A STUDY OF MEDIA POLARIZATION WITH AUTHORSHIP ATTRIBUTION METHODS

by

Yifei Hu

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Julia Rayz, Chair

Department of Computer and Information Technology

Dr. Baijian Yang,

Department of Computer and Information Technology

Dr. Jin Kocsis,

Department of Computer and Information Technology

Approved by:

Dr. John A. Springer

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Media polarization is a serious issue that can affect someone's views, ranging from a scientific fact to the perceived results of a presidential election. The media outlets in the United States are aligned along political spectrum representing different stances on various issues. Without providing any false information (but usually by omitting some facts), media outlets can report events by deliberately using the words and styles that favor particular political positions.

This research investigated the U.S. media polarization with authorship attribution approaches, analyzing stylistic differences between the left-leaning and right-leaning media and discovering specific linguistic patterns that made the news articles display biased political attitudes. Several models of authorship attribution were tested while controlling for topic, stance, and style, and were applied to media companies and their identity within a political spectrum. Style features that were compared included semantic and/or sentiment-related information, such as stance taking, with features that seemingly do not capture it, such as part of speech tags. The results demonstrate that a successful classification of articles as left-leaning or right-learning is possible regardless of their stance. Finally, we provide an analysis of the patterns that we found.

CHAPTER 1. INTRODUCTION

1.1 Motivation

The motivation for this thesis came from a description of a recent political event. In June 20, 2020, President Trump held his campaign rally at Tulsa. A few days later, several staff members who joined the rally tested positive for Coronavirus (COVID-19). CNN and Fox News reported this event on the same day with totally different attitudes that served for different political inclinations. CNN reported that “Attendees at Trump‘s rally were not required to wear a mask or practice social distancing… administration officials at the rally did not wear masks, though campaign manager Brad Parscale was seen in one.” Fox News reported that “All rally attendees were given temperature checks before heading through security and offered face masks and hand sanitizer, though the campaign emphasized that wearing a face covering was optional.”

The two media companies were describing the same event but there was very little overlap in their description. CNN emphasized that the campaign team did seemingly nothing to protect the attendees and only the campaign manager wore a mask. Fox News, however, described the actions the campaign team took to seemingly protect the attendees. The actions included temperature check, face masks offered by the campaign team, and hand sanitizer available. When reading such opposite descriptions, one may question whether CNN or Fox News reported any of the details with distortion. The answer is negative: both media companies were telling the truth; nevertheless, their perception of truth was described to emphasize different salient points, for different political inclinations, and in different writing styles.

The media polarization issue has been studied for decades. Fox News is known to be a so-called right-wing media, while CNN represents the opinion from the left-wing of the American Politics. This means that usually Fox News highlights events from the point of view of the American Republican Party, while CNN sides with the opinions expressed by the Democratic Party. In the media and journalism fields of study, researchers have been observing and reporting the media bias with case studies and analysis. However, in the natural language processing, there is a very limited number of research publications that systematically measured the difference in term of “style” between different media companies or different political side.

This research explores the stylistic difference between different groups with authorship attribution approaches. The question this research is interested in answering is: using computational authorship attribution approaches, can computers learn to differentiate the stylistic characteristics of news articles, relative to their political spectrum?

1.2 Significance

Stylometry is usually thought of as ”the science of inferring characteristics of the author from the characteristics of documents written by that author” (Joula, 2008) Yet, many authors that may share similar worldview or values may have similar style or stance. While there is an abundant and comprehensive research on stance recognition (Hasan et al., 2013 ,& Boltužić et al., 2014), among all the research in the authorship attribution and stylometry area, none have investigated the stylometry features for different organizational groups, in particular, for media companies. Most quantitative research in the media and journalism area measured the media bias from the perspective of coverage, truthfulness, or the correlation between each piece of news information, while the works in authorship attribution area primarily focused on the stylometric features of individuals or authors with similar physical characteristics (gender, age, etc.). This research is filling this gap by studying the media polarization with computational authorship attribution approaches and linguistic knowledge.

1.3 Research Question

The question central to this research is:

- What are the shared unique stylometry features within each U.S. media group or political side that can be identified by computational authorship attribution approaches?

1.4 Assumptions

The assumption for this study is:

- The 3rd-party research organizations provide mostly accurate conclusions about the political leaning of media outlet.

1.5 Limitations

The limitations for this study include:

- This research only focused on the mainstream English-language media companies in the United States.
- This research used stylistic features proposed in the literature to study the stylometry of each individual author. Thus the classification models should ideally classify the texts mainly based on the stylistic features of the authors.

1.6 Definitions

In the broader context of thesis writing, we define the following terms:

Media Group: The media companies like CNN, Fox News, ABC, NBC, etc.

WSJ: The Wall Street Journal, a media company.

NYT: The New York Times, a media company.

Breitbart: The Breitbart News, a far-right media company.

Vox: The Vox Media, an American digital media company.

CNN: Cable News Network is a news company.

CNN Model: Convolutional neural network model, a class of deep learning models.

Token: In the area of natural language processing for English, a token usually means a word.

1.7 Summary

Media polarization is an emerging phenomenon which has a profound influence on how audiences perceive the news events. By analyzing the patterns of writing styles from different media outlets, it may be possible to understand the general patterns of the biased news information. Authorship attribution is an area studying the style of text. This work applies authorship attribution approaches to study the media polarization in the U.S. media outlets.

CHAPTER 2. LITERATURE REVIEW

This chapter provides a review of the literature relevant to stylometry, authorship attribution, and media polarization.

2.1 Authorship Attribution and Stylometry

We start with literature in authorship attribution and stylometry domains.

2.1.1 Authorship attribution

According to Joula (2008), authorship attribution is the science of inferring characteristics of the author from the characteristics of documents written by that author. Joula stated that the history of authorship attribution data can be traced back to the Old Testament if we broadly define the question "what can I tell about a person from the language he uses" as the earliest authorship attribution application. Joula defined 3 main problems in authorship attribution:

1. The closed class: given a particular sample of text known to be one of a set of authors, determine which one.

2. The open class: given a particular sample of text believed to be by one of a set of authors, determine which one.

3. Stylometry (or profiling): determine the properties of the authors of a sample of text(s).

Stamatatos (2009) defined other tasks of authorship attribution:

1. Author verification: decide whether a given text was written by a certain author.

2. Plagiarism detection: find similarity between two texts.

3. Author profiling or characterization: extract information about the age, education, sex, etc., of the author of a given text.

4. Detection of stylistic inconsistencies as may happen in collaborative writing.

Looking at these tasks, the applications of authorship attribution seem broad. An ideal authorship attribution scenario should be a small closed set of candidate authors with abundant training texts for each of the authors. The training texts should contained enough information for the authorship attribution algorithms to learn the characteristics of the individual author. However, the real-life authorship attribution tasks could encounter many scenarios that were less ideal. Koppel (2009) described 3 main common hinders:

1. The profiling problem: it's difficult to find enough demographic or psychological information about the author.

2. The needle-in-a-haystack problem: too many authors with too few writing samples.

3. The verification problem: how to determine a suspect is or is not the author.

Due to the difficulties encountered in real-world authorship attribution applications, researchers explored more than 1,000 different linguistic features (Abbasi et al., 2008) to adapt the dynamic scenarios. Each of the linguistic features contained certain level of predictive information about the authorship, but sometimes the features weren't stable or robust to provide consistent performance on some tasks. Some linguistic features (like vocabulary richness and average sentence length) were related to certain characteristics (age, gender, education level, native language, etc.) of the authors. Applying computational approaches to these features allowed to make inference about the authorship based on given text data. More details of the features and computational authorship attribution approaches is discussed in Section 2.2.

2.1.2 Stylometry

Within authorship attribution papers the term "stylometry" is mentioned quite frequently. Stylometry is defined as "the use of statistical methods in the analysis of literary style" (Holmes, 1998). Stylometry is closely related to authorship attribution if one uses Joula (2008) description of the 3rd problem in authorship attribution, which stated that "authorship Attribution is a near-synonymous term of 'stylometry'".

In some papers, the term “stylometry” is indeed a near-synonym to “authorship attribution”. For instance, Abbasi et al. (2008) described one of the stylometric analysis tasks as “compare anonymous texts against those belonging to identified entities, where each anonymous text is known to be written by one of those entities”. This matched the description of the closed class task defined by Joula (2008).

From stylometric analysis, many attributes of the authors can be recovered from the texts. These attributes include age, gender (Goswami et al., 2009), native languages (Koppel et al., 2005), whether the author has dementia (Le et al., 2011), whether someone is writing deceptive online reviews (Ott et al., 2011), and whether a paper is written in the style of a conference or a workshop (Bergsma et al. 2012). As people are getting more aware of the importance of privacy, there are also applications trying to hide the authors’ attributes by obfuscating the writing style (Emmery et al., 2018).

At first, the study of style initially focused on natural language. Later, stylometry also expanded to programming languages (computer codes). Caliskan-Islam et al. (2015) used stylometric analysis to deanonymize programmers via the source code and proved that stylometry can be used on structured languages. If we extend the definition of stylometry to a broader level, then music (Westcott, 2006) and paintings (Qi et al., 2011) can also be analyzed with similar approaches.

2.1.3 Topic, genre and style

Computational approaches of authorship attribution are text classification methods in nature. When applying the approaches, it is possible that the algorithms classify the texts based on not only the style but also the topic, genre, or other elements of the texts. In other words, what needs to be analyzed is whether there are any topic-free stylometric features containing predictive information about the authorship.

Kešelj et al. (2003) performed an authorship attribution experiment on seven different authors: Emily Bronte, Edgar Rice Burroughs, Lewis Carroll, John Cleland, Charles Dickens, H. Ryder Haggard, Washington Irving, and Williams Shakespeare. The result was nearly perfect with some parameter tuning. It should be noted, these 7 authors lived in different eras and countries, and their books had very different plots. However, Kešelj et al. reported on the text classification algorithm without controlling for the topic or genre, which meant that the authors could be potentially identified not by style alone. In this thesis, we have to question confounding factors that lead to high accuracy of stylometric classification.

The discussion about topic, genre and style has been addressed by other papers. Mikros et al. (2007) reported that many stylometric variables were discriminating topics rather than authors. The conclusion suggested that when performing authorship attribution experiments on multitopic corpora, the researchers should be extremely careful about the impact of the topic.

Moreover, Sarawgi et al. (2011) performed several authorship attribution experiments on various genres and topics and found statistical evidence of gender-specific language styles beyond topics and genres, although the token-level language models had the tendency to learn topic words on top of the basic stylometric cues.

Shrestha et al. (2017) suggested that character-level features like character-n-grams learned the style of an author. Applying convolutional neural network and character-n-grams can identify verbal fillers and structural features of the given authors.

2.1.4 Stance vs. style

Besides topic and genre, another variable in texts is the stance. According to the Merriam-Webster dictionary, stance is defined as an intellectual or emotional attitude. When people choose to take a stance, the action is called "stancetaking" which is viewed as a social action that shares the speaker's view of an object with their audience, sometimes inviting listeners to take their own stance as well (Englebretson, 2017). In the world of journalism, two news articles can have the same topic and genre but opposite stances.

There are very few papers exploring the boundary between topic, stance, and style. Most authorship attribution or stylometry papers simply ignored the existence of stance. Among the papers exploring both stance and style, many believed that there was no explicit boundary between the two elements. Kiesling (2009) stated that "repeatable linguistic styles emerge out of stancetaking strategies that prove repeatedly relevant and useful for particular speakers in particular kinds of interactions". Jaffe (2009) also suggested that "a personal style is created through habitual stancetaking". Both papers indicate that stance and style are closely related. More importantly, style is the concrete and embodied expression of stance. The analysis of stylometric features also learned the stance difference consciously or unconsciously.

2.2 Approaches for authorship attribution

There are more than 1,000 linguistic features and various computational approaches in the authorship attribution area. Based on the type of task, the type of language, the genre of the texts, the length of the texts, the number of authors, and the size of the data set, different features and approaches may produce the best results.

In many research papers, the term "authorship attribution" was used with varied indications. In some papers, the authorship attribution experiments were strictly referring to the stylometric analysis, which studied one or multiple linguistic features in the texts and tried to correlate specific characteristics (age, gender, etc.) of the authors with the linguistic features. However, other papers called the general text classification experiments as "authorship attribution" where the candidate authors were the labels and the main goal was to find the author for a given piece of text.

2.2.1 Linguistic features

Joula (2008), Koppel et al. (2009), and Stamatatos (2009) introduced different types of linguistic features that can be used in authorship attribution experiments. Both statistical approaches and machine learning based approaches are built based on the fundamental linguistic features. Below are some linguistic features mentioned in multiple authorship attribution papers:

- Vocabulary
  - Word length
  - Sentence length
  - Vocabulary richness
  - Word frequency

- Syntactic
  - Part-of-speech (POS)
  - Function words
  - Punctuation
- N-gram
  - Word N-gram
  - Character N-gram
- Structure
  - Layout
  - Font
  - Figure placement
  - Indentation
- Anomaly
  - Spelling mistakes
  - Cultural differences ("trunk" and "boot", "color" and "colour")

Among all the linguistic features above, some are worth noting. Function words are defined as "a set of words that do not carry any semantic information" (Stamatatos, 2009). The importance of function words is related to the discussion between topic and style. For authorship attribution tasks, the words without semantic information can reflect the author's stylometric features. In style-based classifications, function words often carry predictive information. In topic-based classifications, function words are often excluded from the features since these shallow information are not related to the topics.

The part of speech (POS) feature is usually annotated automatically by part of speech taggers. Søgaard (2010) proposed an approach for Part of speech tagging that achieved about 97% accuracy. The high tagging accuracy allowed the POS to be a feature in most authorship attribution experiments.

The vocabulary richness as a feature has been mentioned in multiple review papers. Wimmer and Altmann (1999) proposed that vocabulary richness was related to factors like theme and style of the text, intelligence level, age, and other characteristics of the author. However, Hoover (2003) suggested that vocabulary richness was an unreliable feature. “Different texts and even different sections of a single text by one author are almost as different in vocabulary content as are texts by different authors.” Even for the same author, the vocabulary richness can be inconsistent over time. The best situation to utilize vocabulary richness as a feature is when there is a small and extremely various group of texts. However, in such a limited use case, the necessity of doing authorship attribution is highly questionable.

Average sentence length is another common linguistic feature in authorship attribution. It is easy to be calculated either manually or by computer algorithms. Kruh (1982) used the anomalous average sentence length of the Beale manuscripts as an evidence to argue they were forgeries.

One last feature worth mentioning is the n-gram. An n-gram is a sequence of N words or characters. For example, “for example” is a bi-gram of words and “fo” is a bi-gram of characters. N-gram is a powerful feature in many authorship attribution tasks. Word-based n-grams can capture the most frequently-used word combinations of an author, which is a supplementary feature for word frequency. The character n-gram is even more popular in authorship attribution research since it can capture errors (for instance, both “goverment” and “enviroment” are missing the “nme” tri-gram) and words with similar roots (for instance, “author”, “authors”, and “authorship” shared some character N-grams). For some Asian languages (like Chinese), the distinction between word and character is vague, which makes the process of word tokenization less easy and accurate. For Chinese and Japanese, there is an extra step called word segmentation which needs to be applied before the tokenization step (Xue, 2003 & Mochihashi et al., 2009). In such cases, character n-gram suits the situation much better. (Statmatatos, 2009)

Researchers have already identified more than 1,000 different features (Abbasi et al., 2008). With all those linguistic features, selecting and combining the features to achieve the best performance have become an independent research question. Stuart et al. (2013) compared multiple linguistic feature combinations and discovered that the features cannot be infinitely accumulated. When combining more than 5 features, the accuracy started to decrease.

In the next few subsections, we discuss computational approaches for authorship attribution.

2.2.2 Statistical Approach

The earliest and most fundamental authorship attribution approach is to use statistical inferences. In the previous subsection, we have discussed various linguistic features often used in authorship attribution experiments. The statistical approaches use one or multiple linguistic features and apply statistical tests to quantify the differences of the selected features in different documents. Research suggests that no single feature is robust enough to produce a good result (Joula, 2008). However, combining multiple features can make a significant difference (Abbasi et al., 2005).

Burrows (2002) analyzed the frequency of the 150 most frequent words in the corpus and calculated a z-score (the distance from the mean) for each word. Based on the z-score, Burrows defined a measure called “Delta” - “the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text”. The value of Delta represents the similarity between two text groups.

Peng et al. (2004) proposed an approach that combined the statistical n-gram models and the Naïve Bayes algorithm. This approach was tested on several authorship attribution tasks and multiple languages. The result suggested that “basic statistical language modeling ideas might be more relevant to other areas of natural language processing than commonly perceived”

<u>2.2.2.1 Support Vector Machine</u> Support Vector Machine (SVM) (Cortes et al., 1995) is a machine learning algorithm invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. It was later refined and published by Cortes et al. in 1995. The SVM algorithm can be used for both classification and regression problems.

The main motivation behind SVM is that if the data points cannot be separated in a low dimension, they are separable in a higher dimensional. There are multiple higher-dimensional hyperplanes that may separate the data points, therefore the goal of the SVM algorithm is to find the hyperplane with the largest margin (the distance from the hyperplane to the nearest data point on each category).

SVM is known to have the scalability problem, meaning it does not work well with large data size. However, in most authorship attribution tasks, the amount of training data is very limited, thus SVM works fine in such situations.

In the research conducted by Diederich et al. (2003), SVM performed well on various tests. SVM does not require the process of feature selection and can handle relatively "large" sizes of input data by authorship attribution standards. For example, the dataset used by Diederich et al. consisted of 2652 documents which was larger than many other authorship attribution datasets (compared to the dataset used by Kešelj et al. (2002) which only contained seven books) but still relatively small compared with the datasets used in deep learning experiments. In an authorship attribution experiment on German newspapers, SVM achieved close to 100% precision for the majority of the authors. SVM also showed great robustness even when the topics were not strictly controlled.

Ouamour et al. (2012) used Sequential Minimal Optimization based Support Vector Machine (SMO-SVM) to perform authorship attribution on ancient texts written by 10 Arabic travelers and achieved 80% accuracy. In their experiment, the SVM classification was performed with different n-gram features. The experiment results showed that character n-gram performed much better than word n-gram.

2.2.3 Neural Networks

The nature of authorship attribution task is very similar to text classification. In authorship attribution tasks, we classify the authors based on their texts. Neural Networks are known to achieve high accuracy in text classification tasks.

Zhang et al. (2015) published a paper on general text classification using character-level convolutional neural networks. The input text is encoded into a one dimensional vector which is then passed to a convolutional neural network (CNN). By comparing the character-level CNN with word-level CNN, and Long Short-Term Memory (LSTM) models, Zhang et al. suggested that "the character-level CNN is an effective method".

Shrestha et al. (2017) adopted a similar idea and used character level n-gram with convolutional neural networks for authorship attribution on short texts. Shrestha et al. tested the approach with different n-gram size, different numbers of authors, and different number of short texts written by each author. The result indicated that the n-gram CNN approach can work on various situations. The accuracy for 50 authors with 1000 tweets each was 76.1%. Shrestha et al. also analyzed what types of character n-gram were learned by the CNN model. By comparing the n-grams of a bot-like author and a human author, the differences in style were identified. The bot-like author tended to follow a "Title: URL" (": [U]") pattern while the human author favored some special verbal fillers like "uhm" or "um".

Bagnall (2016) used multi-head recurrent neural networks to cluster similar authors. Alsulami et al. (2017) used LSTM for source code authorship attribution on 10, 25, and 70 different authors. The results showed that the LSTM models outperformed the SVM and Random Forest models. Due to the choice of dataset and the evaluation metrics, it is difficult to directly compare the recurrent neural network (RNN) models with other CNN models.

2.3 Media Polarization

This section reviews the literature in the media and journalism area.

2.3.1 U.S. media polarization

Jurkowitz et al. (2020) quantitatively analyzed the U.S. media polarization in 2020. They investigated 30 major news sources with a range of different measures. According to the findings, of the 30 news sources, Democrats trust 22 of them, while Republicans only trust 7 of them. Most Republicans place their trust solely on Fox News, while Democrats trust multiple sources including CNN, NBC News, ABC News, CBS News and PBS.

According to Figure 2.1, some of the neutral news sources are New York Post, CBS News, ABC News, NBC News, USA Today, and WSJ. Fox News and 5 other news sources are right-leaning, while the rest of the news sources are left-inclined.

Allsides Media (2020) also classified the media based on their political lean. Comparing Figures 2.1 and 2.2, one can notice many agreements between the two ratings. For instance, USA Today and WSJ are neutral in both ratings. CNN and Vox are left. Fox News and Breitbart are right.

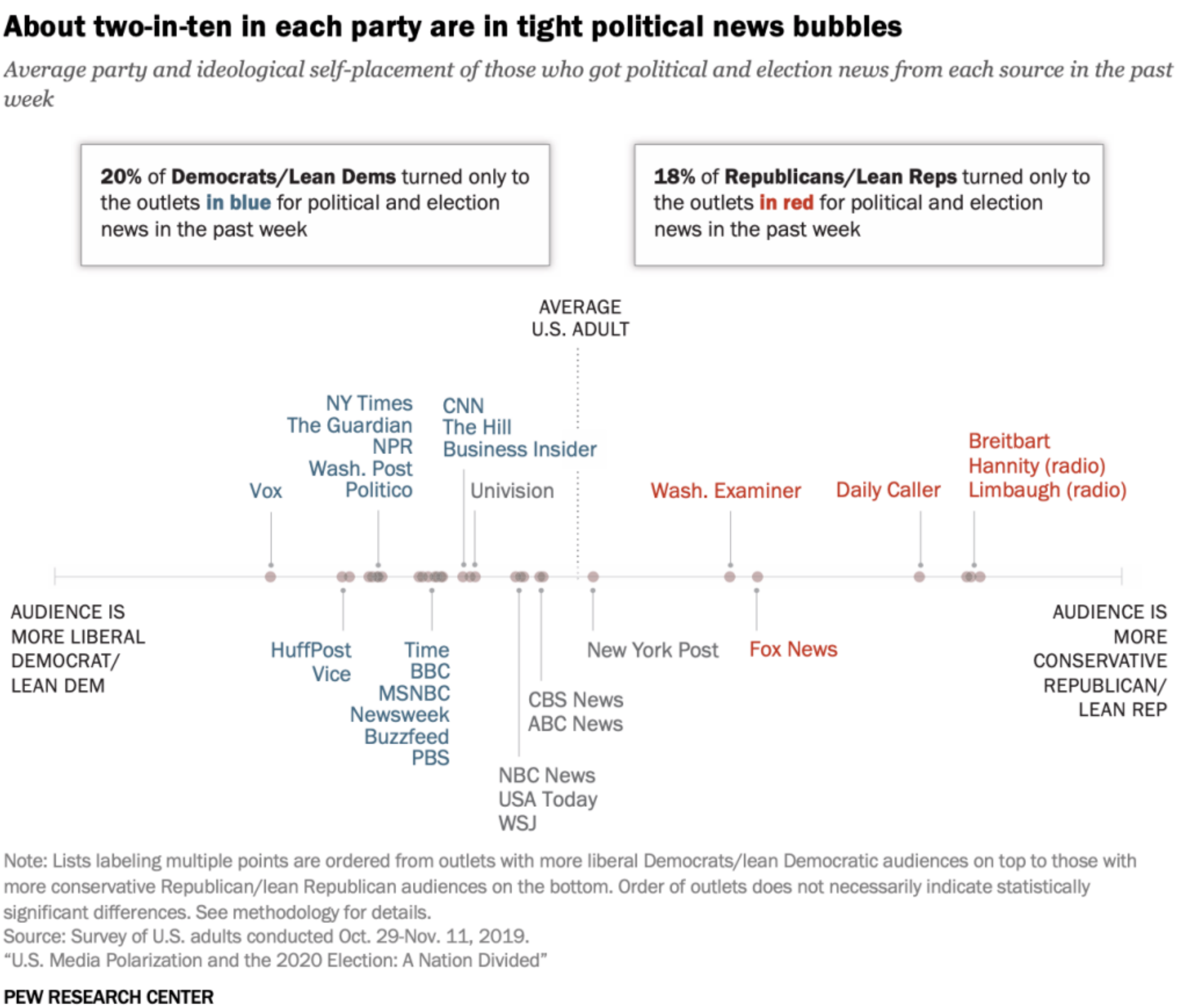

Figure 2.1. Average audience placement of each news outlet based on party and ideology. (Jurkowitz et al., 2020)

Figure 2.2. Allsides Top Online News Media Bias Ratings (Allsides, 2020)

2.3.2 Reasons for bias in the media

Posner (2015) pointed out that competition made the polarization issue even more obvious. The great example is the competition between CNN and Fox News. Fox expanded dramatically from 1996 to 2000 (DellaVigna et al., 2007), which caused CNN to shift to the left. Fox gained right-inclined viewers while CNN had to lose the same viewers in the competition with Fox News. In this situation, CNN decided to make its content more appealing to the remaining viewers, the left-leaning viewers, by embracing their political preferences. As more news sources joined the competition, the polarization issue is also becoming more inevitable.

As Posner explained, the liberals wanted to read liberal newspapers and the conservatives conservative ones, since people didn't want their beliefs to be challenged or being in a state of doubt. In fact, Annenberg Public Policy Center's survey showed that 43% of the respondents preferred the news organizations to have a decidedly political point of view (Overholser, 2016).

Xiang et al. (2007) viewed the media bias as the consequence of political advertising and public relations activities. According to Xiang et al., there might be a market for extreme media positions. The extreme stances were able to mitigate the inconsistency between media positioning and the candidate's preferred position.

Fox News is not the only media organization that intensifies the competition and caused more polarization. Hopkins et al. (2013) also mentioned MSNBC, which was established under a strategic partnership between NBC and Microsoft, had the similar impact on the left-inclined audience, making them become even more left-leaning.

2.3.3 The consequence of media polarization

The bias in each media group can impact the audience's political lean and further impact the voting statistics. DellaVigna et al. (2007) studied the expansion of Fox News and its relation to the voting data, claiming that there was a significant effect of exposure to Fox News on voting. "Towns with Fox News have a 0.4 to 0.7 percentage point higher Republican vote share in the 2000 Presidential elections, compared to the 1996 elections."

Groeling et al. (2008) investigated the partisan bias on ABC, CBS, NBC and Fox News. The result showed that "ABC, CBS, and NBC all appeared to favor good news for Clinton and bad news for Bush, while Fox appeared to favor the reverse". For CBS and NBC, they simply chose not to cover the bad news about Clinton.

On topics like climate change, the viewers of different news sources also had different beliefs. Feldman et al. (2011) suggested that the viewers of CNN/MSNBC emerged with different beliefs about climate change than viewers of FOX News. "the opportunity for consensus-building and cooperation on global warming—as well as on other critical issues of the day—diminishes".

Krosnick et al. (2010) even stated that the frequent viewers of Fox News were less likely to accept scientists' views of global warming due to the skeptical messages spread by Fox News possibly.

The bias between the left-wing and right-wing media can also be reflected from some statistical measurements. Harmon et al. (2009) studied the word frequency of FOX, CNN, and other new groups' news coverage on the Iraq War. The statistics showed that FOX, compared to CNN, was more likely to use the pro-war terms.

From 1987 to 2003, the diversity index of media's attention to different interest groups grew from 22.0 to 46.0 (Binderkrantz, 2011). The interests groups with clear political positions easily gained more media attention. The media polarization divided viewers' perception in political, economical, and scientific areas by spreading extreme point of view against the competitors with different political inclinations.

2.3.4 Media Polarization and the 2016 election

Donald Trump had a surprising win in the 2016 election while the majority of the news outlets predicted the opposite results. The left-learning and right-leaning media outlets behaved dramatically differently on the news coverage about the candidates. According to Faris et al. (2017), the right-learning media concentrated on pro-Trump content significantly before the election, while the left-leaning media maintained the objectivity of journalism and covered broader topics. Francia (2018) also suggested that Trump gained advantage from the media by dominating the discussions on social media platforms. Each of Donald Trump's tweets could become a news topic of the day.

In the battlefield of media, Donald Trump benefited the most from the "Fake News" (cases of deliberate presentation of false or misleading claims as news (Gelfert., 2018)) which spread false information that was mostly pro-Trump or anti-Clinton. Allcott & Gentzkow (2017) even pointed out that "Donald Trump would not have been elected president were it not for the influence of fake news". Based on the Facebook data used in Allcott & Gentzkow's paper, pro-Trump fake news had spread 3 times more than the pro-Clinton fake news.

The motivation of fake news was primarily attacking the opposing party (Vargo et al., 2018). The quick spread of fake news was also boosted by some traditional cable news outlets.

For instance, the fake reporting of Hillary Clinton's leaked speech to Wall Street banks was first generated by "RealTrueNews", a website producing fake news. This fake news was soon picked up by Fox News which is a traditional cable news outlet. Later, the same fake news was reported by multiple right-leaning websites and caused great public exposure.

As we discussed in the previous section, the bias in media can impact the political lean of the public and the 2016 election is one of the examples. To eliminate or mitigate the media bias, we need to come up with metrics that can measure the bias. The next section is going to discuss how to measure the bias in media.

2.3.5 Measuring the bias in media

There are some existing methods to measure the media bias. D'Alessio et al. (2006) introduced the meta-analysis with d'. d' measures the difference in coverage. If a newspaper gives the Democratic candidate 20% more coverage, d' would be 0.20. The statement bias can also be measured with the similar algorithm.

D'Alessio et al. also introduced p which was defined as the Gatekeeping Bias. The Gatekeeping Bias (p) can be calculated by the following formula 2.1:

$$p = (stories/n)/events \tag{2.1}$$

In the formula, stories is the total number of stories generated, events is the number of campaign events isolated, and n is the number of media outlets (in each of these cases, newspapers) in the sample.

Guruji (2018) used natural language processing techniques to measure the polarization in media. The news articles were clustered with K-means Clustering method, which was an unsupervised learning technique, on top of some document representation approaches. The result showed that news articles written by the same author(s) shared more common features. However, the result did not capture the polarization between the major U.S. media outlets.

An et al. (2012) proposed a novel approach to measure and visualize the media bias by analyzing the follow links on Twitter. Figure 2.3 showed the result of the experiment. The most of the left-wing media outlets were clustered next to each other. However, Fox News was located between The Washington Times and The U.S. News, which did not match the commonly acknowledged political position of Fox News.

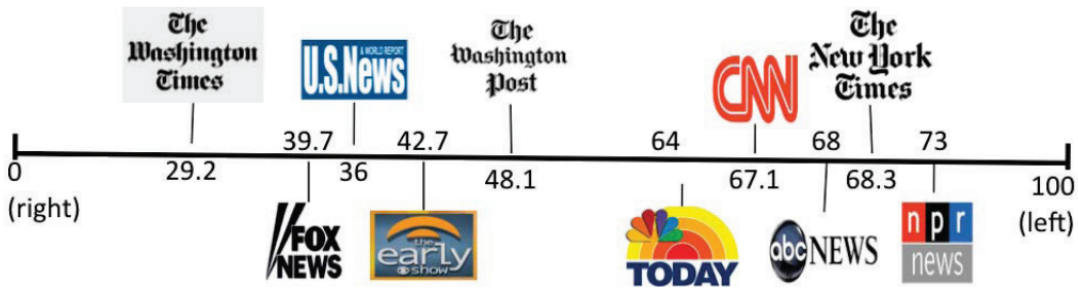

Figure 2.3. Media outlets clustering results (An et al., 2012)

2.4 Summary

This chapter provided a review of the literature relevant to authorship attribution, stylometry and media polarization. The field of authorship attribution and stylometry provides quantitative approaches to analyze texts data with respectful precision and flexibility. Both predictive and descriptive information can be extracted from large amount of texts with various types of algorithm. Media polarization is an area where more and more research projects have started to rely on quantitative computational approaches to find patterns and make inferences. This study conducted a series of experiments on media polarization with authorship attribution approaches.

CHAPTER 3. DATASET AND METHODOLOGY

This chapter provides a description of the dataset and methodology that were used in the research.

3.1 Dataset

The following sections will discuss the detailed population and sample information about the dataset used for this experiment.

3.1.1 Population

The dataset we used was retrieved from Kaggle.com, which contains more than 140,000 news articles from 15 major news sources. The number of articles of published by each news source and the political lean of each news source (Jurkowitz et al., 2020) are shown in TABLE 3.1.

Out of 143,000 articles, 104,001 of them were published between June 2016 to June 2017. More information pertaining to distribution of publication date is shown in Figure 3.1. There were articles published before June 2016 or after June 2017. However, the distribution was very sparse.

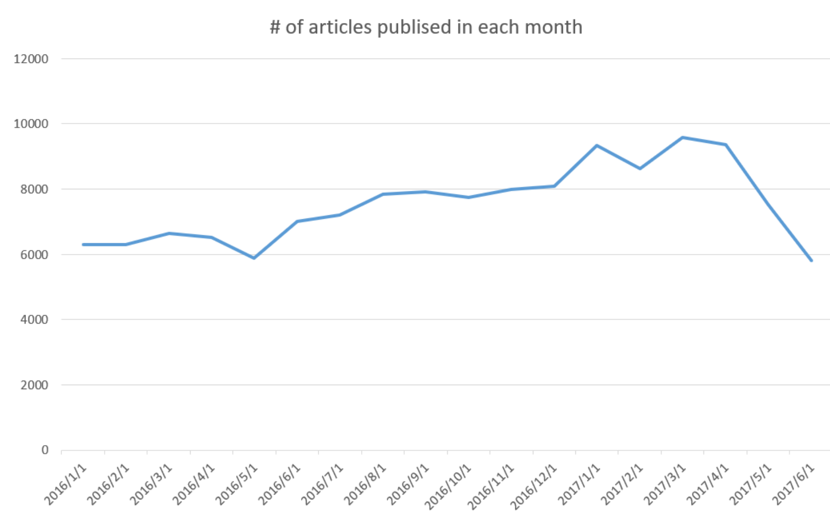

Figure 3.1. # of articles published in each month

Table 3.1. *Dataset Statistics - Overview*

media group	# of article	political leaning	selected
Breitbart	23781	R	7076
New York Post	17493	R	N/A
NPR	11992	L	N/A
CNN	11488	L	1306
Washington Post	11114	L	2963
Reuters	10710	C	N/A
Guardian	8681	L	N/A
New York Times	7803	L	813
Atlantic	7179	L	1579
Business Insider	6757	N/A	N/A
National Review	6203	R	N/A
Talking Points Memo	5214	N/A	N/A
Vox	4947	L	1236
Buzzfeed News	4854	L	586
Fox News	4354	R	1567

Table 3.2. *Dataset Statistics - Article Length*

Total # of Articles	142570
Average length	779
Standard deviation	798
Min length	1
25%	360
50%	621
75%	965
Max length	51861

Table 3.3. *Dataset Statistics - Names Mentioned from January to November 2016*

name	**# of occurrence**	**# of article**	**Mentioned by the Left**	**Mentioned by the Right**
Hillary	41492	16727	N/A	N/A
Clinton	117700	18106	N/A	N/A
Hillary Clinton	29770	15735	3493	5201
Donald	51449	23638	N/A	N/A
Trump	241018	25211	N/A	N/A
Donald Trump	42309	21750	3955	6540

Out of the 143,000 articles in the dataset, around 50% of them have length between 360 and 965 words (Table 3.2). The length of the news articles varies, therefore in Section 4.1.3 we tested how the article length could impact the performance of different authorship attribution approaches.

TABLE 3.3 shows the number of times the presidential candidates' names were mentioned from January to November 2016. In either political inclination, Donald Trump received more exposure than Hillary Clinton. This is a form of coverage bias discussed in he literature (D'Alessio et al., 2006). In Section 3.13, we explain how do we control the topic of the news articles and why the names of the presidential candidates are important. In column 4 (which shows the number of news articles related to the 2016 election for each media company used in this study), the left-wing media companies are CNN, Vox, Atlantic, and New York Times. The right-wing media companies are Breitbart and Fox News. On both political inclinations, Donald Trump had more exposure than Hillary Clinton in the media.

Table 3.4. *Sample News Article*

ID	17283
Title	House Republicans Fret About Winning Their Health Care Suit - The New York Times
Publisher	New York Times
Author(s)	Carl Hulse
Publication Date	2016-12-31
Content	WASHINGTON — Congressional Republicans have a new fear when it comes to their health care lawsuit against the Obama administration: They might win.

3.1.2 Sample

Each sample contains the following attributes: the ID of the news article, the title of the news article, the publisher, the author(s), the publication date, and the content. TABLE 3.4 shows one sample news article from New York Times.

Different publishers have different standards for formats. For instance, New York Times tends to put all upper-cased location in the beginning of the content. This feature is very likely to be captured by the algorithms, and according to the literature, how authors formatting the texts is also one of the stylistic features in authorship attribution.

3.1.3 Filtering and Partition

We created the training dataset with steps decribed below.

Firstly, we extracted the news articles related to the 2016 election from January to November 2016. The way to filter the articles is by matching the keywords: Donald Trump and Hillary Clinton. If an article contains either keyword, it will be selected for this study. If we only use the last names of the presidential candidates, their family members (husband, children, etc.) would likely to be included. Using full names might excluded some relevant articles, but can improve the precision of filtering.

Secondly, we labeled the publishers' political inclinations based on Jurkowitz et al. (2020). To make sure the data was well balanced, we selected CNN, New York Times, Vox, Atlantic, Buzzfeed News, and Washington Post as the left-wing group. Fox News and Breitbart were chosen as the right-wing group. The final distribution of the data is shown in TABLE 3.5. The left-wing and right-wing have approximately the same number of articles.

To check whether the news articles were highly similar in topic, we used principal component analysis (PCA) to visualize the word vectors on a 2-D space.

Table 3.5. *Data Distribution*

	# of articles	**percentage**
Left-wing	8490	49.53%
Right-wing	8649	50.47%
Total	17139	100%

Figure 3.2. PCA on media inclinations

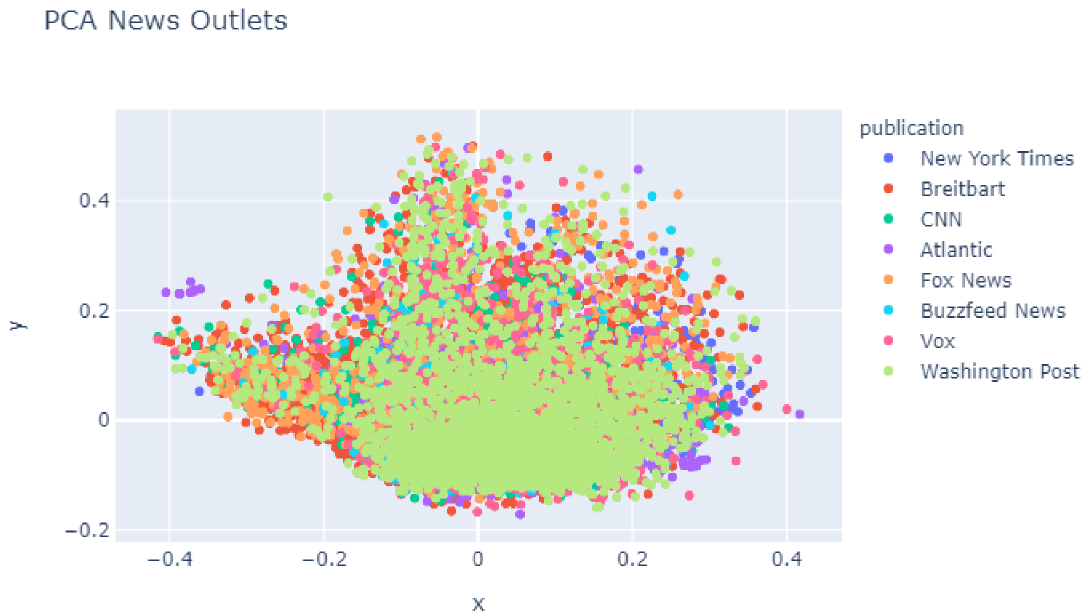

Figure 3.3. PCA on publishers

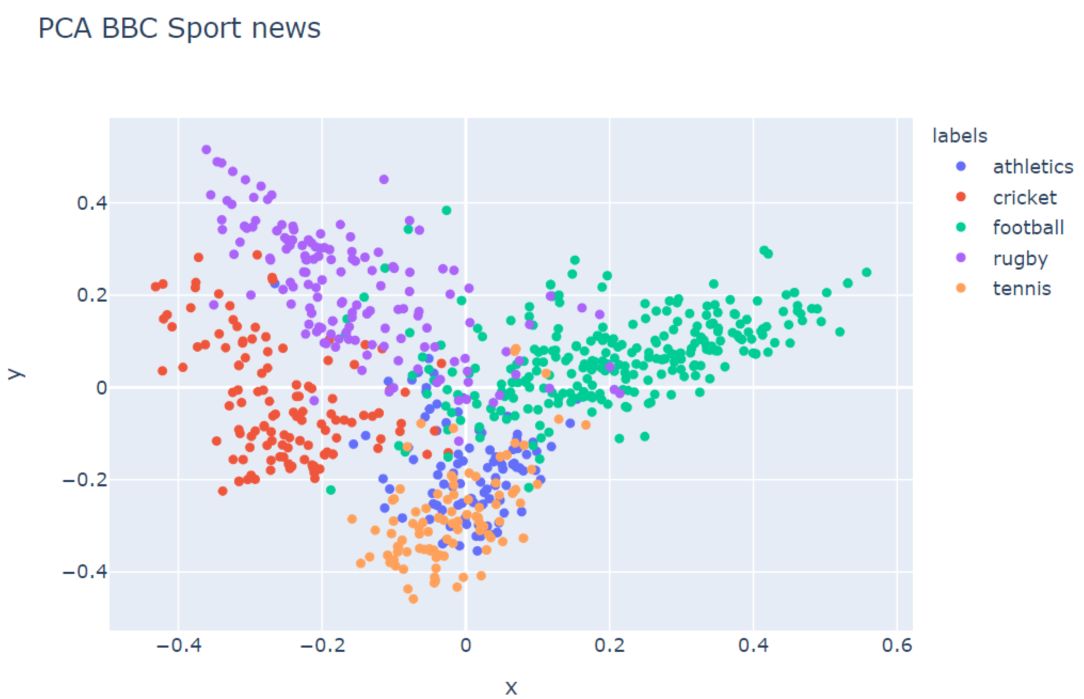

Figure 3.4. PCA on different topics

Figure 3.2 shows the PCA graph for different media inclinations of the resulting dataset. Figure 3.3 shows the PCA graph for different publishers. For each publiser, the top 10000 word-level (single word) features were converted into the TF-IDF vector. The overlaps of the dots with different colors indicate the news articles are similar. For comparison, Figure 3.4 is the PCA graph for news articles published by the same publisher (BBC News) with different topics. The dots on Figure 3.4 are clustered based on different topics and each of the clusters is separated from others clearly. The PCA graphs indicate that the the news articles selected in our study are highly similar in topic, therefore the classification algorithms should rely more on the stylometric differences.

For machine learning models, cross validation is a common step to reduce the variation of the results. We will apply cross validation to all machine learning models.

NPR (Left), New York Post (Right), and Reuters (Center) were excluded from the training dataset purposely. If we assume that there are shared stylometric features in each political inclination, we can test this assumption by running the classification models on news articles published by media outlets that were not in the training dataset.

3.1.4 Text Pre-processing Procedures

The news articles in the dataset were crawled from the Internet and possibly contained unwanted information like URLs, HTML tags, special characters, etc. The classification models also require certain level of text-cleaning before training. Thus, text pre-processing is necessary to be applied. The most common text pre-processing techniques include:

- Remove URLs
- Remove HTML tags
- Remove emojis
- Remove special characters and non-English words
- Convert to lowercase
- Lemmatization

- Remove duplicated spaces
- Remove the names of the media companies
- Remove certain names (Trump, Clinton, etc.)

Due to the variations in formats and the noise generated during data collection, applying all techniques above can be challenging and time consuming. For each of the news articles, we removed special characters and the names of the media companies as the basic text pre-processing steps. Then we used lemmatization, lowercase, and removal of names in Section 4.1.2 and observed how the text pre-processing procedures could impact the performance of different models. The text pre-processing algorithms were provided either by NLTK (Bird et al., 2009) or Stanza (Qi et al., 2020).

3.2 Methodology

The following sections will discuss the experiment design and the measure for success.

3.2.1 Experiment Design

Figure 3.5. shows the pipeline of this study. We conducted authorship attribution experiments on individual authors, media publishers, and two sets of publishers grouped by their political inclinations. We hypothesize that the style of individuals determines the style of organizations or groups. By comparing the salient features and results we can test this hypothesis and can come closer to an understanding whether stylometric classification must take stance into account.

We used support vector machine, logistic regression, random forest, and deep learning models to perform the authorship attribution tasks. The features were vectorized using the TF-IDF algorithm so that they could be passed into the machine learning models for further classification. To avoid selection bias during the data partition stage, we applied 5-fold cross validation on every machine learning experiment and reported the average score in the result section.

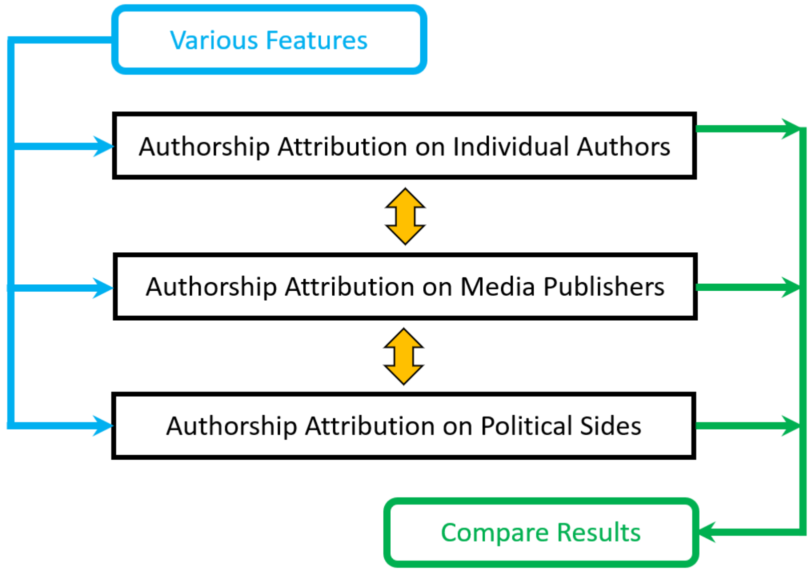

Figure 3.5. Pipeline

As there is no clear boundary between style, stance, or topic, in order to run a clean experiment we control for topic by following the approaches outlined in Section 3.1.3, and select word-level features, character-level features, and part of speech (POS) features (POS tags can be produced by Stanza (Qi et al., 2020)) that we believe contain different amount of non-style information. Based on common sense, word-level features can contain sentiment or stance information while POS features most certainly do not. POS features can, however, contain style information. Character-level features break down some of the word choice information (carrying stance) but still contain more non-style information than POS features. By combining different features with the same classification models, we expected to learn how much the style differences contributed to the classification by comparing the different results generated by each type of features.

We also introduced two extra factors in this study: text pre-processing and article length. As we mentioned in the previous section, the news articles could contain different kinds of noise and their lengths varied dramatically from several hundred tokens to thousands of tokens. To comprehensively understand the capability of our models, we performed controlled experiments on those two factors and observed their impact on the results.

Most of the machine learning models provided convenient ways to extract the highest weighted features. Features accounting for the differences between the political inclinations explain the results of the machine learning models. By looking into the highest weighted features in the machine learning algorithms, we expected to learn some specific stylometric patterns favored by each political inclination. The analysis of features could reflect the degrees of explainability of the machine learning models.

3.2.2 Measure of Success

The classification models measure the results by the accuracy score. The accuracy of the classification is measured as the total number of correct classified news articles divided by the total number of news articles. For machine learning approaches, shuffling and cross-validation should also be applied to make sure the results are not affected by selection bias during the sampling stage.

One common type of baseline is to compare the classification accuracy with the chance of random guess. For authorship attribution with individual authors, the chance of random guess is low and the difficulty of the task is high (because there can be at most 400 different candidate authors). For authorship attribution on media companies and the political leanings, the chance of random guess is higher (12.5% for media companies, 50% for political leanings), therefore better classification accuracy are expected accordingly. The results from each authorship attribution task should reflect the difficutly level of the task as well. Overall, the higher the classification accuracy is, the larger the stylistic differences should be.

Beyond the classification accuracy, the more important measure of success is that what the classification results are based on. The statistical approaches and the neural network models rely on different features. By comparing the results from different approaches, some linguistic features that dominate the result should be identified with a thorough analysis.

CHAPTER 4. RESULTS AND DISCUSSIONS

This chapter describes the results of the experiments outlined in the previous Chapter.

4.1 From individual to group style

We first perform the experiments on individual authors. We applied support vector machine (SVM) models with various features to perform authorship attribution tasks on 100 individual authors. Table 4.1 shows the results from word-level, character-level, and Part-of-speech (POS) features.

Table 4.1. *Authorship Attribution with 100 Authors*

Feature	Accuracy
Single Word	61.66%
Word-2-gram	**64.56%**
Word-3-gram	61.13%
Char-2-gram	29.35%
Char-3-gram	57.03%
Char-4-gram	**59.43%**
POS-2-gram	29.41%
POS–3-gram	50.74%
POS–4-gram	**56.91%**
POS–5-gram	55.62%

The results show that the word-level features generally performed better than character-level features and POS features. This is not surprising as we expect authors to be identified by certain words or phrases that are characteristic of their writing. SVM with word bi-gram feature achieved 64.56% accuracy. Character 4-gram and POS 4-gram have similar accuracy. With 100 authors, the chance of random guess is 1%. The accuracy achieved by the models easily exceeded random guess.

The 17,000 news articles in the dataset are written by 1677 different authors. We thus wanted to find out if the same method and features used in Table 4.1 also worked on authorship attribution tasks with different number of authors. We thus wanted to compare the method for a smaller number of authors and a larger number of authors.

The mean number of articles published by each author is 9.26, but the median is only 1, which means more than half of the authors only published one article. We exclude authors that do not have enough samples, and experiment with the top 400 authors. We also opportunistically select 50 to be a number tested for the smaller scale.

Each of the top-400 most productive authors published at least 5 news articles (which is convenient for 5-fold cross validation), while each of the top-50 authors published at least 33 articles. All of these articles were used in our experiments. We applied the SVM model with the best performing word-level and character-level features on authorship attribution tasks with different number of authors (see Table 4.2).

Table 4.2. *Authorship Attribution on 50, 100, and 400 Authors*

	chance	word-2-gram	char-4-gram
50 authors	2.00%	77.31%	72.56%
100 authors	1.00%	64.56%	59.43%
400 authors	0.25%	48.76%	44.72%

The results suggest that authorship attribution tasks with fewer authors typically produced higher accuracy, which is consistent with previous findings and matches the results from various past literature. The accuracy decreased when increasing the number of authors from 50 to 100 to 400. On the other hand, the difference between random guess and the calculated accuracy actually improved as the number of authors increased. The results also show that the word bi-grams still performs better than character 4-grams.

Table 4.3. *Authorship Attribution on 8 Media Companies*

Feature	Accuracy
Word-2-gram	**80.34%**
Char-4-gram	73.63%
POS-4-gram	74.61%

Table 4.4. *Statistical models (SVM, Random Forest(RF), Logistic Regression(LR)) for word, character and POS n-grams*

	2-gram	3-gram	4-gram
Word-SVM	**0.9178**	0.9003	0.8751
Word-RF	0.8851	0.8732	0.8669
Word-LR	0.9087	0.9038	0.8827
Char-SVM	0.8375	0.9040	**0.9052**
Char-RF	0.8015	0.8071	0.8405
Char-LR	0.8101	0.8769	0.8895
POS-SVM	0.8041	0.8494	**0.8657**
POS-RF	0.7800	0.7763	0.7802
POS-LR	0.7737	0.8312	0.8489

The style of each political inclination (assuming binary classification) comes from the style of each media company supporting the similar political views. The SVM model with the best performing features in previous tasks can also differentiate the media companies, as shown in Table 4.3. Out of the eight media companies (shown in Table 3.1 Column 4), the chance of random guess is 12.5%.

Table 4.3 showed how the word-level, character-level and POS features performed on the task of identifying the publisher of a news article. There are only eight different media companies in our dataset, thus the task is considerable easier than before. The best performing feature (word bi-grams) provided above 80% accuracy in this task, and, again, corresponds to word choices. One should notice, however, that the accuracy does not decrease by a large amount when we consider POS or character 4-grams. In other words, assuming that POS does not carry any stance information, the media companies do have differences in style.

4.1.1 The style of political inclinations

The previous experiments supported our assumption that the style of individual authors can shape the style of an organization. The goal of this study is to discover the style features that characterize political inclinations. Therefore, we trained multiple machine learning models to classify the political inclinations of the news articles (see Table 4.4).

From the results shown in Table 4.4, it is clear that word-level features can differentiate the political inclinations well. The best-performing model (Support vector machine and word 2-gram) achieved 91.78% accuracy on a balanced dataset. Considering the way (which we described in Section 3.1.3) the news articles were labeled, 91.78% can be considered an almost perfect result.

Word-level features were classified as stylometric features according to Joula (2008), but Joula did not specify whether the word-level features also contain information more than just style. One can argue that some of the semantic information was also well preserved in the word-level features, and such semantic information reflect not only the writing style but also the topic, sentiment, and other characteristics that are beyond style.

Breaking the words into character n-grams is one way to reduce some of the semantic information preserved by the words. Table 4.4 showed that the best performing character-level features (character 4-grams) still achieved an excellent accuracy (90.52%). Shrestha et al. (2017) suggested that character-level n-grams were able to capture the stylistic features (like verbal filters) of the texts, but Shrestha et al. did not mention whether the character n-grams also stored semantic information. Zhang et al. (2015) used character-level features to classify the topics of the documents and showed that character n-grams performed just as well as word-level features in terms of identifying the topic of a given text. This makes it plausible that char level n-grams, while lower in accuracy, still carry some features of stance-taking.

Character N-grams potentially decomposed some semantic information, but the results from character-level features and word-level features were still highly similar. To further explore the role of style and stance in identifying the political inclinations, we converted words to their parts of speech (POS) and used POS n-grams in order to dissociate the semantic information from the texts. Part-of-speech carries mostly syntactic information which is closely related to writing style. Examples below demonstrate how a regular sentence looks in its POS form.

“Thursday night, Democrats did what party people do in Brooklyn.”

“PROPN NOUN PUNCT PROPN VERB DET NOUN NOUN VERB ADP PROPN PUNCT”

The results of classification using POS n-grams are shown in two bottom rows of Table 4.4. The results strongly indicate that there was a significant stylometric difference between the left-leaning and right-learning media outlets. After removing the semantic (stance-carrying or topic-carrying) information of the news articles, the classification accuracy only decreased by approximately 4.5% (comparing with the best performing word 2-gram support vector machine model). The results suggest that the semantic differences played a relatively minor role during the classification process.

4.1.2 Text Pre-processing is Optional

For the word-level and character-level features, we can apply different types of text pre-processing procedures to observe the effect of normalize the texts. Text pre-processing procedures are unnecessary for Part-of-speech features because the final Part-of-speech sequences remain the same with or without text pre-processing. Below are the text pre-processing procedures we applied to the raw text:

- Converting all words to lower cases
- Lemmatization
- Removing entity names (replacing politicians' names and news publishers' names with an out-of-vocabulary name - Perotinus)

We applied text pre-processing procedures to the best models in the previous tests. Table 4.5 shows the effect of the text pre-processing procedures.

The results in Table 4.5 show that the models are not affected by the text pre-processing procedures we introduced above. The same word-level bi-gram model and character-level 4-gram model only suffered 0.7% and 1.4% reduction in accuracy after applying all 3 text pre-processing procedures to the raw text. Therefore, the text pre-processing procedures are highly optional in the experiments similar to ours.

Table 4.5. *Accuracy of best models after text pre-processing*

	Word 2-gram SVM	Char 4-gram SVM
No Cleaning	0.9178	0.9052
lower case + Lemmatization	0.9078	0.8949
lower case + Lemmatization + removing names	0.9099	0.8910

4.1.3 Length as a Variable

The length of the input text is also a factor we wanted to explore because the lengths of the news articles vary dramatically from several hundred words to thousands of words. Observing how length affects the performance helps us understand the strength and weakness of the models. In this project, we randomly selected a segment with certain length from a news article and repeated this process five times. Repeating the randomized process could potentially reduce the selection bias compared with selecting from the beginning or the ending of the text.

Table 4.6 shows the results of how the length can affect the performance. Luyckx & Daelemans (2011) indicated that short text fragments was a challenge in authorship attribution. Our results also indicate that longer input texts contain more features that can be used for classification.

Table 4.6. *Best Models with Different Input Length*

	Word 2-gram SVM	POS 5-gram
100 tokens	0.7381	0.6747
200 tokens	0.7811	0.7129
400 tokens	0.8528	0.7841
500 tokens	0.8800	0.8059
Full Length	0.9178	0.8720

4.1.4 Neural Network Models

We applied neural network models along with character-level N-gram features because both Zhang et al. (2015) and Shrestha et al. (2017) used the character-level features in their studies and confirmed the effectiveness of the character-level features. The neural network models have the following hyper-parameters:

- The type of the neural network (convolutional or recurrent)
- The type of the feature (size of the character-level N-gram)
- Input length (INPUT_LENGTH)

- Embedding dimension (EMB_DIM)
- Network Layers
- The type of each layer (GRU, MaxPooling, Dense, Dropout)
- Number of neurons in each layer
- Activation function in each layer (use SoftMax for the last layer and ReLU for all dense layers)
- Training epochs (consider early stop if the models show apparent sign of overfitting)
- Kernel size and stride size (for convolutional neural networks only)

Figure 4.1 and 4.2 show the neural network structures for both CNN and RNN.

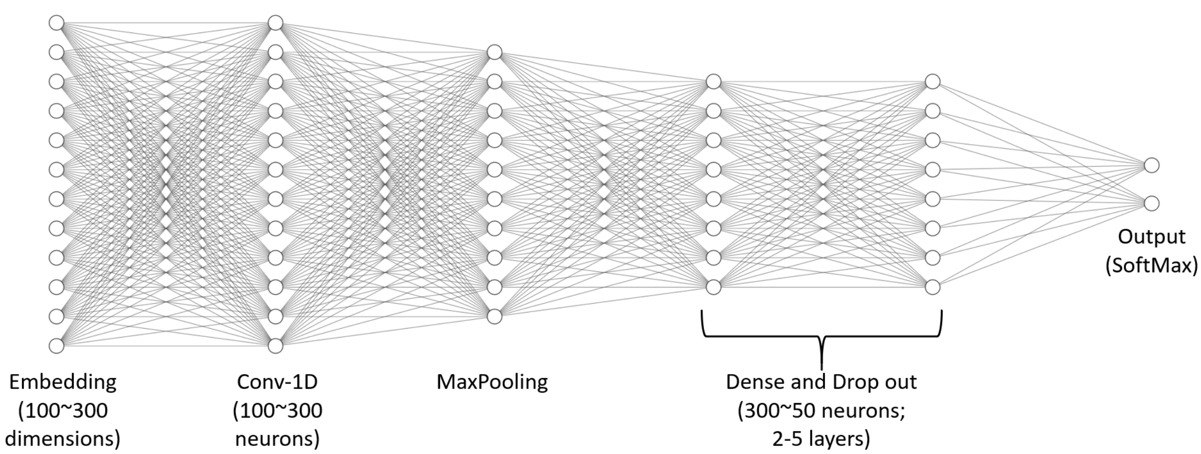

Figure 4.1. CNN Model Structure

Due to the complexity of the hyper-parameters and limited computation power we had, only limited combinations were tested. Table 4.7 shows the results from neural network models with different hyper-parameter settings. We will explain the logic behind the hyper-parameter tuning process in the next paragraph.

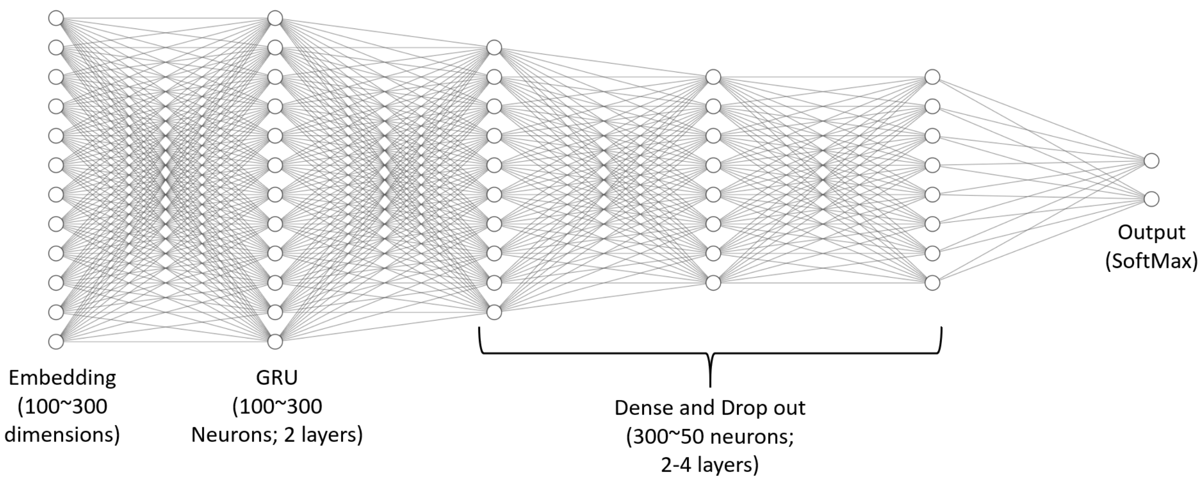

Figure 4.2. RNN Model Structure

We started with two large models on character-level bi-gram features. The input length for both models was 600 tokens (bi-grams). Any input articles with less than 600 tokens would be padded with placeholders (in this case, "0"s). The CNN and RNN models had different layers therefore the parameter sizes were different. Even though the RNN model had more parameters, it performed slightly worse than the CNN models. Both Models suffered from overfitting issue, thus they were only trained for five epochs.

Table 4.7. *Neural Network Models with Different Hyper-parameters*

Model	Feature	INPUT_LENGTH	EMB_DIM	Layer	Parameter	Epochs	Accuracy
CNN	Char-2-gram	600	300	9	3.70 M	5	79.54%
RNN	Char-2-gram	600	300	8	4.33 M	5	77.04%
CNN	Char-2-gram	600	100	6	1.07 M	10	77.38%
RNN	Char-2-gram	600	100	6	1.14 M	5	77.37%
CNN	Char-2-gram	300	100	9	1.70 M	10	78.47%
CNN	Char-2-gram	300	100	6	1.07 M	10	77.53%
CNN	Char-3-gram	600	300	9	3.70 M	10	83.32%
CNN	Char-3-gram	600	300	6	3.07 M	5	82.79%
CNN	Char-3-gram	1000	300	9	3.70 M	10	84.21%
CNN	Char-3-gram	1500	300	9	3.70 M	10	84.49%
CNN	Char-3-gram	3000	300	9	3.70 M	10	85.00%
CNN	Char-4-gram	1500	300	9	3.70 M	10	**86.01%**

The second group of comparison was on two smaller models. We removed some dense layers, reduced the embedding size, and used fewer neurons in each layer. The RNN model and CNN model were overall in similar size. Their performances were also similar. However, the RNN model took about five times longer to train and to do inference because of its sequential structure. Since the RNN models did not show better performance than the CNN models but took much longer in training, we decided to only continue with CNN models. For the third group of tests, we reduced the input length and observed how it could impact the performance for both large and small CNN models. The changes in results suggested that shorter input length did not help increase the performance of the large model. It did increase the performance of the small model by less than 0.2% but the small model still performed worse than the large model.

The fourth group of the comparison switched to character-level tri-gram features. At the same setting, model with the tri-gram features gained about 4% increase in performance. By comparing the large and small models, the large one performed slightly better. From now on, we would continue with the large models (with 9 layers and 3.7 million parameters).

We kept increasing the input length and gained small increase each time. This phenomenon matched with what we learned from the previous experiments (shown in Table 4.7). However, we could not infinitely increase the input length because 3,000 tokens already surpassed the maximum length for many articles. Also, as the input length increased from 600 to 3,000, the model took 3 times more training time.

The last test was on the character-level 4-gram features. Compared with the same model using tri-grams, model with 4-grams gained about 1.5% increase in accuracy. By observing the differences between bi-grams, tri-grams, and 4-grams, it was possible to keep increasing the performance by using larger N-grams. However, the purpose of using character N-gram features was to reduce the semantic information in words and try to restrict the classification to be based on mainly stylometric features. Using larger N-grams directly conflicts with the goal of the project.

So for, the hyper-parameters that could improve the performance were the input length and the size of the N-gram. Since we could not infinitely increase either of them and the computation assumption grew rapidly, we decided to stop the deep learning tests at this point.

One of the reasons to explain why the traditional statistical machine learning methods performed better than the sophisticated deep learning methods would be that we failed to find the better hyper-parameter settings for the neural network models. It could also be attributed to the characteristics of the dataset which did not favor the neural networks.

Support vector machine (SVM) and other statistical machine learning models are known to work well with small dataset. After a training-testing split, Our training dataset only contains about 13,000 samples, which seems to be inadequate for deep learning. Although both Zhang et al. (2015) and Shrestha et al. (2017) applied convolutional neural network (CNN) models with character-level features and achieved respectful results, there were 2 conditions we did not have in our experiments.

Zhang et al. (2015) tested the CNN models on several document classification datasets, and even the smallest dataset was 10 times larger than our training dataset. Figure 4.3 and 4.4 showed that the CNN model converged fast and started to overfit after 2-3 epochs of training. While our smaller neural network models contained only 1 million parameters and the larger models contained more than 4 million parameters, they still overfitted fairly quickly. One of the possible ways to fight against the overfitting problem would be increasing the size of the dataset, which we currently are not capable of. Shrestha et al. (2017) used CNN models on Twitter texts which belonged to a different genre. The Twitter texts were informal, thus the character-level CNN models captured verbal fillers (like *"Uhm.."*) and internet expressions (like *"XD"* and *"lol"*) which did not exist in formal news articles.

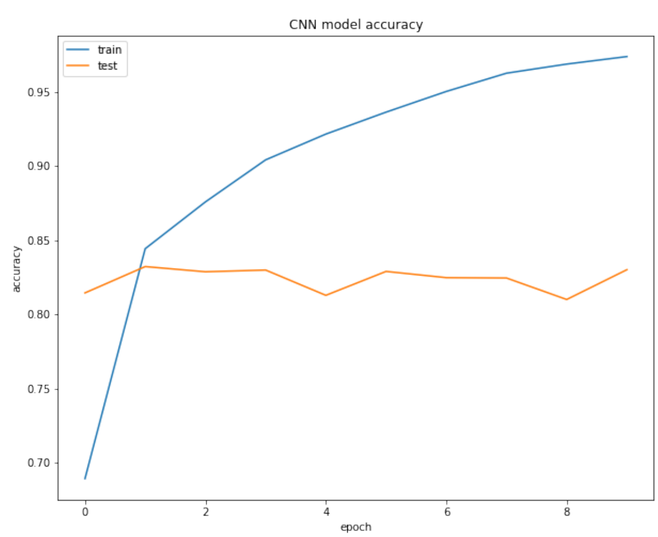

Figure 4.3. CNN model accuracy

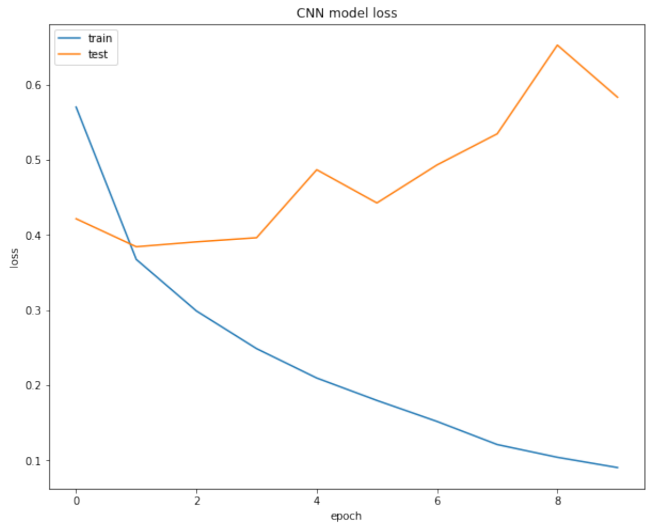

Figure 4.4. CNN model loss

Table 4.8. *Keywords Counts for Both Inclinations*

Keywords	Left-leaning	Right-leaning
Mr. Trump	1201	9010
Donald Trump	12043	16939
Mrs. Clinton	343	3293
Hillary Clinton	10193	11499

4.2 Analysis on Features

The following sections will discuss the the descriptive information revealed from various authorship attribution methods.

4.2.1 Word and word N-grams

Even though the classification results on Part-of-speech tags suggested the existence of stylometry differences between the media outlets, we still wanted to find out what specific stylometric differences caused the division.

Word frequency is one of the features that can be easily interpreted. Table 4.8 shows the word frequency differences we manually selected from the results. The right-leaning media tended to use titles to address the presidential candidates. Similar patterns also happened in other titles (like *"President."*, *"Senator."*, *"Secretary."*, etc.) which were favored by the right-leaning media. Addressing a person with or without title is only the personal preference of the author, which does not change the overall meaning of the sentence.

The word N-gram coefficients in Table 4.9 showed some concerning results. Among the 20 highest-weighted features, many of them (bold Italic entries) seemed to be purely format features. Since the news articles were crawled from the web pages, features like *"read rest"* and *"follow twitter"* were treated as part of the content of the news articles. Such minor noise was difficult to detect and clean during text pre-processing, but it did play a very important role in the classification algorithm.

Table 4.9. *Potential Format Features*

feature	**weight**		**feature**	**weight**
read rest	4.795745		***com follow***	2.776307
follow twitter	4.397071		***latest coverage***	2.737312
news com	4.057425		***headline 2016***	2.737312
frontrunner politician	3.526376		***latest headline***	2.736402
news daily	3.503085		election biggest	2.734587
press contributed	3.300977		name politics	2.734587
prediction map	3.056871		***see latest***	2.699488
democrat party	3.039265		biggest name	2.643088
nominee politician	2.958779		politics see	2.597096
illegal alien	2.821554		mainstream medium	2.522064

Further analysis shows that the word-level bi-gram *"follow twitter"* occurred in 829 left-leaning news articles, vs. 316 articles from the right-leaning media. The similar features (related to Twitter) were also the highly weighted features in authorship attribution tasks on individual authors. The possible explanation could be the authors put their own Twitter account in the web pages and the web crawlers (spider) treated them as part of the content. Besides the format features, some keywords (like *"illegal alien"*) strongly indicate the political lean of the new articles due to particular word choices that are characteristic of a single party.

Table 4.10. *POS-N-Gram Coefficients from Linear SVM*

weight	feature		weight	feature
5.91	NUM PUNCT NUM PUNCT		2.95	PROPN PUNCT PUNCT VERB
4.62	VERB adj NOUN SYM		2.87	PROPN PROPN PUNCT PROPN
4.62	VERB PROPN PROPN ADP		2.67	PUNCT VERB PROPN PROPN
3.92	VERB PROPN PUNCT NOUN		2.63	VERB NOUN NOUN SYM
3.81	PRON ADP PROPN ADP		2.55	NOUN ADP DET NUM
3.58	DET NUM NOUN ADP		2.41	ADP PROPN ADP PROPN
3.48	PROPN PROPN NUM NOUN		2.32	PUNCT PUNCT PUNCT PRON
3.26	PROPN NUM NOUN NOUN		2.31	NUM NOUN ADP DET
3.00	DET VERB PROPN VERB		2.24	DET NOUN ADV PUNCT
2.96	PROPN NOUN PROPN PROPN		2.19	PUNCT DET VERB PROPN

4.2.2 Part-of-speech N-grams

Table 4.10 shows the highest-weighted part of speech N-grams from a linear support vector machine model. We further selected the top-5 features and calculated their occurrences in each media outlets' articles (Table 4.11).

We cannot precisely interpret the corresponding semantic meanings behind the Part-of-speech N-grams without manually locating their positions in the raw text, but some of them still point to the format artifacts. For example, the feature *"VERB ADJ NOUN SYM"* represents "verb adjective noun symbol" which is likely to be a specific format feature used by Fox News exclusively since this feature appeared 333 times in Fox News articles and only 8 times in other articles. In 259 of the Fox News articles, *"VERB ADJ NOUN SYM"* simply means *"See Latest Coverage →"* - an HTML element that is not part of the content. Other features are more likely to be universal but are preferred by one political inclination. *"VERB PROPN PUNCT NOUN"* (which represents "verb proper-noun punctuation noun") appeared in articles from all media outlets. However, the right-leaning media uses this feature 47% more frequently than the left-leaning media.

Table 4.11. *POS-N-Gram Occurrences*

	NUM PUNCT NUM PUNCT	**VERB ADJ NOUN SYM**	**VERB PROPN PROPN ADP**	**VERB PROPN PUNCT noun**	**PRON ADP PROPN ADP**
Fox News	181	333	296	180	53
Breitbart	1665	6	3080	223	582
CNN	104	0	262	38	20
Atlantic	240	0	408	61	26
NYT	121	1	398	12	10
Buzzfeed	60	0	155	16	11
Vox	173	0	277	57	23
Wa. Post	443	1	599	90	52

4.2.3 Validations

Reuters, New York Post (NYP) and NPR were excluded from the training dataset. Since we already have the best-performing models, we would like to validate the models with the Reuters, NYP, and NPR data, previously unseen by these models. Using the best-performing models we trained in this study, we labeled the previously unseen news articles published by Reuters, NYP, and NRP. For each news article, the models provided a label (0 or 1) that represented a political inclination. If more than 66% of the news articles were labeled as either right-leaning or left-leaning, the media company would be labeled with the same political inclination. Otherwise, the media company would be categorized as "center" (neutral). Table 4.12 shows how the classification models label these media outlets.

Table 4.12. *Models on Holdout Dataset*

	Reuters	New York Post	NPR
AllSides	C	R	L
Pew Research	N/A	C	L
Char-level SVM	R	C	L
Word-level SVM	R	R	L
POS SVM	R	R	L

The word-level feature and POS feature reached agreement on the labels. However, none of the classifier labeled the all 3 media outlets the same as the 3rd-party research organizations did. The next section will provide more in-depth analysis based on the above results.

4.3 Model Generalization

Table 4.12 shows how the models performed on unseen dataset. The labels for NPR and New York Post are acceptable since the labels match the conclusions from the literature. New York Post received different labels (center and right) and both of them were supported by different literature. The label for Reuters was less ideal. As a foreign media company, Reuters is supposed to be less biased, yet our models labeled more than 80% of its articles was right-leaning. We did not perform extra tests on more unseen dataset, but based on the results above, we might need to realize that our models may not perform as well on unseen dataset. One of the potential reasons could be the models rely heavily on the format features which are publisher specific. On unseen data, the models trying to compare the similarity of format against the training data. To clarify, according to Joula (2008), format is a type of stylometric feature but it does not explain any politics-related views or bias. Further tests need to be conducted to verify our speculation.

Another variable that is related to the generalization capability is the length of the input. Both neural network models and traditional models seem to prefer longer input texts. The models were trained on online news articles which mostly had 300 to 1000 words (according to Table 3.2). However, in social media platforms, the news tends to be shorter. For example, here is one of the news post from Reuters on Twitter (an abstract of Trevor Hunnicutt, Steve Holland, and Jeff Mason's news article published on November 14, 2020):

"President Trump falsely claimed victory over Democratic rival Joe Biden with millions of votes still uncounted in a White House race that will not be decided until a handful of states complete vote-counting over the next hours or days."

The post only contains 1 sentence, 39 words, and 191 characters, but it is still counted as news and published by one of the traditional cable news companies we studied in this research. The news outlets also publish long reports (like "The President's Taxes" published by New York

Times on September 27, 2020) which have thousands of words. Although our models prefer longer documents, such extra-long articles may exceed the input length restrictions for some models. The variations in form and length are likely to be the weakness of our proposed models.

However, our experiments on text-cleaning did suggest the names of the politicians were almost negligible in the classification algorithms. Based on this finding, we can further speculate that the models trained on the 2016 election news articles should also work on news articles in other election years. One way to test this assumption could be using the current models (and weights) to classify news articles in different years. However, we cannot complete this step due to the difficulties in data collection.

4.4 Difficulties encountered

Due to commercial, legal and possibly political reasons, the majority of the news outlets do not provide clean and computational-friendly news archives. Researchers can only acquire the news articles from the web pages which could contain various forms of noise. Crawling the news articles from the web pages are neither efficient nor convenient. As we discussed in the previous section, the noise in the dataset greatly affected the results. Nevertheless, due to the variations of the forms, it could be considerable time-consuming to locate and remove all the noise. Before we are able to develop more accurate and more robust text pre-processing technologies, the issue of data cleanness could continuously undermine the authorship attribution research.

Another difficulty we encountered in this project was the disagreements in the past literature. In this fast-changing world, the conclusions drawn from the political science research can be either conflicting or outdated. Pew Research is one of the very famous organizations in the politics and media area. The conclusions of media bias from Pew Research were considered to be the gold standard in this project. Pew Research (Jurkowitz et al., 2020) labeled New York Post as the least biased media, while another website (AllSides) labeled New York Post as a far-right media. Both Pew Research and AllSides website published reliable media polarization ratings which are cited by many papers in the politics and journalism area. The main difference between the 2 rating systems is that AllSides has an extra metric which collects opinions about the rating results from the visitors of their website. AllSides rated New York Post as a right-leaning media and 5920 out of 10565 users agreed on the result by November 11, 2020.

4.5 Potential Applications

Our models extracted a great number of format features which contribute significantly to the classification results. In this project, those format artifacts were viewed as one of the stylometric features. However, if we look at the results from a different angle, our models worked fairly well in extracting the format artifacts. Such format artifacts might be categorized as unwanted noise in other text classification tasks focusing on topic, sentiment, or stance. Format is not related to any previous elements but can still impact the classification results. It is possible that we change the objective of our models and use them to detect and remove the format artifacts when they are not needed.

As we mentioned in the Model Generalization section, our models were trained on the 2016 election data and it would be valuable to learn how they perform on the 2020 election. We might be able to learn if some of the media outlets became more biased or more neutral by observing how the ratio of the labels on each political inclination changes. For example, if Fox News had 70% of the news labeled as right leaning in 2016 and 90% labeled as right leaning in 2020, we could possibly infer that Fox News becomes more biased in 2020. The writing styles of each media outlet could change after 4 years because people's writing styles constantly vary in different period in history.

Devlin et al. (2018) proposed a pre-trained language model "BERT" which soon became a popular direction in the natural language processing area. The BERT-like models use the transformer architecture which performed extraordinarily well in text generation and other tasks. Such models are trained on large corpora which contain texts that have specific political leanings. If the training corpora are biased, the texts generated by the models can also be biased. The models are often used as the foundation for fine-tuning and secondary research. The political bias within the models can produce further influence on the related research. Since we already have the tools to label the political position of a given document, we might be able to use them to find out the political leanings of the pre-trained language models.

CHAPTER 5. CONCLUSION

This study applied authorship attribution techniques to study the political leanings of the media outlets. We hypothesized that the style of a particular political side came from the shared style of each media companies supporting similar political views, which in turn can be traced to the styles of individual authors working for the same company. The results from authorship attribution experiments on individual authors, media companies, and political sides showed consistency in terms of features and accuracy, which confirmed our hypothesis.

In this study, we hypothesize that writing style is an independent element in news articles. The boundary between style and other elements (like topic and stance) in writing is vague. Our experiments showed that even without any semantically related information, the difference between the political sides still provided enough predictive information for classification. Further analysis showed that the pattern of some of the non-semantic features can be linked to particular phrases that were used by individual media outlets as they formatted their articles. The format artifacts were counted as stylometric features in the literature, but they seemed to be political-neutral and publisher-specific. Our tests on previously unseen dataset suggested that models relying on format features did not generalize very well, at least on the example of Reuters.

The research is completed in the middle of the 2020 election. Political events arising in this special period make this study seem more valuable. Media polarization has a profound impact on every aspect of the world. If we can identify the appearance of the polarization issue, understand its cause, and be aware of its consequence, we can then propose technical, legal, and even constitutional measures to detect it and govern it. This work is a step leading to a direction with possibilities and potential to counter the impact of media polarization, starting with analyzing the writing style of the political-inclined news articles.

REFERENCES

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 20(5), 67-75.

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS), 26(2), 1-29.

Adomaite, L., & Kačerauskas, M. (2020, June). People Are Laughing At How Hollywood Portrays Other Countries In This Accurate Meme. Retrieved November 08, 2020, from https://www.boredpanda.com/different-places-hollywood-movies

AllSides Media Bias Ratings. (n.d.). Retrieved July 04, 2020, from https://www.allsides.com/media-bias/media-bias-ratings

Alsulami, B., Dauber, E., Harang, R., Mancoridis, S., & Greenstadt, R. (2017, September). Source code authorship attribution using long short-term memory based networks. In European Symposium on Research in Computer Security (pp. 65-82). Springer, Cham.

An, J., Cha, M., Gummadi, K., Crowcroft, J., & Quercia, D. (2012, May). Visualizing media bias through Twitter. In Sixth International AAAI Conference on Weblogs and Social Media.

Bagnall, D. (2016). Authorship clustering using multi-headed recurrent neural networks. arXiv preprint arXiv:1608.04485.

Bergsma, S., Post, M., & Yarowsky, D. (2012, June). Stylometric analysis of scientific articles. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 327-337). Association for Computational Linguistics.

Binderkrantz, A. S. (2012). Interest groups in the media: Bias and diversity over time. European Journal of Political Research, 51(1), 117-139.

Bird, S., Klein, E., Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Boltužić, F., & Šnajder, J. (2014, June). Back up your stance: Recognizing arguments in online discussions. In Proceedings of the First Workshop on Argumentation Mining (pp. 49-58).

Budd, K., & Desmond, S. A. (2014). Sex offenders and sex crime recidivism: Investigating the role of sentence length and time served. International Journal of Offender Therapy and Comparative Criminology, 58(12), 1481-1499.

Burrows, J. (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. Literary and linguistic computing, 17(3), 267-287.

Caliskan-Islam, A., Harang, R., Liu, A., Narayanan, A., Voss, C., Yamaguchi, F., & Greenstadt, R. (2015). De-anonymizing programmers via code stylometry. In 24th USENIX Security Symposium (USENIX Security 15) (pp. 255-270).

Cortes, Corinna; Vapnik, Vladimir N. (1995). ”Support-vector networks”. Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018

D’Alessio, D., & Allen, M. (2000). Media bias in presidential elections: A meta-analysis. Journal of communication, 50(4), 133-156.

DellaVigna, S., & Kaplan, E. (2007). The Fox News effect: Media bias and voting. The Quarterly Journal of Economics, 122(3), 1187-1234.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. Applied intelligence, 19(1-2), 109-123.

Emmery, C., Manjavacas, E., & Chrupała, G. (2018). Style obfuscation by invariance. arXiv preprint arXiv:1805.07143.

Feldman, L., Maibach, E. W., Roser-Renouf, C., & Leiserowitz, A. (2012). Climate on cable: The nature and impact of global warming coverage on Fox News, CNN, and MSNBC. The International Journal of Press/Politics, 17(1), 3-31.

Gelfert, A. (2018). Fake news: A definition. Informal Logic, 38(1), 84-117.

Goswami, S., Sarkar, S., & Rustagi, M. (2009, March). Stylometric analysis of bloggers’ age and gender. In Third international AAAI conference on weblogs and social media.

Groeling, T. (2008). Who’s the fairest of them all? An empirical test for partisan bias on ABC, CBS, NBC, and Fox News. Presidential Studies Quarterly, 38(4), 631-657.

Guruji, P. (2018). Polarization in Media: Perception and Reality (Doctoral dissertation, Purdue University).

Harmon, M., & Muenchen, R. (2009). Semantic framing in the build-up to the Iraq war: Fox v. CNN and other US broadcast news programs. ETC: A Review of General Semantics, 12-26.

Hasan, K. S., & Ng, V. (2013, August). Extra-linguistic constraints on stance recognition in ideological debates. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 816-821).

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. Literary and linguistic computing, 13(3), 111-117.

Hoover, D. L. (2003). Another perspective on vocabulary richness. Computers and the Humanities, 37(2), 151-178.

Hopkins, D. J., & Ladd, J. M. (2013). The consequences of broader media choice: Evidence from the expansion of Fox News. Available at SSRN 2070596.

Jaffe, A. (Ed.). (2009). Stance: sociolinguistic perspectives. OUP USA.

Juola, P. (2008). Authorship attribution (Vol. 3). Now Publishers Inc.

Jurkowitz, M., Mitchell, A., Shearer, E., & Walker, M. (2020, May 08). 1. Democrats report much higher levels of trust in a number of news sources than Republicans. Retrieved July 04, 2020, from https://www.journalism.org/2020/01/24/democrats-report-much-higher-levels-of-trust-in-a-number-of-news-sources-than-republicans/

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003, August). N-gram-based author profiles for authorship attribution. In Proceedings of the conference pacific association for computational linguistics, PACLING (Vol. 3, pp. 255-264). sn.

Kiesling, S. F. (2009). Style as stance. Stance: sociolinguistic perspectives, 171.

Koppel, M., Schler, J., & Zigdon, K. (2005, August). Determining an author's native language by mining a text for errors. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 624-628).

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 60(1), 9-26.

Krosnick, J. A., & MacInnis, B. (2010). Frequent viewers of Fox News are less likely to accept scientists' views of global warming. Report for The Woods Institute for the Environment. http://woods. stanford. edu/docs/surveys/Global-Warming-Fox-News. pdf.

Kruh, L. (1982). A basic probe of the Beale cipher as a bamboozlement. Cryptologia, 6(4), 378-382.

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. University of Chicago Press.

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary and Linguistic Computing, 26(4), 435-461. & Daelemans:2011& Daelemans:2011& Daelemans:2011

Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. Literary and linguistic Computing, 26(1), 35-55.

Mikros, G. K., & Argiri, E. K. (2007, July). Investigating Topic Influence in Authorship Attribution. In PAN.

Mochihashi, D., Yamada, T., Ueda, N. (2009, August). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (pp. 100-108).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557.

Ouamour, S., & Sayoud, H. (2012, June). Authorship attribution of ancient texts written by ten arabic travelers using a smo-svm classifier. In 2012 International Conference on Communications and Information Technology (ICCIT) (pp. 44-47). IEEE.

Overholser, G. (2006). On behalf of journalism: A manifesto for change. Philadelphia: Annenberg Public Policy Center.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. Information Retrieval, 7(3-4), 317-345.

Posner, Richard A. “Bad News.” The New York Times, 31 July 2005. NYTimes.com, https://www.nytimes.com/2005/07/31/books/review/bad-news.html.

Qi, H., & Hughes, S. (2011, May). A new method for visual stylometry on impressionist paintings. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2036-2039). IEEE.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Sarawgi, R., Gajulapalli, K., & Choi, Y. (2011, June). Gender attribution: tracing stylometric evidence beyond topic and genre. In Proceedings of the fifteenth conference on computational natural language learning (pp. 78-86).

Shrestha, P., Sierra, S., González, F., Montes, M., Rosso, P., & Solorio, T. (2017). Convolutional Neural Networks for Authorship Attribution of Short Texts. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 669–674. https://www.aclweb.org/anthology/E17-2106

Søgaard, A. (2010, July). Simple semi-supervised training of part-of-speech taggers. In Proceedings of the ACL 2010 Conference Short Papers (pp. 205-208).

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35(2), 193-214.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556.

Stuart, L. M., Tazhibayeva, S., Wagoner, A. R., & Taylor, J. M. (2013, October). On identifying authors with style. In 2013 IEEE International Conference on Systems, Man, and Cybernetics (pp. 3048-3053). IEEE.

Westcott,R. (2006). Making Hit Music into a Science. news.bbc.co.uk, http://news.bbc.co.uk/2/hi/5083986.stm?ls.

Wimmer, G., & Altmann, G. (1999). On vocabulary richness. Journal of Quantitative Linguistics, 6(1), 1-9.

Xiang, Y., & Sarvary, M. (2007). News consumption and media bias. Marketing Science, 26(5), 611-628.

Xue, N. (2003, February). Chinese word segmentation as character tagging. In International Journal of Computational Linguistics Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing (pp. 29-48).

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).