gain (Shu et al., 2017).

Several organizations have come to the same conclusion that both factuality and political bias play a large role in the spread of fake news, and they have created websites to manually annotate news for factuality and/or bias, including Media Bias/Fact Check (Dave Van Zandt, 2015), Allsides (AllSides, 2020), and Politifact (PolitiFact, 2007). Despite the efforts of these organizations, though, public awareness of bias in media remains low (Elejalde et al., 2018). We think this is partly due to a change in how news is being consumed, with social media increasingly becoming the primary news source for consumers (Shu et al., 2017). While social media is undoubtedly a powerful tool for discovering news and sharing it with others, it has also vastly increased the amount of published news and the speed at which it spreads. Manual fact checking and bias labeling simply cannot keep up with rate of content creation well enough to stem the damage of fake news. On top of that, fake news has actually been shown to spread six times faster than real news (Vosoughi et al., 2018), which makes the task of stopping fake news through manual labelling even more difficult. The most obvious next step in preventing the spread of fake news is to automate fact checking and bias prediction, and there are a growing number of benchmark tasks being created to stimulate this field (e.g. FEVER for fact extraction and verification (Thorne et al., 2018) and Hyperpartisan SemEval 2019 for political bias prediction (Kiesel et al., 2019)).

In our work, we leverage methods that have been developed on these benchmark tasks to create a system for predicting media sources' political bias and factuality levels from natural language. Through political bias labeling, we aim to provide some context on the partisan window that people may be viewing information from, and through factuality labeling our goal is to identify misinformation. Additionally, we study approaches that aim to provide transparency in fake news detection methods in an effort to subvert the psychological biases humans have against viewpoints that conflict with their own views. We discuss related methodology in the fields of both fake news detection and rationalization in Chapter 2, before discussing the development of our bias and factuality prediction system in Chapter 3. Chapter 4 describes our experiments in providing transparency in fake news detection models, and we conclude our work and note possible future directions to explore in Chapter 5.

Chapter 2

Related Work

2.1 Bias and Factuality Prediction

In recent years, fake news has become very well-studied due to increasing public awareness about the problems it poses. To provide a background on the current state of the field, we first highlight previous and current methods for detecting fake news, and justify why we focus on predicting bias and factuality at the media source-level using only natural language. We then provide background on the methods that are used in our study.

2.1.1 Previous Methods

The methods currently used to detect fake news fall into three general categories: feature-based, graph-based, and propagation-based (Kumar and Shah, 2018). They are defined as follows:

- **Feature-Based:** methods that rely on linguistic information extracted from news-related text (Horne and Adali, 2017; Kumar et al., 2016; Potthast et al., 2018)

- **Graph-Based:** methods that study networks of user interactions with news (Akoglu et al., 2010; Beutel et al., 2013)

- **Propagation-Based:** methods that model the differences between how real vs. fake information spreads (Tripathy et al., 2010; Nguyen et al., 2012)

While all three of these methods have their merits and comparable performances, we believe that graph and propagation-based methods rely on modeling fake news indirectly and are not ideal as a result. The indirect features they are modeled on, user graphs and information flow, are largely influenced by bot accounts (Nied et al., 2017), which may even be what these models exploit to detect fake news. However, it has been shown that bots are not actually responsible for the overall spread of fake news - they only accelerate it (Davis et al., 2016). This makes graph and propagation-based methods susceptible to adversaries who decide to change the behavior of these bots. Due to these shortcomings, we decide to study feature-based models which focus on the language that gives fake news its viral and damaging power as we believe it is the most robust predictor of fake news.

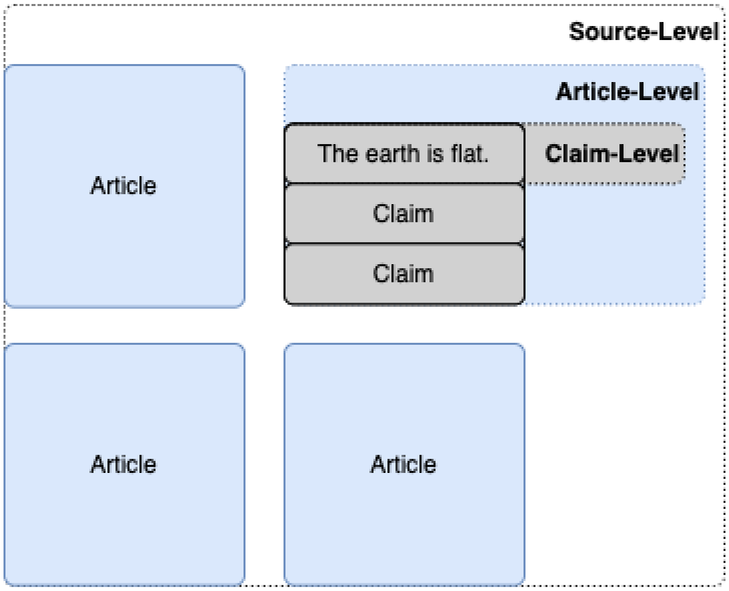


Figure 2-1: An illustration of the different scopes that bias and factuality prediction are performed on. The largest scope is the source-level, which consists of a collection of articles from a media source. Next is the article-level, made up of a collection of claims. Finally, the smallest scope is the claim-level, which is a single assertion. (Note: In this figure we do not include external information that may be used to perform predictions at each of the three levels.)

For bias and factuality prediction, feature-based models operate on three different levels of granularity. From smallest to largest scope, these are the claim-level, article-level, and source-level, and an illustration of these scopes can be found in Figure 2-1.

Claim-Level

At the claim-level, previous methods are mostly centered around predicting factuality, and heavily rely on stance-detection - whether ground-truth sources of information agree or disagree with a claim. Some approaches have used manually fact-checked claims as the ground-truth source for stance-detection (Mukherjee and Weikum, 2015), whereas others use comment-based discussion on social media as proxy for the ground-truth (Kochkina et al., 2018; Dungs et al., 2018). Yet another avenue for stance-detection draws relevant sources of information from the Web to serve as the ground-truth to detect stance against (Mukherjee and Weikum, 2015; Baly et al., 2018b). In the contexts of article-level and source-level scopes, stance-detection has also been useful under the hypothesis that trustworthy articles and sources will tend to agree with truthful claims and disagree with false claims (Mukherjee and Weikum, 2015; Popat et al., 2018). However, as we increase the size of our scope past the claim-level, approaches begin to model bias in addition to factuality and make use of extra information that is available at the larger scopes.

Article-Level

At the article-level, many methods for bias and factuality prediction rely on extracting features from language. In a study across three datasets related to real vs. fake news, Horne and Adali (Horne and Adali, 2017) found that fake news tends to use shorter, simpler, and more repetitive language, and a later work from Horne et al. creates a toolkit to exploit some of these tendencies (Horne et al., 2018b). The features they analyze include complexity, structure, bias and others which we make use of in our research and will describe in more detail in Section 2.1.2. Language is similarly used by Potthast et al. to predict factuality through a stylometric analysis of fake news using N-grams, readability scores, word frequencies, and features specific to the news domain (ratios of quoted words, number of external links, avg. paragraph length) (Potthast et al., 2018). Other studies use only raw text to model factuality, feeding it directly to an LSTM (Rashkin et al., 2017), or more recent language models. In some of these studies on factuality prediction, not only is political bias a useful feature (Horne et al., 2018b), but it is also predictable using some of the exact same methods (Horne et al., 2018a). Furthermore, article-level bias prediction is a task that has also been studied in its own right, using n-grams, lexical features, vocabulary richness, and readability scores (Saleh et al., 2019), as well as through latent representations of article language content extracted from attention-based models (Kulkarni et al., 2018). For both article-level bias and factuality prediction, source-level reliability has been found to be an informative feature (Karadzhov et al., 2017), and can incorporate information external to articles like a source's social media presence and third-party descriptions of the source. However, the field of source-level reliability is understudied compared to the two smallest granularities of bias and factuality prediction, which is why we focus our study on it.

Source-Level

Not only are source-level bias and factuality predictions understudied and useful for article-level predictions, but source-level predictions are also useful by themselves. As noted by Baly et al. (Baly et al., 2020), bias and factuality prediction at the smaller granularity levels can be a computational challenge, and potentially still too slow to effectively prevent the spread of fake news. Profiling entire media sources, however, can provide a good indication on whether newly published material is reliable or not, without the need to verify each claim or each article. Besides this feature of source-level prediction, it also allows us to draw information from more mediums to inform predictions. A majority of the work done at the source-level was developed by Baly et al. in 2018 and 2020 (Baly et al., 2018a, 2020), in which they train Support Vector Machine classifiers on features not only extracted from articles, but also from YouTube, Twitter, Facebook, and Wikipedia data. They hypothesize that YouTube presents additional media published from a source that could be useful for predictions, that Twitter and Facebook can help analyze the audience of a media source, and that Wikipedia provides a third-party view of a media source. Our work builds a system using some of the methodology described by Baly et al. (Baly et al., 2020), and we further describe some of the tools used for extracting features from these data channels in Section 2.1.2.

2.1.2 Methods Used in our Work

Baly et al. use many of the same methods mentioned in the article-level approaches to extract features from all of their selected data channels. These tools are comprised, in part, by the lexicons described by Horne et al. (Horne et al., 2018b) in their News Landscape (NELA) toolkit, but a majority of them rely on latent representations extracted from language models - this is an increasingly used method as language models have become much more powerful in recent years. We make use of both types of feature extraction toolkits in our efforts to build a system based on the methods from Baly et al., so we describe them in more detail here.

The NELA toolkit extracts features that have proven to be useful across a wide range of studies in political bias and factuality prediction, and a discussion of these studies can be found in the paper from Horne et al. (Horne et al., 2018b). NELA extracts the following categories of features:

- **Structure:** Part of speech counts, linguistic features (function words, pronouns, prepositions, etc.), and clickbait title classification (Chakraborty et al., 2016)

- **Sentiment:** Sentiment scores from VADER (Hutto and Gilbert, 2015), and emotion and happiness scores from other lexicons (Recasens and Jurafsky, 2013; Mitchell et al., 2013)

- **Topic-Dependent:** Lexicons to differentiate scientific fields and others to distinguish personal concerns

- **Complexity:** Lexical diversity, readability scores, avg. word and text lengths, and # of cognitive process words
- **Bias:** Several bias lexicons (Recasens and Jurafsky, 2013; Mukherjee and Weikum, 2015) and subjectivity measures (Pang and Lee, 2004)
- **Morality:** Lexicon-based measures of morality (Lin et al., 2018) and features from the Moral Foundation Theory (Graham et al., 2009)

While these features have previously been useful in other studies, and we include them in our own, lexicon-based approaches are giving way to methods using neural language models for feature extraction. Instead of using a rule or vocabulary-based means of feature extraction, neural language models develop latent representations of words which allow them to model some aspect of language (typically the probability of seeing specific sequence of words). These latent representations can provide powerful embeddings of semantics, syntax, and task-specific features at the word, sentence, or even article level. In fact, neural language models that are used for factuality and bias prediction may develop latent representations of the features extracted from lexical based methods. Our study makes use of a few different methods for retrieving these embeddings.

The first method, Global Vectors for Word Representation (GloVe) (Pennington et al., 2014), leverages global vocabulary statistics, as well as local contexts, in a log-bilinear regression model for the unsupervised learning of word representations. GloVe is used by Baly et al. in their 2018 study, and is replaced by newer methods in their 2020 work. Though it is true that GloVe is an older approach for obtaining word embeddings, and has since been outperformed on natural language understanding tasks, we believe there is still some benefit in including it in our study as it has a fundamental difference from the newer embedding methods we experiment with. This difference is that GloVe produces static embeddings, meaning each word has a single, fixed representation.

The newer embedding methods we explore, produce contextual embeddings - each word can have different representations based on the contexts in which it is used.

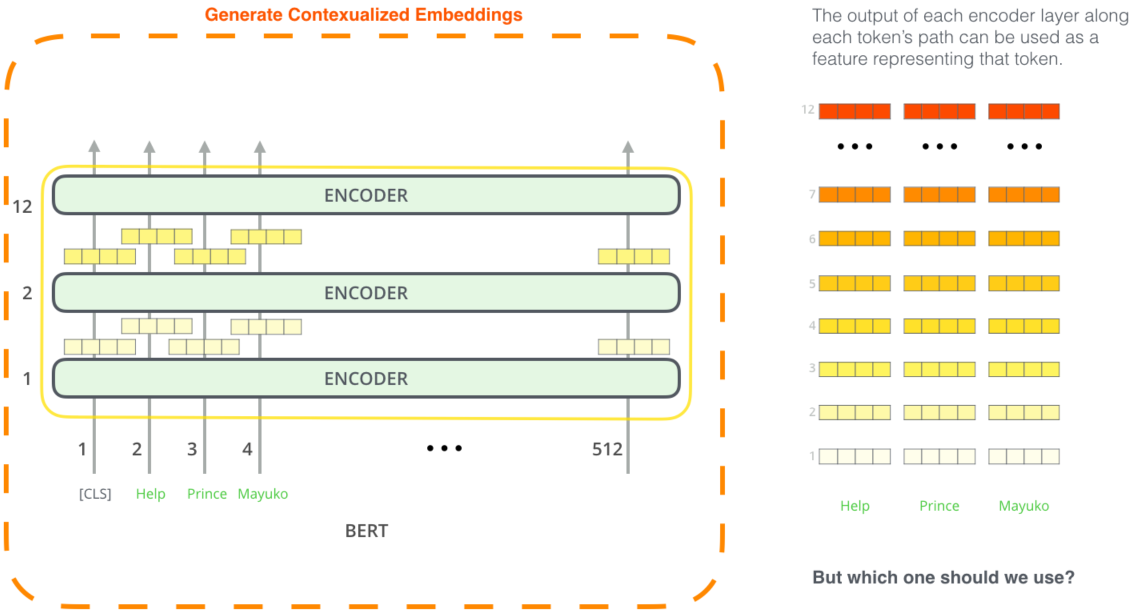

Generate Contexualized Embeddings
12
2
1
ENCODER
ENCODER
ENCODER
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko
BERT
The output of each encoder layer along each token's path can be used as a feature representing that token.
12
7
6
5
4
3
2
1
Help
Prince
Mayuko
But which one should we use?

Figure 2-2: A visualization of the transformer encoder layers within BERT, and the embedded representations that are input and output from each encoder. Embeddings for an input sequence can be retrieved at each encoder layer, and may behave differently depending on which layer they are extracted from (Alammar, 2018).

Methods that we leverage for retrieving these contextual embeddings are based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), a recently developed, transformer-based (Vaswani et al., 2017) model that achieved state-of-the-art performance on a range of natural language understanding tasks. BERT consists of 12 stacked transformer encoder layers, and is pre-trained using two techniques which enable it to develop its powerful bidirectional representations. The first technique is masked language modeling, where BERT predicts randomly masked tokens at any position in the input. A result of this pre-training task is that BERT learns to use context before *and* after the masked word to make its prediction, which gives BERT its bidirectional property. Additionally, BERT is pre-trained on a next sentence prediction task - whether sentence A is followed by sentence B. This technique encourages BERT to reason about the relationships between sentences as well. The resulting pre-trained model can be further fine-tuned to a variety of natural language processing tasks, and we describe how we suit it to our research when we go through our methods in Section 3.2.2.

From either pre-trained or fine-tuned BERT models, token embeddings can be retrieved from each of the 12 encoder layers within BERT, as seen in Figure 2-2. At each layer, these embeddings incorporate information from all the other tokens in the input window, which is what makes these representations contextualized (i.e. dependent on the other tokens in the input text). In our work, we choose to extract embeddings from the second-to-last layer, as it has been shown that deeper layers offer more contextualized embeddings (Ethayarajh, 2019), and the last layer may be biased towards BERT's pre-training objectives (Baly et al., 2020).

Though BERT is very useful for producing embeddings for some of our data, it is poorly-suited to some of the other data channels we use. For example, some data is better suited to embeddings produced at the sentence-level rather than the token-level embeddings offered by BERT. In these cases, we use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), a variant of BERT, instead. SBERT is trained using siamese and triplet networks on top of BERT that encourage it to produce semantically-meaningful embeddings at the sentence-level.

In our work, we experiment with the a subset of the toolkits used by Baly et al. in their 2018 and 2020 works (Baly et al., 2018a, 2020) in order to create a political bias and factuality prediction system. We specifically avoid the toolkits and data from the YouTube and Facebook data channels, as they did not contribute much in the methodology they present. Our system, data, and experiments are described in detail in Chapter 3.

2.2 Rationalization

In addition to building a system for political bias and factuality prediction, we also study rationalization methods for explaining the predictions produced by models relevant to fake news. We include rationale in our work because we believe that without explanations on why an article as has been classified a certain way, whether it is related to bias or factuality prediction, people are unlikely to believe anything opposing their own views (Nyhan and Reifler, 2010). Therefore, in order for an automated bias and factuality prediction system to make a real impact on the spread of fake news, we believe that it must offer some level of transparency to its users.

Though older language-based approaches to political bias and factuality prediction have not explicitly studied rationalization for the purpose of presenting it alongside predictions, the methods are well-suited for transparency. Methods that use part-of-speech tagging, punctuation and stop-word counts, and measures of syntactic complexity have found that fake news is simpler and more repetitive (Horne and Adali, 2017), and can highlight the structural and complexity-related features that are used to classify real vs. fake news. Stance-based methods that reason about the factuality of an article by comparing its claims to known factual claims (Popat et al., 2018) can provide the stances of the articles and the facts they agree and disagree with. Lexicon-based approaches have found that fake news tends to contain emotional, praising, implicative, and perspective specific language (Recasens and Jurafsky, 2013; Horne et al., 2018b). These methods can highlight vocabulary found within their lexicons to offer transparency on their predictions. Finally, n-gram-based models have found that the most highly weighted phrases are those dealing with divisive topics like "trump" and "liberals" and contain dramatic cues like "breaking" (Rashkin et al., 2017). In order to provide an explanation of their predictions, they can list the most predictive n-grams.

Unlike older language-based methods related to fake news detection, though, newer embedding-based methods, like some we use to build our bias and factuality prediction system, operate as black-boxes and are much less transparent as a result. We aim to increase these models' transparencies for fake news detection, and specifically focus on explaining predictions from BERT using attention-based rationale, and gradient-based rationale. Both the attention and gradient-based methods we use produce *extractive rationales*, meaning a model's predictions are explained using a subset of that model's inputs.

2.2.1 Attention

Attention has become a widely used mechanism within the field of natural language processing due to its ability to model global dependencies better than older, hidden state-based recurrent neural network approaches. Conceptually, it operates by computing a weighted sum of input representations which allows models to focus on especially informative inputs, and ignore others, when making predictions (Bahdanau et al., 2015). In addition to improving a range of neural language models' performances, the input weightings that attention computes can be extracted to offer some transparency on the model's predictions (i.e. which inputs are highly weighted when making a prediction). Many existing studies have used this property of attention to explain and interpret their models, including using it to highlight relevant medical details alongside diagnosis prediction (Mullenbach et al., 2018), to provide sentence summaries (Rush et al., 2015), among many other uses (Hermann et al., 2015).

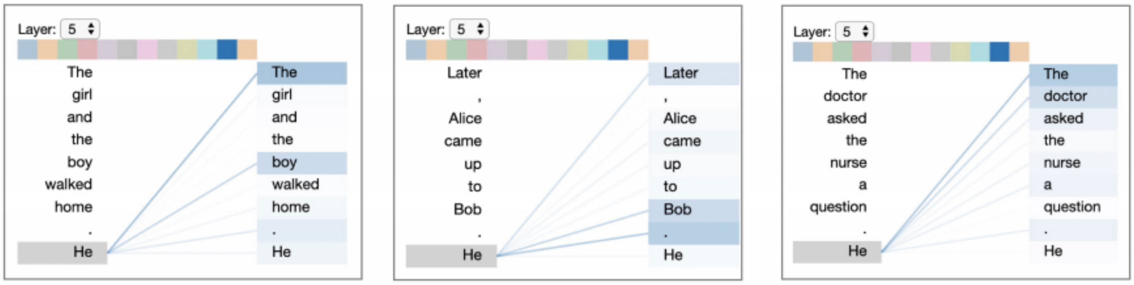

Layer: 5
The girl and the boy walked home . He
The girl and the boy walked home . He
Layer: 5
Later , Alice came up to Bob . He
Later , Alice came up to Bob . He
Layer: 5
The doctor asked the nurse a question . He
The doctor asked the nurse a question . He

Figure 2-3: A visualization of how BERT's attention heads highly weight relevant tokens for a specific task. In these examples, created using the bertviz tool, the attention heads show specializations towards coreference resolution (Vig, 2019).

Early attention-based models, which worked in combination with recurrent neural networks, have since been replaced with models based only on attention. Vaswani et al. introduced the Transformer, which involves stacked layers consisting of several attention mechanisms (also known as attention heads) (Vaswani et al., 2017). The model we use to study rationalizing predictions for fake news, BERT (Devlin et al., 2019), consists of 12 of these Transformer layers, and we analyze the attention heads within them. We are not the first to do so (though we are not aware of any other works operating in the context of fake news), as other studies have performed extensive analysis on how BERT's attention heads specialize linguistically and behave differently across layers of the model (Clark et al., 2019). Additionally, tools have been created purely for the sake of visualizing BERT's different attention heads and increasing the model's transparency (Vig, 2019). An example of these visualizations can be seen in Figure 2-3. Although the news articles we use in our rationalization study are too large to visualize in the same way as Figure 2-3, we create other visualizations to demonstrate highly attended to regions of articles during classification. Our attention-based methods and results are described in Section 4.1.

2.2.2 Gradient-Based Approaches

Though attention is a plausible route for rationalizing predictions related to fake news, the faithfulness of attention distributions to model predictions is still a debated topic in the field (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), which is why we also explore gradient-based approaches as an alternative. In general, gradient-based rationalization methods rely on estimating how much a change in a model's input will impact its output, where inputs that induce the largest changes are thought to be more important in a model's prediction. There are several different approaches to estimating these gradients, but a survey of them by Adebayo et al. (Adebayo et al., 2018) shows that several of these methods produce rationale that are invariant under model and label randomization (i.e. unrelated to the prediction task), and are inadequate as a result. However, pure gradients, involving simply calculating the gradient of the output with respect to the input, are shown to provide legitimate rationale, and the methods we explore are based on them.

We are not aware of other works that explore gradient-based approaches for rationalizing predictions in the domain of fake news detection, but these methods have been used extensively for other tasks, especially in the field of computer vision. Some studies have rationalized digit and object-recognition tasks (Baehrens et al., 2010; Simonyan et al., 2014) to show validity in using pure gradients, and others have since used them to explain predictions like medical imaging diagnosis (Margeloiu et al.,

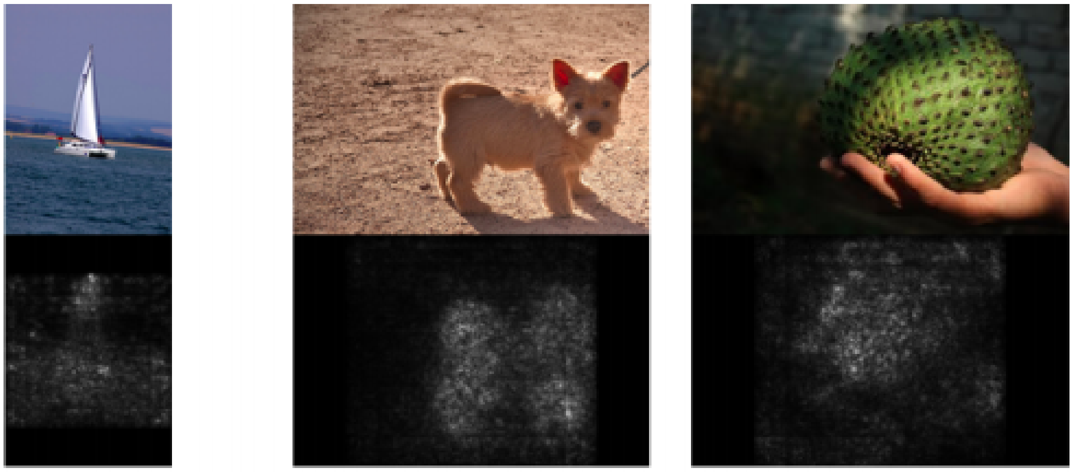


Figure 2-4: An example of pure gradients being used to provide transparency on an image classification task. We can observe that the larger gradients (bottom) correspond to the regions of the original images (top) that contain the object of interest (Simonyan et al., 2014).

2020). In the field of computer vision, these methods produce nice visual rationale for their predictions, as seen in Figure 2-4, but they are also being used to interpret language models and visualize important linguistic inputs as well - the recently released Language Interpretability Tool incorporates these methods (Tenney et al., 2020). We describe our experiments using pure gradient rationales, and show some of our own visualizations in Section 4.2.

Chapter 3

Bias and Factuality Prediction

In this chapter, we describe the bias and factuality prediction system that we implemented using the methodology described by Baly et al. (Baly et al., 2020). We first explain the data and assumptions that the system relies on, before outlining the process of training the necessary models for the system, and the experiments we conducted in order to try and best replicate the results stated by Baly et al. (Baly et al., 2020). However, when we present our results, we note that our findings do differ, and provide some hypotheses of possible causes. Finally, using our trained models, we outline the system's process of making predictions on unseen data, using only media sources' URL's as input.

3.1 Data and Assumptions

This system leverages three different channels of data to make its bias and factuality predictions: what was written by the media source, who is in the media source's audience, and what was written about the media source. The first channel, what was written by the media source, refers to articles that the source has published, and how it defines itself on social media (if the source has a social media account). While it is obvious that the material published by the media source will contain its inherent biases and some level of factuality, we assume that we can find indications of this bias and factuality from linguistic features in this data. For the next channel of data, who is in the media source's audience, we use the followers of the source on social media. More specifically, we analyze the bios of the media source's followers with the assumption that a source's followers tend to agree with the views presented by the source, and that the followers may express their own views within their bios. The final channel of data, what was written about the media source, refers to the media source's Wikipedia page, if it has one. Here, we assume that Wikipedia presents an independent view of the media source, and that this view may contain comments about the media source's bias and/or factuality.

The methodology that our work is based on (Baly et al., 2020) includes data from the original media source, Twitter, YouTube, Facebook, and Wikipedia. However, their work showed that the features extracted from YouTube and Facebook, about a media source, did not have any significant benefit in the system's predictive ability for both bias and factuality. As a result, we decide to exclude those two data channels from our experiments, and only use the original media, Twitter, and Wikipedia to retrieve the necessary data.

In order to train our bias and factuality prediction system given the data channels explained above, we also need a dataset of media sources labeled by both bias and factuality. Media Bias/Fact Check (MBFC) (Dave Van Zandt, 2015) provides a database of media sources manually labeled for both these tasks, from which we obtained 1,737 labeled media sources. Although manual annotation does add an element of subjectivity, we find this problem unavoidable and believe that Media Bias/Fact Check maintains a sufficient level of objectivity through its transparency and correction policies. The MBFC database ranks bias on a 7-point scale: extreme-left, left, left-center, center, right-center, right, and extreme-right. Factuality is rated on a 6-point scale: very-low, low, mixed, mostly-factual, high, and very-high. Like Baly et al.'s approach (Baly et al., 2020), we decide to condense both our bias and factuality ratings to 3-point scales (bias: left, center, right - fact: low, mixed, high) due to loosely defined conditions for the additional labels. After retrieving labels for these sources, we then aggregated the URL's of each media source's homepage, and Twitter and Wikipedia pages if they exist. A few examples from our labeled training dataset are shown in Table 3.1.

Name	**Bias**	**Fact**	**URL**	**Twitter**	**Wikipedia**
Huffington Post	Left	High	huffingtonpost.com	HuffPost	wiki/HuffPost
CNN	Left	Mixed	cnn.com	cnni	wiki/CNN
True Activist	Left	Low	trueactivist.com	TrueActivist	N/A
Foreign Affairs	Center	High	foreignaffairs.com	foreignaffairs	wiki/Foreign_Affairs
The Wrap	Center	Mixed	thewrap.com	thewrap	wiki/TheWrap
The Onion	Center	Low	theonion.com	theonion	wiki/The_Onion
Alt News Media	Right	High	altnewsmedia.net	AltNewsMedia	N/A
Daily Mail	Right	Mixed	dailymail.co.uk	MailOnline	wiki/MailOnline
News Wars	Right	Low	newswars.com	N/A	N/A

Table 3.1: Examples from our dataset including the media sources, their bias and factuality labels, and their corresponding URL's.

3.2 Training the System

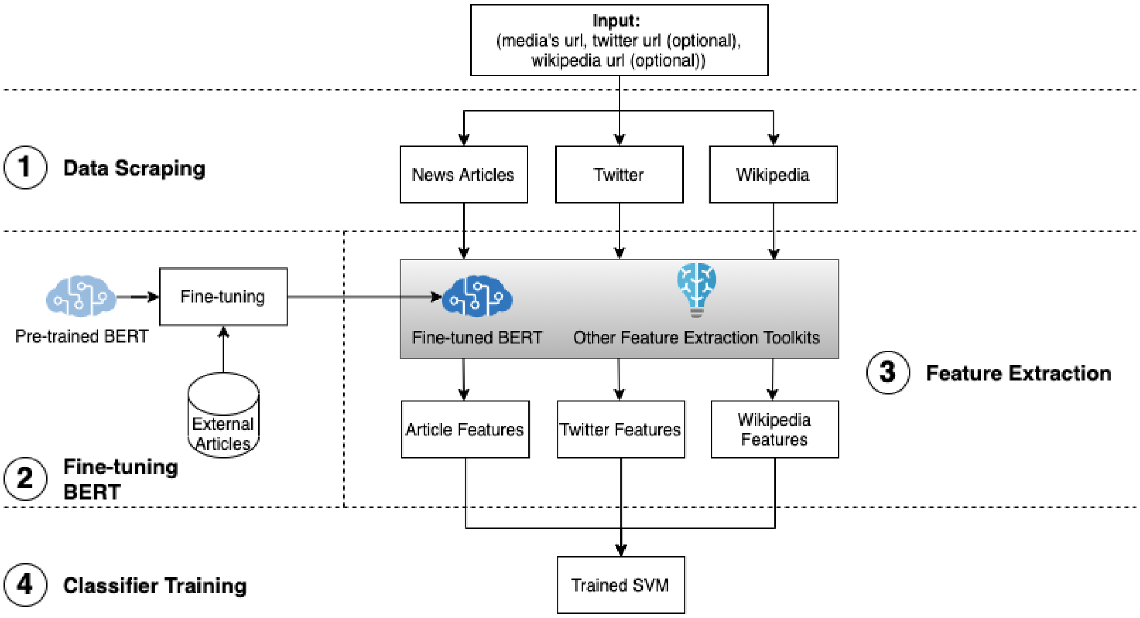


Figure 3-1: An outline of the training process for the bias and factuality prediction system. First, we scrape relevant language data from the URL's of the media sources in our dataset, and fine-tune BERT models to extract features specific to the tasks of bias and factuality prediction. Next, we use these fine-tuned BERT models, among other pre-built tools, to extract features from the scraped language data. Finally, we train Support Vector Machines to predict bias and factuality using these extracted features.

Given the data described in the previous section, we can train the necessary models for the bias and factuality prediction system through the process outlined in Figure

3-1. As an overview of the process, we first scrape news articles, Twitter profiles, and Wikipedia pages from the provided media sources' URL's. Next, we fine-tune BERT models for predicting either bias or factuality from news articles, and use these fine-tuned BERT models, among other pre-built toolkits, to extract features from our scraped textual data. Finally, we use these extracted features to train Support Vector Machine (SVM) classifiers for both tasks of bias and factuality prediction. We describe this process in further detail in the next few sections, and break it down into four main steps which are reflected in Figure 3-1. These steps are: 1) Data scraping 2) Fine-tuning BERT 3) Feature Extraction 4) Classifier training.

3.2.1 Data Scraping

In Section 3.1, we described the necessary inputs for our bias and factuality prediction system - the media sources' URL's, and optionally the media sources' Twitter and Wikipedia URL's. However, since our system's models operate on natural language, the first step of our system must be to retrieve forms of natural language from the input URL's.

From the media sources' URL's, we scrape news articles that the source has published using the `newspaper3k`[1] library. We further process each scraped article to determine if they are suitable for predicting bias and factuality. Specifically, we focus on articles containing political vocabulary, as these tend to be more representative of a media source's bias and factuality levels, and articles that we determine as non-political are discarded. The other main condition that articles must satisfy is that they are of some minimum length in order to have enough language for our models to extract meaningful features from. After we scrape and filter articles from a media source, we then decide if we have obtained enough information to include the media source in the training process. We believe it is possible to have a set of articles that are not representative of a media source if we have less than five articles for

[1] The `newspaper3k` library provides functionality to crawl a media source's website and download articles from it that have been cleaned of any formatting and HTML (i.e. condensing information down to natural language only).

the source. So, if a media source does not meet this criteria, we remove it from our dataset. From the original 1,737 sources obtained from MBFC, we have 1,197 sources remaining after filtering. The class distributions amongst this dataset can be seen in Table 3.2.

Bias		**Factuality**	
Left	405	Low	194
Center	363	Mixed	289
Right	429	High	714

Table 3.2: Our dataset's distribution of classes for bias and factuality

After filtering our dataset, we then move on to scraping the Twitter and Wikipedia pages for the media sources that have them[2]. From Twitter, we scrape the media sources' self-written descriptions, as well as other account metadata (e.g. is it verified, how many followers, how many posts, etc.). Additionally, we scrape the bios of the media sources' followers. From Wikipedia, we scrape all the textual information written about the media source. Across the 1,197 sources in our dataset, we obtained Twitter data for 82.1% of them (983 sources), and Wikipedia data for 67.5% of them (808 sources).

3.2.2 Fine-tuning BERT

Once the necessary data has been retrieved, we move on to preparing the tools that will be used to extract features from this data. Similar to the methodology presented by Baly et al., we make use of BERT (Devlin et al., 2019) to extract features from the text we have collected. BERT has achieved state of the art performance across a wide range of natural language understanding tasks, and has been shown to produce powerful contextualized embeddings of language in deeper layers of the model (Ethayarajh, 2019). For this reason, we believe BERT offers a valuable avenue to featurize our text data. Furthermore, we also have the ability to adapt BERT to the tasks of bias or factuality prediction, which will provide us with features that are even more

[2] We use the `python-twitter` library for scraping Twitter and the `wikipedia` library for scraping Wikipedia. Both libraries are python wrappers around the original organizations' APIs.

specialized to these tasks.

We adapt our BERT model to the task of bias or factuality prediction through a process called fine-tuning. During fine-tuning, we train BERT to predict either the bias or factuality of individual news articles (note that fine-tuning is at the article-level, not the source-level), and since we are interested in both tasks (bias and factuality) we fine-tune two different BERT models - one for each task. To fine-tune BERT using article-level data, we train the model with an additional linear layer and softmax layer on top of the [CLS] token that is output from the final transformer as seen in Figure 3-2. As input, we feed the first 510 WordPieces (BERT's tokenization method) of each article. For the task of bias prediction, we train BERT to classify each article with the same three labels we are using for source classification - left, center, and right. However, for predicting article-level factuality, we train BERT to classify each article simply as high or low factuality - Baly et al. mention that the mixed label does not make sense in the context of a single article.

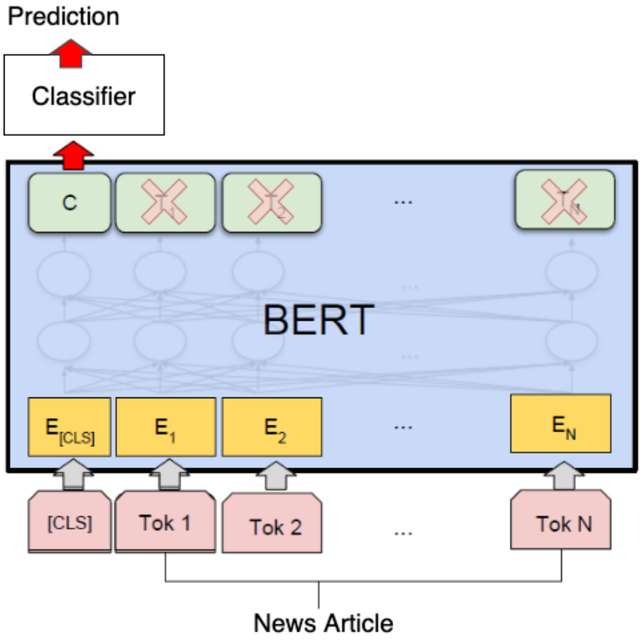


Figure 3-2: Diagram of the fine-tuning architecture use to adapt BERT to the tasks of bias and factuality prediction. We train a linear layer on top of the [CLS] embedding output from the 12 transformer encoder layers of the pre-trained BERT model (Devlin et al., 2019).

In order for the features produced by the fine-tuned BERT models to generalize well, we must draw the news articles for fine-tuning, and their corresponding labels, from a dataset external to the sources used for training our final classifiers. There are several options available for this dataset, and we find in our experiments that the fine-tuning process has a significant impact on the system's predictive ability, with the main variations in the process being a result of our choice of dataset. The datasets we try in our experiments include a partition of our 1,197 collected sources, data from the Hyperpartisan SemEval task (Kiesel et al., 2019), and data from Allsides.com (AllSides, 2020). Our labels for the first two options are derived using distant supervision, where articles are assigned the same label as the media source they come from (a common method for labeling large news datasets (Nørregaard et al., 2019)), and the labels for the latter option are on the article-level - Allsides labels the data they provide at the article-level. The resulting differences between using these datasets for fine-tuning are described during our experiments in Section 3.3.

3.2.3 Feature Extraction

After fine-tuning our BERT models, we move on to the feature extraction process, which transforms the textual data we have scraped into numeric features that can be used to train our final SVMs. In addition to the BERT (Devlin et al., 2019) models we have fine-tuned for bias and factuality prediction, we make use of other feature extraction toolkits out-of-the-box. These include the News Landscape Toolkit (NELA) (Horne et al., 2018b), Global Vectors for Word Representation (GloVe) (Pennington et al., 2014), and Sentence-Bert (SBERT) (Reimers and Gurevych, 2019). We describe each of these toolkits in further detail, as well as the feature extraction process for each type of data, below.

News Articles

For extracting features from news articles published by a media source, we not only make use of NELA and BERT, used in the work by Baly et al. (Baly et al., 2020) that our methods are based on, but also use GloVe, which is used in a previous work by Baly et al. (Baly et al., 2018a). From each article, the NELA toolkit computes features relating to structure, sentiment, topic, complexity, bias, and morality. After computing these for each article, we average the extracted NELA features across all articles for a specific media source to move from article-level features to the source-level (we are interested in predicting bias and factuality of entire media sources).

To retrieve features from our BERT models, trained for bias or factuality classification, we average the word representations extracted from the second-to-last layer of BERT - Baly et al. cite this as common practice as the final layer may be biased to BERT's pre-training objectives. Furthermore, to translate from the article-level to the media source-level, we again average the retrieved embeddings across all articles we have gathered for a specific media source.

GloVe, like BERT, is a method for extracting embeddings from our news articles. However GloVe provides static embeddings, whereas BERT's embeddings are contextual. In their most recent work, Baly et al. state that using BERT embeddings *instead* of GloVe embeddings produced a large boost in performance (Baly et al., 2020), but we decide to include the static embeddings from GloVe in our experiments to see if there is complementary information between the two embedding types. Just like the processes of extracting features using NELA and BERT, GloVe provides embeddings at the article-level, which we then average across all articles from a media source.

Twitter

Within our Twitter data, we have two different sets of information: the media source's own Twitter profile, and the bios of the media source's followers. From the media source's Twitter profile, we first extract some metadata about the profile including if it is verified, its followers and friends count, how many posts have been written and favorited, where and when it was created, etc. - these are the only non-linguistic features used in our system. Additionally, we embed the bio of the media source's account using SBERT. Unlike in the case of news articles, there is not enough data among the media sources' bios to fine-tune a BERT model to produce task specific embeddings. So instead, we rely on a sentence-level embedding produced by the pre-trained SBERT model. Regarding the bios of a media source's followers, we do have enough data to fine-tune BERT, but we run into a different issue. The distant supervision we used to retrieve bias and factuality labels for news articles is ill-suited for classifying Twitter bios - there is much more variance on whether or not a media source's followers agree with the source's views. As a result, we again embed these bios using SBERT, and average them across a media source's followers.

Wikipedia

The process for extracting features from a media source's Wikipedia page is identical to embedding news articles using BERT. Wikipedia pages tend to have a similar structure as the news articles written by the media sources. So, we embed the Wikipedia pages using the BERT models that are fine-tuned for bias and factuality prediction of news articles.

3.2.4 Classifier Training

After extracting numeric features from each of our data channels, we can now train our two SVMs - one to classify media sources' bias, and the other to classify media sources' factuality. As input, we concatenate the extracted features from all or a subset of our data channels (described in more detail in Section 3.3), using the features extracted from BERT fine-tuned for bias prediction for training the bias classifier, and the features extracted from BERT fine-tuned for factuality prediction for training the factuality classifier.

Following the methodology from Baly et al., we train and evaluate our SVMs using 5-fold cross-validation, and perform a grid search over the cost parameter, C, and the kernel coefficient, γ, at each step of cross-validation. Finally, we select the best SVMs using the macro-F1 score - the F1 score averaged across each class, as both the bias and factuality datasets are not balanced.

3.3 Experiments and Results

During development of this system, we try a few variations of our training process, and we report our findings on them in this section. Our experiments broadly fit into two groups. The first experiment we discuss relates to the dataset used in fine-tuning our BERT models for feature extraction - we choose to vary this step because our BERT models produce the most informative features for bias and factuality prediction. Our second experiment determines which configuration of features results in the best performing classifier for either bias or factuality prediction, and is helpful for selecting which features to use in our final system.

Dataset Variation for Fine-tuning

As mentioned in Section 3.2.2, the datasets we use to fine-tune BERT include a partition of our scraped MBFC dataset, the Hyperpartisan SemEval dataset, and the Allsides.com dataset. However, for the task of factuality prediction, we are limited to using the first two datasets, as Allsides does not provide factuality labels. We also note that for generalization purposes, we exclude any media sources used in fine-tuning from the data used to train the final classifiers.

To test which dataset was most useful during the fine-tuning process, we fine-tune BERT models on each dataset of interest for the task of bias or factuality prediction, and then measure how useful each BERT model is on the downstream task of source-level prediction. Although we can measure the performance of the article-level bias or factuality predictions during fine-tuning, we choose to measure the performance on the downstream SVM classifiers because it directly aligns with our goal of source-level bias or factuality prediction. Furthermore, measuring performance at the article-level during fine-tuning may be misleading, as it is possible that BERT may overfit the articles and media sources it is fine-tuned on, and not generalize well when extracting features for unseen media sources. So, in order to examine the differences between datasets used for fine-tuning, we fine-tune a BERT model on each dataset of interest, carry out feature extraction from articles using each BERT variation, and finally train the bias or factuality classifier using the extracted features from each model. We report the macro-F1 and accuracy of each variation in Tables 3.3 and 3.4.

Fine-tuning Dataset	**Macro-F1**	**Accuracy**
MBFC Partition	74.57	74.53
Hyperpartisan SemEval	76.13	76.15
Allsides	70.20	70.30

Table 3.3: Source-level **political bias prediction** results with features extracted from articles using BERT models fine-tuned on different datasets.

Table 3.3 reports the fine-tuning variation results for political bias prediction, and we observe that fine-tuning on the Hyperpartisan SemEval dataset significantly outperforms the others. We hypothesize that the differences in performance are due to the number of different sources present in each fine-tuning set. The work from Baly et al. reports using ∼30, 000 articles from 298 sources for fine-tuning, and we aimed to use a similar amount of articles in our fine-tuning processes, but this resulted in differing numbers of sources. Our dataset variations have the following statistics:

- **MBFC Partition:** ∼30, 000 articles, ∼700 sources
- **Hyperpartisan SemEval:** ∼27, 000 articles, ∼130 sources
- **Allsides:** ∼35, 000 articles, ∼70 sources

We believe that the BERT model fine-tuned on the Hyperpartisan SemEval dataset is able to extract more generalizable features to the unseen sources at classification time than the Allsides dataset, as there is a larger variety of sources present during fine-tuning. The reason we do not observe the same trend between the Hyperpartisan SemEval dataset and MBFC partition, is due to the exclusion of fine-tuning sources during classifier training. When it comes to training the SVM for the MBFC partition, we hypothesize that excluding these 700 sources from fine-tuning significantly reduces the information the classifier receives during training. As a result, the Hyperpartisan SemEval dataset strikes a better balance than the other variations between fine-tuning a generalizable BERT model, and providing enough sources to train the SVM.

Fine-tuning Dataset	**Macro-F1**	**Accuracy**
MBFC Partition	60.97	70.94
Hyperpartisan SemEval	61.34	70.99

Table 3.4: Source-level **factuality prediction** results with features extracted from articles using BERT models fine-tuned on different datasets.

Table 3.4 reports the fine-tuning variation results for factuality prediction, and though the Hyperpartisan SemEval dataset still performs the best, we observe a much smaller gap in performance than we did for bias prediction. We also believe the cause of this smaller gap is due to the amount of sources present in the datasets. Recall from Section 3.2.2 that we fine-tune BERT to predict either high or low for article-level factuality. As a result, we exclude the sources from our datasets that were labeled as mixed. Though Baly et al. do not report their dataset statistics after this filtering, ours are as follows:

- **MBFC Partition:** $\sim$23, 000 articles, $\sim$540 sources
- **Hyperpartisan SemEval:** $\sim$18, 000 articles, $\sim$90 sources

We hypothesize that removing sources from the Hyperpartisan SemEval dataset reduces the generalizability of the BERT model fine-tuned on it, but it still ends up performing slightly better than the MBFC partition due to having more sources to train the SVM.

Feature Ablation

To determine which combination of features extracted from our data channels (articles, Twitter, and Wikipedia) results in the best classifiers for bias and factuality, we conduct ablation studies for bias prediction and for factuality prediction (similar to Baly et al.). We first examine which features and feature combinations are most useful within each data channel. The combinations of article features are most interesting to us, as we are curious about the impact of combining the static embeddings from GloVe and the contextual embeddings from BERT (as mentioned in Section 3.2.3). After examining the features within each channel, we then combine the best performing features from each data channel to see if they contain complementary information and result in a better classifier. Finally, we compare the results of our classifiers trained on different feature combinations to the results of the most comparable configurations presented by Baly et al. - our results differ from theirs and we give hypotheses on possible causes.

		Our Results		**Baly et al. Results**	
#	**Features**	**Macro-F1**	**Accuracy**	**Macro-F1**	**Accuracy**
1	Articles: NELA	66.34	66.48	64.82	68.18
2	Articles: BERT	76.13	76.15	79.34	79.75
3	Articles: GloVe	68.56	68.64	N/A	N/A
4	Articles: BERT+GloVe	75.45	75.49	N/A	N/A
5	Articles: BERT+NELA	74.82	74.84	81.00	81.48
6	Articles: BERT+GloVe+NELA	74.33	74.37	N/A	N/A
7	Twitter: Profile	49.93	50.23	59.23	60.88
8	Twitter: Followers	65.55	66.29	62.85	65.39
9	Twitter: Profile+Followers	62.98	63.66	N/A	N/A
10	Wikipedia: BERT	53.47	54.37	64.36	66.09
11	**A+T: rows 2 & 8**	**76.38**	**76.81**	**84.28**	**84.87**
12	A+W: rows 2 & 10	75.12	75.21	81.53	81.98
13	A+T+W: rows 2, 8 & 10	75.61	75.77	83.53	84.02

Table 3.5: An ablation study of our features' predictive ability on the task of **political bias prediction**. In rows 11-13, we combine the best performing features from each data channel - (A) stands for articles, (T) stands for Twitter, and (W) stands for Wikipedia. We also include the results from Baly et al. (Baly et al., 2020) for comparison (some of our features behave differently than Baly et al. but we still report their best reported results in rows 11-13).

Table 3.5 shows the results from our ablation study on political bias prediction and the comparable results from Baly et al.. In rows 1-10, we try features and feature combinations within each data channel, and in rows 11-13, we combine the best feature combinations from each channel.

Within the features extracted from articles, we find that only using BERT-extracted features produces the best results (row 2), and that GloVe features, whether by themselves, or in conjunction with BERT, result in worse performance (rows 3-6). This leads us to believe that for the task of political bias prediction, GloVe embeddings do not offer any new information on top of what is provided by the BERT embeddings. Our findings do agree with Baly et al. that the article data channel provides the most predictive features for political bias. However, we contrastively observe that

NELA features worsen performance when used in conjunction with BERT (rows 2 and 5), and that our BERT features significantly underperform their results (row 2). We believe that differences in the dataset used for fine-tuning BERT are the cause of both of these discrepancies. Though we do not know which specific dataset Baly et al. used to fine-tune BERT, we do know that they had over 160 additional sources present in their fine-tuning set without significantly reducing the information available for training their classifier. The BERT model fine-tuned on their dataset may have generalized better than ours as a result, and could have potentially diverged from extracting similar information as the NELA toolkit.

Regarding the Twitter data channel, we find that features from the media sources' Twitter followers are much more informative than the media sources' own Twitter pages. The reason these features behave differently could be due to a high variance within Twitter profiles. While each media source's own profile may or may not express their bias with high variance, this variance becomes much lower when averaging across all the Twitter profiles of a media source's followers. If this is the case, it also provides a potential explanation on why our Twitter profile features underperform Baly et al. (row 7) while our Twitter follower features outperform. Baly et al. collected Twitter information for 72.5% of their media sources, whereas we collected it for 82.1%. The additional high variance information of the media source's Twitter profiles may have further hurt performance, while the low variance, averaged media source Twitter followers, may have helped performance.

The only observation regarding our Wikipedia data channel is that it underperforms Baly et al.'s results. Since we use the same fine-tuned BERT model to extract features from Wikipedia as we used to extract features from articles, we believe that our lower performance is also due to a difference in the dataset used for fine-tuning.

When testing feature combinations across data channels, we always included our best performing article features (row 2), as they were by far the most predictive features. We first tried combining article and Twitter features (row 11), which resulted in the best performing classifier of political bias - the same is reported by Baly et al.. Although it is not reflected in our table, Baly et al. also found that classification only benefited from using Twitter follower features and not from media source's Twitter profile features. The additional feature combinations we tried (rows 12 and 13) follow similar trends to those reported by Baly et al. and we attribute the performance gaps to the causes we state for why individual data channels underperform.

		Our Results		**Baly et al. Results**	
#	**Features**	**Macro-F1**	**Accuracy**	**Macro-F1**	**Accuracy**
1	Articles: NELA	53.65	66.20	55.54	62.62
2	Articles: BERT	61.34	70.99	61.46	67.94
3	Articles: GloVe	60.09	70.89	N/A	N/A
4	Articles: BERT+GloVe	63.20	73.05	N/A	N/A
5	Articles: BERT+NELA	62.22	72.39	59.34	64.82
6	**Articles: BERT+GloVe+NELA**	**63.40**	**73.05**	N/A	N/A
7	Twitter: Profile	42.35	61.31	49.96	56.71
8	Twitter: Followers	49.10	64.51	42.19	58.45
9	Twitter: Profile+Followers	55.11	65.73	N/A	N/A
10	Wikipedia: BERT	39.87	62.44	45.74	55.32
11	A+T: rows 6 & 9	61.10	71.92	65.45	70.40
12	A+W: rows 6 & 10	61.37	71.92	**67.25**	**71.52**
13	A+T+W: rows 6, 9 & 10	62.75	72.77	64.14	69.36

Table 3.6: An ablation study of our features' predictive ability on the task of **factuality prediction**. In rows 11-13, we combine the best performing features from each data channel - (A) stands for articles, (T) stands for Twitter, and (W) stands for Wikipedia. We also include the results from Baly et al. (Baly et al., 2020) for comparison (some of our features behave differently than Baly et al. but we still report their best reported results in rows 11-13).

Table 3.6 shows the results from our ablation study on factuality prediction and the comparable results from Baly et al.. Similar to our political bias ablation study, we try features and feature combinations within each data channel in rows 1-10, and combine the best features from each channel in rows 11-13. Since our factuality dataset is unbalanced, we make our comparisons between feature combinations using macro-f1 as our performance metric.

Similarly to Baly et al. and our political bias prediction results, we also find that our news article data channel provides the most informative features for factuality prediction. In our study, a combination of all the article features (BERT, GloVe, and NELA) produces the best classifier (row 6). Though our features extracted using BERT still provide the best individual results (row 2), we observe that both NELA and GloVe improve results (rows 4-6), unlike in our political bias prediction study.

We believe that this is because our BERT model, fine-tuned for factuality prediction, produces less powerful contextual embeddings than in the bias case - recall that our factuality fine-tuning dataset contains fewer articles and sources than our bias fine-tuning dataset. As a result, the NELA features and static, GloVe embeddings are able to contribute new information to the classifier. We also note that our finding of BERT and NELA features being additive (row 5) contradicts the results from Baly et al., but we attribute that, as well as our better overall performance using article features, to our BERT model learning different features than Baly et al.'s model during fine-tuning.

Our results regarding the Twitter data channel trend similarly to those we reported in our political bias study, with the media sources' Twitter profiles being less informative than the media sources' Twitter followers - we again believe that Twitter profile variance is the cause. We do see, though, that Twitter generally produces less informative features for factuality prediction than it had for bias prediction.

The Wikipedia channel significantly underperforms relative to Baly et al.'s results. Fine-tuning differences could again be the cause, like in the political bias prediction case, but we are more skeptical of this as our results from BERT in the article data channel are comparable to Baly et al. (row 2). Another possible cause for this performance difference could be similar to the what we theorize for the Twitter case. Specifically, how these Wikipedia pages describe media sources may have a high variance in whether or not they offer any predictive information. As a result, different cross-validation splits may result in significantly different performance.

Unlike in Baly et al.'s ablation study, combining features across data channels only hurts performance relative to article-only features in our case, and our best performing classifier only uses a combination of all article features. The differences we cited for Twitter and Wikipedia discrepancies could be the cause.

3.4 Making Predictions

After going through the process of training the necessary models and selecting the most informative features for the bias and factuality prediction system, we can now use it to predict on unseen sources. Making predictions on new data is very similar to the training process, except instead of fine-tuning new BERT models, or training new SVMs, we use the models already trained on our labeled datasets. First, our system takes a media source's homepage URL, and optionally its Twitter and Wikipedia URL's, as input, and uses these to scrape the relevant natural language data. Next, we use the BERT models fine-tuned for bias and factuality, and the other best-performing toolkits, to extract features from the scraped data. Finally, these features are concatenated and fed through the bias or factuality SVM classifiers to obtain the final predictions.

Chapter 4

Rationalization

During our work in building a system to make source-level political bias and factuality predictions, we noted that the field of fake news detection using language is becoming decreasingly transparent because the newer, deep-learning based approaches are more obfuscated than their lexicon-based predecessors. However, we believe that model explainability, especially when it comes to fake news detection, may be crucial for automated systems to be seen as credible in the eyes of the public. As a result, though the rationalization methods we explore are not directly linked to the system we describe in Chapter 3, we are hopeful that our analysis will encourage the field of fake news detection to maintain its efforts in transparency.

Our study in rationalization focuses on extractive rationales, or rationales that explain a model's predictions using selected subsets of the input data. Unlike our system that operates on the source-level, we choose to focus on article-level predictions for our rationalization study. We reason that there is too much information being aggregated at the source-level to offer an extractive rationale that is understandable to a user, but at the article-level, we can, much more concretely, highlight specific language that contributes the most to the prediction being made.

For our article-level predictions, we use a BERT model that is fine-tuned using the same process described in Section 3.2.2. Instead of extracting embeddings from this model, like we did in our source-level system, we simply use the article-level predictions from BERT itself. Additionally, since we are now working at the article-level for our predictions, we believe it is important to use article-level data for training our model. So, we fine-tune our BERT model using the article-level labelled political bias data from Allsides (AllSides, 2020), and ignore the datasets we obtained using distant supervision when training our source-level system. Using Allsides data restricts our rationalization study to the task of political bias prediction (Allsides does not contain article-level factuality labels), but we believe that politically biased language lends itself better to the qualitative analysis we perform anyways, and note that the same extractive rationale methods we explore could be used for factuality prediction, given the proper dataset.

4.1 Analysis of BERT's Attention

We first explore extractive rationalization through an analysis of the attention heads that compose BERT. Attention heads, intuitively, determine the relative weighting of inputs that create the most informative outputs for a downstream task (in this case political bias prediction). So, it is possible to interpolate which input words within an article are most heavily weighted by each attention head. Our assumption is that during fine-tuning, attention heads will learn to highly weight the words within articles that are most relevant to the task of political bias prediction, and the most heavily weighted words could serve as a viable explanation for the model's prediction. Examples of what these explanations might look like can be seen in Figures 4-1 and 4-2, where words in dark red have been more heavily weighted by a specific attention head - this article was predicted as right-leaning.

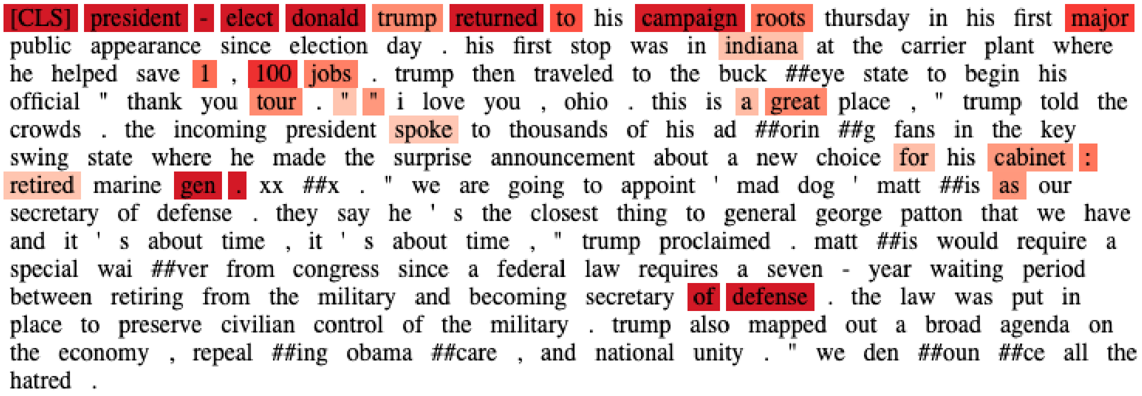
For a given article we wish to classify and extract a rationale from, we feed the maximum amount of possible tokens (512)[1], starting from the beginning of the article, into BERT. As the input article is run through BERT, each attention head computes 512 attention weights for each of the 512 tokens in the sequence (a total of 512x512 weights), and we save these computed weight values at each attention head in BERT.

[1] BERT can operate on a maximum of 512 tokens, and since we need to include two special tokens ([CLS] and [SEP]) at the start and end of our input, we can use 510 tokens from an article.

Analyzing 512 weights for each of the 512 tokens in our sequence is infeasible though, as this data contains too much information to provide a suitable explanation. However, we believe it is sufficient to only look at the weights of one specific token in our input sequence, the special [CLS] token (1x512 weights). Recall from Section 3.2.2 that BERT is fine-tuned for classification by training a fully-connected layer on top of the [CLS] embedding representation output from BERT's transformer encoders. Since this [CLS] representation is the only information the final layer in our classifier uses to make its prediction, we believe that the attention weights used to calculate the representation of the [CLS] token contain the most relevant information in making political bias predictions.

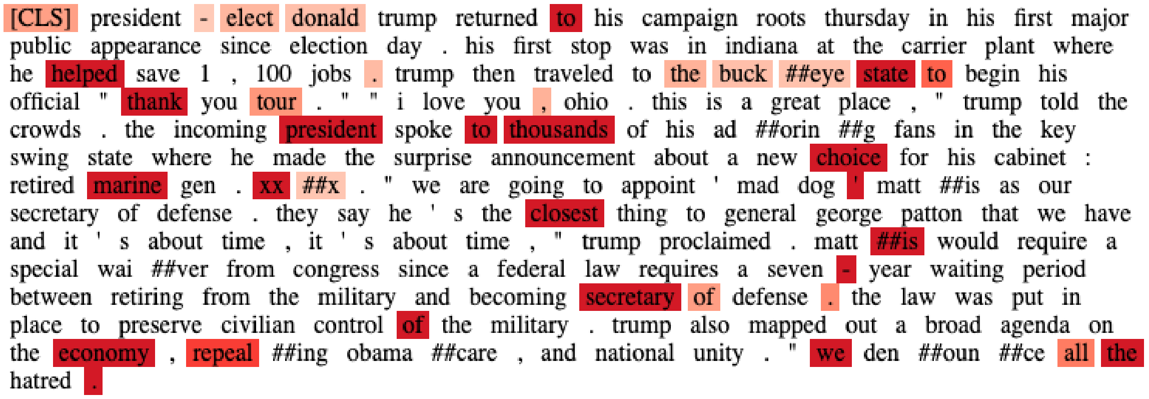
Though we have retrieved the attention weights from BERT and reduced them to an amount more tractable for analysis, there is still an abundance of information, as we obtain weights from each of BERT's 144 attention heads (there are 12 attention heads in each of BERT's 12 transformer encoder layers). However, other works analyzing BERT's attention heads have shown that different heads learn to specialize to specific components of syntax (Clark et al., 2019) - some heads focus on structure, co-reference resolution, etc., while others specialize in semantics. Though structure and other syntactical nuances may play a role in BERT's predictions in political bias, we focus on finding heads specializing in semantics, as these convey the most human-understandable information.

While our visualizations across all attention heads on articles from different political ideologies is too much information to include, we show some examples of our analysis for an article classified as right-learning by our model. The trends we highlight for this example article hold for different articles we analyzed as well as across the right, center, and left political ideologies. In Figures 4-1 and 4-2, we visualize heavily weighted words from attention heads selected from the middle layer and last layer of BERT respectively. We choose to show attention heads from different layers to compare how different layers of the model behave, and selected these specific heads because the words they highly weight are semantically aligned with the criteria Allsides uses for labelling right-learning articles. Specifically, defense-oriented words



[CLS] president - elect donald trump returned to his campaign roots thursday in his first major public appearance since election day . his first stop was in indiana at the carrier plant where he helped save 1 , 100 jobs . trump then traveled to the buck ##eye state to begin his official " thank you tour . " " i love you , ohio . this is a great place , " trump told the crowds . the incoming president spoke to thousands of his ad ##orin ##g fans in the key swing state where he made the surprise announcement about a new choice for his cabinet : retired marine gen . xx ##x . " we are going to appoint ' mad dog ' matt ##is as our secretary of defense . they say he ' s the closest thing to general george patton that we have and it ' s about time , it ' s about time , " trump proclaimed . matt ##is would require a special wai ##ver from congress since a federal law requires a seven - year waiting period between retiring from the military and becoming secretary of defense . the law was put in place to preserve civilian control of the military . trump also mapped out a broad agenda on the economy , repeal ##ing obama ##care , and national unity . " we den ##oun ##ce all the hatred .

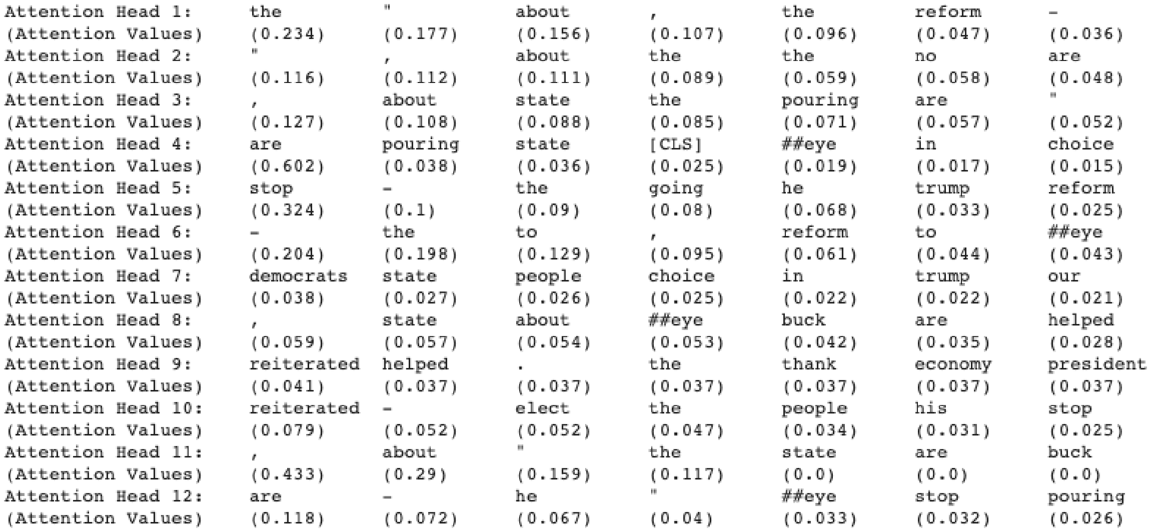
Figure 4-1: Attention visualization from 1st attention head in layer 6 on an article classified as right-leaning (this is a truncated version of the full 512 token input fed into BERT). This attention head highly attends to "president - elect donald trump" and "defense" which align with the right-leaning criteria used by Allsides.



[CLS] president - elect donald trump returned to his campaign roots thursday in his first major public appearance since election day . his first stop was in indiana at the carrier plant where he helped save 1 , 100 jobs . trump then traveled to the buck ##eye state to begin his official " thank you tour . " " i love you , ohio . this is a great place , " trump told the crowds . the incoming president spoke to thousands of his ad ##orin ##g fans in the key swing state where he made the surprise announcement about a new choice for his cabinet : retired marine gen . xx ##x . " we are going to appoint ' mad dog ' matt ##is as our secretary of defense . they say he ' s the closest thing to general george patton that we have and it ' s about time , it ' s about time , " trump proclaimed . matt ##is would require a special wai ##ver from congress since a federal law requires a seven - year waiting period between retiring from the military and becoming secretary of defense . the law was put in place to preserve civilian control of the military . trump also mapped out a broad agenda on the economy , repeal ##ing obama ##care , and national unity . " we den ##oun ##ce all the hatred .

Figure 4-2: Attention visualization from 7th attention head in layer 12 on an article classified as right-leaning (this is a truncated version of the full 512 token input fed into BERT). This attention head highly attends to "president", "marine", "state", and "economy" which align with the right-leaning criteria used by Allsides.

like marine and defense are highly weighted, as well as the name of the right-learning presidential elect (at the time), Donald Trump. However, between the different layers of the model, we observe a significant difference in which words are heavily weighted. In fact, even when comparing semantically-oriented attention heads within the same layer we see a similar gap between which words are weighted. Figure 4-3 shows the 7 highest weighted tokens from each attention head in the last layer, and it shows that there is a high variation between which words are highly attended to, and the magnitude of those weightings.



```
Attention Head 1:       the         "           about       ,           the         reform      -
(Attention Values)      (0.234)     (0.177)     (0.156)     (0.107)     (0.096)     (0.047)     (0.036)
Attention Head 2:       "           ,           about       the         the         no          are
(Attention Values)      (0.116)     (0.112)     (0.111)     (0.089)     (0.059)     (0.058)     (0.048)
Attention Head 3:       ,           about       state       the         pouring     are         "
(Attention Values)      (0.127)     (0.108)     (0.088)     (0.085)     (0.071)     (0.057)     (0.052)
Attention Head 4:       are         pouring     state       [CLS]       ##eye       in          choice
(Attention Values)      (0.602)     (0.038)     (0.036)     (0.025)     (0.019)     (0.017)     (0.015)
Attention Head 5:       stop        -           the         going       he          trump       reform
(Attention Values)      (0.324)     (0.1)       (0.09)      (0.08)      (0.068)     (0.033)     (0.025)
Attention Head 6:       -           the         to          ,           reform      to          ##eye
(Attention Values)      (0.204)     (0.198)     (0.129)     (0.095)     (0.061)     (0.044)     (0.043)
Attention Head 7:       democrats   state       people      choice      in          trump       our
(Attention Values)      (0.038)     (0.027)     (0.026)     (0.025)     (0.022)     (0.022)     (0.021)
Attention Head 8:       ,           state       about       ##eye       buck        are         helped
(Attention Values)      (0.059)     (0.057)     (0.054)     (0.053)     (0.042)     (0.035)     (0.028)
Attention Head 9:       reiterated  helped      .           the         thank       economy     president
(Attention Values)      (0.041)     (0.037)     (0.037)     (0.037)     (0.037)     (0.037)     (0.037)
Attention Head 10:      reiterated  -           elect       the         people      his         stop
(Attention Values)      (0.079)     (0.052)     (0.052)     (0.047)     (0.034)     (0.031)     (0.025)
Attention Head 11:      ,           about       "           the         state       are         buck
(Attention Values)      (0.433)     (0.29)      (0.159)     (0.117)     (0.0)       (0.0)       (0.0)
Attention Head 12:      are         -           he          "           ##eye       stop        pouring
(Attention Values)      (0.118)     (0.072)     (0.067)     (0.04)      (0.033)     (0.032)     (0.026)
```

Figure 4-3: The top 7 weighted tokens, and their magnitudes, from each attention head in the final layer of BERT. Between attention heads we see a large amount of variation besides a few common words like "about", "reform", and "state".

Due to the variance we see between attention heads, we believe that selecting specific semantically-oriented attention heads from BERT does not serve as a good explanation for the model's prediction. However, we still believe there may be some merit in analyzing BERT's attention for extractive rationales. If, instead of comparing attention weightings between specific heads, we use a more global view of attention on the [CLS] token (across all heads and layers), we can see some trends emerge. In Figure 4-4, we plot all 144 attention head's weights for the [CLS] token, and observe that later layers of the model do attend to the same tokens - this is shown by the horizontal streaks. Though individual attention heads may attend to different subgroups of these tokens, which is potentially why we observe large differences when comparing individual attention heads, there seem to be some agreed upon important tokens in aggregate. Furthermore, we confirm that these agreed upon tokens for this example (trump, marine, conservative, etc.) are aligned with the right-learning criteria specified by Allsides, and we believe that more studies on the aggregate attention of BERT could be useful in determining a more robust method for extractive rationalization than analyzing individual attention heads.

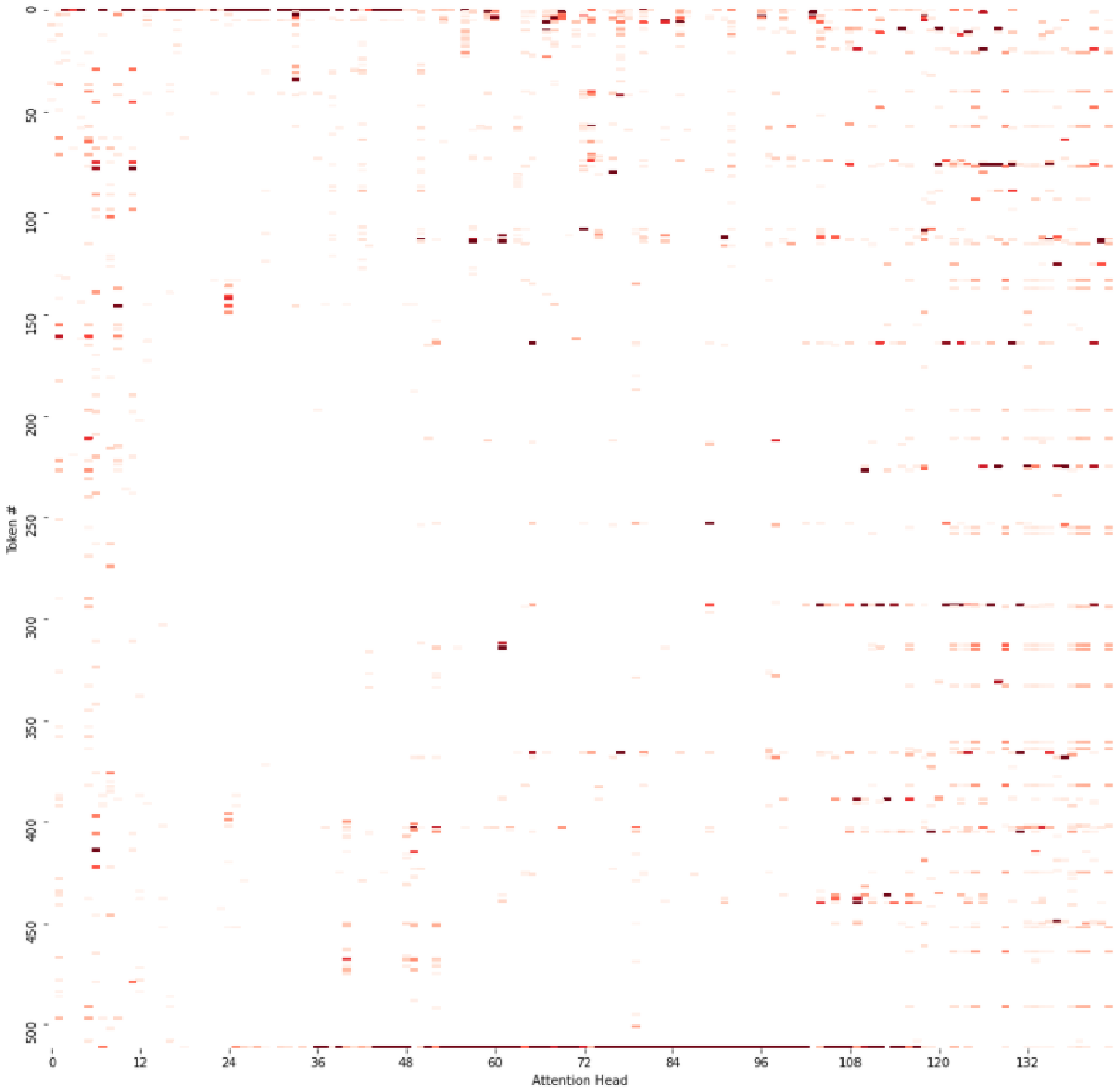


Figure 4-4: [CLS] token's attention weights across all 144 attention heads within BERT. In later layers, we observe some horizontal streaks, indicating that the same tokens are being heavily weighted across different attention heads.

While our study shows using an analysis on attention has potential for extractive rationales, there are some drawbacks to the methodology as well. Not only are individual attention heads poor avenues for rationalization because of their variance and the work required to filter through 144 different attention heads, but there is also no guarantee that the same attention heads will be useful between different articles. As a result, we mentioned that viewing BERT's attention heads in aggregate could help alleviate these issues. However, even in aggregate, it is possible the attention may not be suitable for extracting rationale, as the relationship between attention weights and model predictions is still a debated topic. Jain and Wallace report that attention weights are not correlated with other measures of input importance and that perturbing attention weights does not significantly alter predictions (Jain and Wallace, 2019). However, Wiegreffe and Pinter refute these claims in their own experiments, and though they do not confirm that attention is useful for explanation, they show that the experiments by Jain and Wallace do not disprove its use (Wiegreffe and Pinter, 2019). So, until more research on attention's faithfulness to model predictions is conducted, other extractive rationalization methods are more preferable.

4.2 Gradient-Based Methods

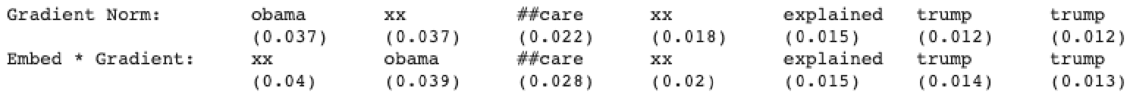
The other methodology we explore for extractive rationalization, gradient saliency, avoids the drawbacks of attention-based approaches and is a more direct measure of input importance. To quantify gradient saliency, we use two different methods, both based on calculating the gradients of our model's outputs with respect to its inputs, and then processing these gradients to retrieve a single importance score per token. The computed importance scores represent how much changing each token impacts the model's output, with the intuition behind using these scores for extractive rationalization being that tokens which can cause large changes in a model's output must be important in making its prediction. An additional advantage for gradient-based approaches is that they provide a single measure of how important each input token is, as opposed to the many channels of attention weights that need to be analyzed in attention-based approaches.

We will refer to the two highly-related, gradient saliency methods we explore for rationalizing political bias predictions as the *gradient norm* and *embed $\cdot$ gradient*. In order to calculate importance scores using either approach, we must define which model outputs we calculate the gradient of, and which input representations we calculate gradients with respect to - we use the same output and input definitions for both approaches. For rationalization we choose to measure our output using the maximum value of the unnormalized class probabilities (also known as logits) output from our BERT classifier. Though it is common to use loss as a measure of output when calculating gradients (as we do during fine-tuning BERT), we believe calculating gradients of the maximum logit with respect to the model's input is more useful for rationalization. Whereas the gradient of the loss measures how changing inputs can *improve* the model's performance, the gradient of the maximum logit measures how much each input changes a model's confidence in its current prediction. As a result, we believe using the maximum logit as our output measure is more aligned than using the loss for rationalizing the model's current prediction, whether it is correct or not. Contrary to our output definition, there is only one option for the inputs we calculate gradients with respect to that both faithfully represents the input tokens and allows for differentiability - the embedded representations of each token in the article currently being classified.

To calculate the gradients of the maximum logit with respect to the input embeddings for a sequence of tokens, x (sequence length x 1), we first embed each token in x to a 768-dimensional space (sequence length x 768). Our embedded sequence is then run through the fine-tuned BERT model to compute the logits, and, finally we calculate the gradients of the maximum logit with respect to the input embeddings using back-propagation. However, because our embedding space is 768-dimensional, each token from x will have an associated 768 gradients - one for each embedding dimension. Since we want a single importance score per token, we must condense these gradients into one representative value, and our two methods differ in how they compute this overall score. The *gradient norm* method uses the euclidean norm of the gradients of the max logit with respect to the input embeddings, measuring the overall magnitude of the gradients for each token. For the *embed · gradient* method, we reason that gradients should contribute more towards importance along the embedding dimensions of tokens which carry the most information. So, this method uses the dot product of our gradients with their respective embeddings. We formalize both these calculations below, where $Emb(x)$ is the embedded representation of our input sequence x, and $\frac{\partial L}{\partial Emb(x)}$ are the gradients of the max logit with respect to the input embeddings.

Gradient Norm: $$\text{Gradient norm} = \left\| \frac{\partial L}{\partial Emb(x)} \right\|$$

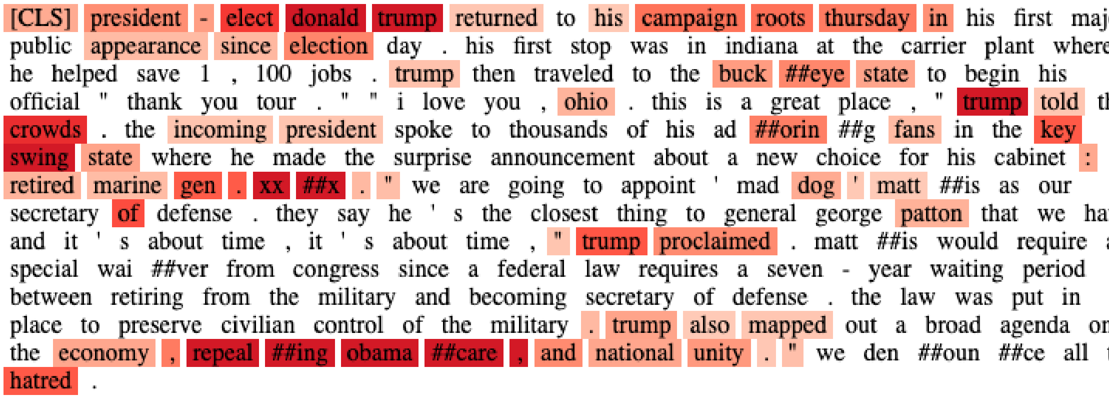
Embed · Gradient: $$\text{Embed} \ \cdot \ \text{gradient} = Emb(x) \cdot \frac{\partial L}{\partial Emb(x)}$$



```
Gradient Norm:       obama     xx        ##care    xx        explained  trump     trump
                     (0.037)   (0.037)   (0.022)   (0.018)   (0.015)    (0.012)   (0.012)
Embed * Gradient:    xx        obama     ##care    xx        explained  trump     trump
                     (0.04)    (0.039)   (0.028)   (0.02)    (0.015)    (0.014)   (0.013)
```

Figure 4-5: The top 7 most important tokens, and their importance scores, using both of our gradient-based methods. Both methods output very similar results.

In Figure 4-5, we compare the top 7 most important tokens, and their magnitudes, as computed by our two different methods. We observe that they are highly similar in the words they deem important, and only have slight differences in magnitudes. Furthermore, when visualizing the importance weighting of words within articles we also see a similarity between the two methods. Figure 4-6 shows the importance weighting for an article using the *embed · gradient* method, and the highlighted words align well with Allsides' criteria for a right-leaning article, showing that defense is important, rejecting public funding for healthcare, and discussing right-learning president-elect at the time, Donald Trump. However, even though these gradient-based methods seem like a useful strategy for extractive rationalization from a qualitative observation, further studies confirming their validity are still needed. A next step confirming these results, for example, could be training a simple Support Vector Machine classifier on top of only the words that were highlighted as important by these methods, and confirming that the classifier still operates with comparable accuracy.



[CLS] president - elect donald trump returned to his campaign roots thursday in his first major public appearance since election day . his first stop was in indiana at the carrier plant where he helped save 1 , 100 jobs . trump then traveled to the buck ##eye state to begin his official " thank you tour . " " i love you , ohio . this is a great place , " trump told the crowds . the incoming president spoke to thousands of his ad ##orin ##g fans in the key swing state where he made the surprise announcement about a new choice for his cabinet : retired marine gen . xx ##x . " we are going to appoint ' mad dog ' matt ##is as our secretary of defense . they say he ' s the closest thing to general george patton that we have and it ' s about time , it ' s about time , " trump proclaimed . matt ##is would require a special wai ##ver from congress since a federal law requires a seven - year waiting period between retiring from the military and becoming secretary of defense . the law was put in place to preserve civilian control of the military . trump also mapped out a broad agenda on the economy , repeal ##ing obama ##care , and national unity . " we den ##oun ##ce all the hatred .

Figure 4-6: Gradient-based importance visualization, using embed · gradient method, on an article classified as right-leaning (this is a truncated version of the full 512 token input fed into BERT). This method finds "donald trump" and "repealing obamacare" as particularly important phrases in classifying this article as right-learning.

4.3 Shortcomings of Extractive Rationales

Even though the extractive rationales in some of our studies were qualitatively good, and even if their validity is further verified, there are still some inherent limitations to using extractive rationales, as well as issues in our own study. While extractive rationales do increase transparency into why models make certain predictions, there is no guarantee that the information which models are using follows the same reasoning that humans use and, as a result, the extracted rationales may be incomprehensible or have different meanings than expected (e.g. exploiting grammatical tendencies, or correlations unrelated to the prediction task). In fact, we believe this is likely what happened in our own study. Though extractive rationales appeared to be reasonable explanations for our model's predictions, other experiments of ours showed that during article-level fine-tuning, BERT was not actually learning predictive language for political bias, but was somehow memorizing the media sources which published the articles.

	Macro-F1	**Accuracy**
Random Split	88.78	88.09
Media Split	32.64	36.45

Table 4.1: Random vs. Media split article-level classification. In the random split experiment, articles were randomly selected across all media sources for the train, development, and test splits. For the media split, we ensured that articles were selected from disjoint sets of media sources for the train, development, and test splits.

We realized this issue in our study when fine-tuning two different versions of our BERT model, one where the articles in the train, development, and test sets were randomly selected across all media sources (Random Split), and another where articles were selected from disjoint sets of media sources between the train, development, and test sets (Media Split). Table 4.1 shows the large gap in performance between the two models, where the model fine-tuned on the media source partition is essentially randomly guessing[2]. Through further analysis, we found that named entities, which we masked out of articles using named-entity recognition (we wanted to ensure media sources did not include their own names within articles for this very reason), were selected as highly important tokens using our methods. Out of the articles that contained some masked named-entity, we found that about 61% contained this mask within the top 10 most important tokens in their extracted rationales (15,799 out of 25,718 articles). Though we attempted to mask out the named-entities within our articles, we suspect that either there was some leakage in the named-entity recognition or the model learned to associate writing styles with specific media sources, and was relying on this to make predictions. As a result, we need to test our extractive rationale methods on other, proven methodologies in the future in order to see how they perform on a robust model.

[2] Note: Using BERT to extract features during our source-level study (Chapter 3) did not appear to have this issue, as we excluded any media sources used during fine-tuning from our classification step, and the classifiers still performed well.

Chapter 5

Conclusion

The general goal of our work is to study potential methods for preventing the spread of fake news. Specifically, we focus on building a system to predict the political bias and factuality of media, as we believe these two properties are crucially linked to the problem of fake news. Several previous works address bias and factuality prediction at smaller scopes like the claim and article-levels, but we center our work around the source-level because it is relatively understudied, useful in predicting at the smaller scopes, and can be used for detecting likely fake news in the instant it is published. Furthermore, the system we develop makes its predictions based on linguistic features, as we believe the viral and damaging power of fake news is derived from its language, and it is the most robust predictor as a result.

The methodology for the source-level political bias and factuality prediction system we build is based on work done by Baly et al. (Baly et al., 2020), and it operates on data from the media sources' websites, and optionally the media sources' Twitter and Wikipedia pages. During development of the system, we experiment with training the necessary models on different subsets of these data sources, and compare our results to a similar ablation study done by Baly et al.. Similarly to their findings, we observe that articles published by a media source produce the most important features for predicting both bias and factuality. However, our system tends to underperform the one described by Baly et al., and differs in which data channels are useful for predictions. We hypothesize that these discrepancies are a result of differences between the datasets used for training the system's models, and believe that our results could be improved by using a dataset containing more labeled media sources. Nevertheless, our system obtains adequate accuracies of 76.81% and 73.05% for the tasks of political bias and factuality prediction respectively, with news articles and Twitter data being the most predictive input combination for the bias task, and only article data being the most informative for the factuality task.

In addition to our development of a system for predicting political bias and factuality, we also study how different rationalization methods can be used to provide transparency in automated fake news detection methods. We believe it is unlikely for people to blindly trust a model's predictions on such a polarizing subject, and that presenting a model's reasoning is a critical step for these systems establishing credibility with the public. The methods we study focus on rationalizing a model's predictions on an article-level, political bias task, as politically-weighted language is well-defined for the qualitative analysis we perform. Using attention and gradient-based approaches, we highlight subsets of a model's input that are deemed important in the prediction it makes. During initial observations, the rationales we extract seem reasonably aligned with left-leaning and right-leaning political ideologies, and both the attention and gradient-based methods appear to produce simple, understandable explanations of a model's predictions. However, further analysis shows significant issues in our article-level bias prediction model which make us skeptical about believing the extracted explanations, and we believe further research is necessary, as a result, to prove the validity of the methods we use.

5.1 Future Work

In our study, we focus on using language to predict bias and factuality, but we believe that combining our methods with graph and propagation-based approaches could further augment our system. Additionally, as we have mentioned, we believe that our system can be improved by providing more labeled media sources for training its models, and, with this in mind, we design our system to be easily retrained on new data. Though there are likely a vast number of other avenues to explore for improving our system's performance, we believe that studying model explainability in the domain of fake news is the most important next step for the field.

Our rationalization study concentrates on extractive rationale methods, but a new field of research, natural language explanation, could potentially provide more understandable explanations behind a model's predictions. The field of natural language explanation aims to model human-like reasoning alongside a prediction task, and this usually takes the form of generating a textual explanation which is later used to make a prediction. With the future goal of applying natural language explanation to fake news detection, we experimented with our own approach to producing explanations on a question answering task.

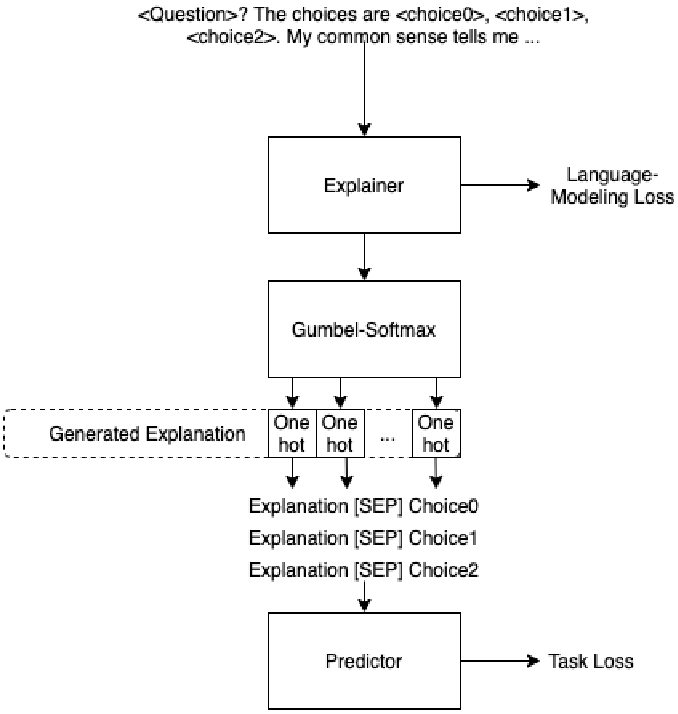

<Question>? The choices are <choice0>, <choice1>, <choice2>. My common sense tells me ...
Explainer
Language-Modeling Loss
Gumbel-Softmax
Generated Explanation
One hot
One hot
...
One hot
Explanation [SEP] Choice0
Explanation [SEP] Choice1
Explanation [SEP] Choice2
Predictor
Task Loss

Figure 5-1: Our model architecture for generating natural language explanations during a question answering task. The explainer component takes the question and answer choices as input, and generates an explanation using gumbel-softmax to sample. This explanation is then used by the predictor to select the proper answer choice to answer the question.

Figure 5-1 shows the architecture of the model we developed for generating explanations during a question answering task[1]. It consists of an explanation component which leverages a language model to generate human-like explanations, and a prediction component that uses these explanations to answer questions. We maintain differentiability through the sampling process of generating an explanation by using the gumbel-softmax (Jang et al., 2017) to allow for task-specific loss to propagate information to the explanation component. Though our model uses a supervised approach to generating its explanations, natural language generation methods are increasingly being refined (Latcinnik and Berant, 2020; Wiegreffe et al., 2020), and our hope is unsupervised methods will be available in the future and can be applied to fake news detection.

[1] We trained this model on the Common Sense Explanations dataset, which extends the Common Sense Question Answering dataset with human annotated explanations (Rajani et al., 2019).

Bibliography

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9505–9515. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf`.

L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Advances in Knowledge Discovery and Data Mining*, pages 410–421, 07 2010. ISBN 978-3-642-13671-9. doi: 10.1007/978-3-642-13672-6_40.

J. Alammar. The illustrated transformer [blog post]. https://jalammar.github.io/illustrated-transformer, 2018.

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236, 05 2017. doi: 10.1257/jep.31.2.211.

AllSides. AllSides Blind Bias Survey. https://www.allsides.com, 2020.

D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, Aug. 2010. ISSN 1532-4435.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, and P. Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium, Oct.-Nov. 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1389. URL `https://www.aclweb.org/anthology/D18-1389`.

R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-2004. URL `https://www.aclweb.org/anthology/N18-2004`.

R. Baly, G. Karadzhov, J. An, H. Kwak, Y. Dinkov, A. Ali, J. Glass, and P. Nakov. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, 2020.

A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 119–130, 05 2013. doi: 10.1145/2488388.2488400.

A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16, 2016.

K. Clark, U. Khandelwal, O. Levy, and C. Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 01 2019. doi: 10.18653/v1/W19-4828.

Dave Van Zandt. Media Bias/Fact Check. https://mediabiasfactcheck.com, 2015.

C. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274, 04 2016. doi: 10.1145/2872518.2889302.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

S. Dungs, A. Aker, N. Fuhr, and K. Bontcheva. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1284`.

E. Elejalde, L. Ferres, and E. Herder. On the nature of real and perceived bias in the main-stream media. *PloS one*, 13(3):e0193765, 03 2018.

K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *EMNLP*, 2019.

J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96 5:1029–46, 2009.

K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

B. Horne and S. Adali. 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 *Proceedings of the International AAAI Conference on Web and Social Media*, volume abs/1703.09398, 2017.

B. D. Horne, W. Dron, S. Khedr, and S. Adali. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 235–238, Republic and Canton of Geneva, CHE, 2018a. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3186987. URL `https://doi.org/10.1145/3184558.3186987`.

B. D. Horne, W. Dron, S. Khedr, and S. Adali. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *ICWSM*, 2018b.

C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 01 2015.

S. Iyengar and K. S. Hahn. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39, 2009. doi: https://doi.org/10.1111/j.1460-2466.2008.01402.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2008.01402.x`.

S. Jain and B. C. Wallace. Attention is not explanation. In *NAACL-HLT*, 2019.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

G. Karadzhov, P. Atanasova, P. Nakov, and I. Koychev. We built a fake news & click-bait filter: What happened next will blow your mind! In *Proceedings of the International Conference on Recet Advances in Natural Language Processing*, pages 334–343. RANLP, 11 2017. doi: 10.26615/978-954-452-049-6_045.

J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2145. URL `https://www.aclweb.org/anthology/S19-2145`.

E. Kochkina, M. Liakata, and A. Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1288`.

V. Kulkarni, J. Ye, S. Skiena, and W. Y. Wang. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1388. URL `https://www.aclweb.org/anthology/D18-1388`.

S. Kumar and N. Shah. False information on web and social media: A survey. *Social Media Analytics: Advances and Applications*, 04 2018.

S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, 04 2016. ISBN 9781450341431. doi: 10.1145/2872427.2883085.

V. Latcinnik and J. Berant. Explaining question answering models through text generation, 2020.

Y. Lin, J. Hoover, G. Portillo-Wightman, C. Park, M. Dehghani, and H. Ji. Acquiring background knowledge to improve moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559, 2018. doi: 10.1109/ASONAM.2018.8508244.

A. Margeloiu, N. Simidjievski, M. Jamnik, and A. Weller. Improving interpretability in medical imaging diagnosis using adversarial training, 2020.

L. Mitchell, M. Frank, K. Harris, P. Dodds, and C. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8:e64417, 05 2013. doi: 10.1371/journal.pone.0064417.

S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, page 353–362, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337946. doi: 10.1145/2806416.2806537. URL `https://doi.org/10.1145/2806416.2806537`.

J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL `https://www.aclweb.org/anthology/N18-1100`.

N. Nguyen, G. Yan, M. Thai, and S. Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 213–222, 06 2012. doi: 10.1145/2380718.2380746.

R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175. URL `https://doi.org/10.1037/1089-2680.2.2.175`.

A. C. Nied, L. Stewart, E. Spiro, and K. Starbird. Alternative narratives of crisis events: Communities and social botnets engaged on social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, page 263–266, New York, NY, USA, 2017. ACM. ISBN 9781450346887. doi: 10.1145/3022198.3026307. URL `https://doi.org/10.1145/3022198.3026307`.

J. Nørregaard, B. Horne, and S. Adali. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *ICWSM*, 2019.

B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32:303–330, 06 2010. doi: 10.1007/s11109-010-9112-2.

B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 271–es, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1218955.1218990. URL `https://doi.org/10.3115/1218955.1218990`.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://www.aclweb.org/anthology/D14-1162`.

PolitiFact. Politifact. https://www.politifact.com/, 2007.

K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3186967. URL `https://doi.org/10.1145/3184558.3186967`.

M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1022. URL `https://www.aclweb.org/anthology/P18-1022`.

N. Rajani, B. McCann, C. Xiong, and R. Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, 01 2019. doi: 10.18653/v1/P19-1487.

H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 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*, pages 2931–2937, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1317. URL `https://www.aclweb.org/anthology/D17-1317`.

M. Recasens and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, volume 1, 01 2013.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference in Empirical Methods in Natural Language Processing*, pages 3973–3983, 01 2019. doi: 10.18653/v1/D19-1410.

A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1044. URL `https://www.aclweb.org/anthology/D15-1044`.

A. Saleh, R. Baly, A. Barrón-Cedeño, G. Da San Martino, M. Mohtarami, P. Nakov, and J. Glass. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1041–1046, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2182. URL `https://www.aclweb.org/anthology/S19-2182`.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Expor. Newsl.*, 19(1): 22–36, Sept. 2017. ISSN 1931-0145. doi: 10.1145/3137597.3137600. URL `https://doi.org/10.1145/3137597.3137600`.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, and A. Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models, 2020. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.15`.

J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL `https://www.aclweb.org/anthology/N18-1074`.

R. M. Tripathy, A. Bagchi, and S. Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 2010.

J. Uscinski, C. Klofstad, and M. Atkinson. What drives conspiratorial beliefs? the role of informational cues and predispositions. *Political Research Quarterly*, 69, 01 2016. doi: 10.1177/1065912915621621.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

J. Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL `https://www.aclweb.org/anthology/P19-3007`.

S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359:1146–1151, 03 2018. doi: 10.1126/science.aap9559.

S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, 01 2019. doi: 10.18653/v1/D19-1002.

S. Wiegreffe, A. Marasovic, and N. A. Smith. Measuring association between labels and free-text rationales, 2020.