Princeton University

Senior Thesis

Classifying News, Satire, and "Fake News": An SVM and Deep Learning Approach

Author: Adam Hare
Advisor: Professor Alain Kornhauser

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor's of Science and Engineering

Department of Operations Research and Financial Engineering
Princeton University

June 2018

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Adam Hare

Adam Hare

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Adam Hare

Adam Hare

Abstract

The problem of false news articles has recently surged to the front of political discussion, particularly in the United States. Misinformation comes from a wide range of sources and major social media companies such as Facebook and Google have taken steps towards reducing the spread of so-called “fake news.” By their nature, many such deceptive news articles are difficult for humans to identify, as there may be conflicting reports, different interpretations of information, or a widespread distortion of the facts as the story circulates. While the notion of objective truth is one best left to the philosophers, it may be possible to make in-roads by studying a related category of articles: satire. Satirical articles often arise as part of the discussion when they are taken as fact by their readers and shared in a way to confuse a large part of the public. Many guides to “fake news" contain mostly websites that claim to be satirical. This is often not easy to verify as the satirical disclaimer may be hidden deep in the website's description. The ability to reliably and automatically categorize this subset of articles from the main body of news would be a useful tool to warn readers not to accept the article as fact.

As standards and norms for politics and society change, so too must standards for news and satire. For this reason, this thesis considers a number of subsets of the data, based on date of publication. By considering both the corpus as a whole and the subsets individually, the goals of this paper are to A) develop an effective machine learning approach to identifying satire and B) see if a changing political and social climate has affected news and satire in a way discernible to a machine learning algorithm.

For the purposes of labeling the data, articles coming from sites explicitly claiming to be satirical are labeled as “satire,” and the rest "serious.” The problem of separating satire from serious news is analogous to separating valid email from spam. For this reason, this paper uses many of the most common techniques from the field of spam filtering such as a Support Vector Machine with a linear kernel. The SVM uses a number of features established other works, chiefly a bag of words, and two new features based on links to Twitter and other websites. This thesis also implements a deep learning approach with a C-LSTM.

The SVM with all features consistently achieved over 99% accuracy, 95% precision, and 96% recall when comparing satirical articles to serious ones. The C-LSTM achieved just under 99% accuracy with about 90% precision and recall. It was found that the "fake news" category is easier to separate from serious news but may share similarities with satire. Lastly, this thesis found that the date of publication is a significant factor in identifying satirical articles and that serious news from the past two years may be more similar to older satirical articles than previously.

Acknowledgments

This thesis would never have been realized without the help of a number of people.

Academically, I want to thank Professor Kornhauser for his thoughtful comments, guidance, and excitement about this project. I want to thank Bernardo Pérez Orozco for helping to inspire my interest in machine learning and for encouraging my pursuit of further education in Computer Science. Thank you to Willow Dressel, the Research Computing team (especially Alexey Svyatkovskiy), and all of those who helped me get the resources I needed to make this a success.

This thesis aside, I would never have made it to or through Princeton without the constant support of my family. Mom, Dad, Emily, you have helped me through every step of the way and I hope my efforts here make them proud. Thank you for always encouraging me to pursue my interests, for trusting me to make the right decision, and for instilling an academic curiosity from an early age. It's gotten me where I am today.

Any acknowledgment would be woefully lacking without mention of my friends from home, from Oxford, and from Princeton. Thank you to Luke, Elizabeth, Jacob, Ptom, Katie, and all the rest from home for giving me a reason to look forward to breaks and Sunday nights. I know no distance will ever be too far for our friendship. Thank you to Simone, James, Amol, Jonathan, Jessica, Haley, and all the other friends I met abroad for all of the unforgettable experiences we shared and continue to share. Thank you to Nick and Kyle for the many hours spent pouring over problem sets. Thank you most especially of my Princeton friends to Courtney and Casandra; you've both helped me learn, grow, and enjoy this place so much that my time here would have been unrecognizably different without you. To all my friends I owe an immense debt of gratitude. You've helped make me who I am today and I can't wait to see where the future will take us.

I feel I owe thanks to Princeton as an institution for all of the opportunities it has provided and all of the ways it has challenged and inspired me. I've had a handful of glimpses of this place as Fitzgerald saw it and they helped make the rest worthwhile. I will truly be leaving with "a love of unseen things that do not die."

Finally, I must thank the artists of the countless of hours of music I've listened to over the course of writing this thesis. This all would have been infinitely more tedious in silence.

This paper represents my own work in accordance with University regulations.

Adam Hare

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"The truth is inconvertible.
Panic may resent it,
ignorance may deride it,
malice may distort it,
but there it is."

Winston Churchill, 1916

1.1 Background & Motivation

Historically, the question of truth has been one without a clear answer. Present two people with the same set of facts and as often as not they will come to different conclusions. Some discussions center around interpretations of events, which could vary significantly without an objective truth. Further complications arise when stories contain fragments of truth but rely on speculation. More facts could come to light that disprove an old story, even if the original story was accurate when it was published. All stories have bias, which can distort facts in to almost unrecognizable forms. These problems arise when dealing exclusively with earnest news stories that, despite inherent bias and unavoidable gaps in knowledge, are trying to accurately report the truth of the matter. It is hard enough for a knowledgeable and motivated person to recognize which sources are reliable and trustworthy even with careful consideration.

If we consider stories meant to be deceptive, the problem seems almost intractable. How does one differentiate a deceptive, untruthful article meant to persuade from a satirical piece meant to amuse? Where does one draw the line between a skewed, biased source and one without any factual backing? Studies have shown that in general, humans perform only about 4% better than random when detecting lies [16]. Additional complications may emerge when generally credible sources misreport or obfuscate the truth. There may even be deep confusion as to whether or not a piece is meant to be taken seriously. To use an example from history, there has been a scholarly debate as to if Machiavelli's *The Prince* is a serious work or a piece of satire. *The Prince* contradicts nearly all of the political philosophy Machiavelli espoused before and after its publication. It is so ruthless that some believe it may have been meant to subtly show the princes of Europe the problems with such governance [20]. In short, drawing these lines is not something most humans can reliably do and is certainly not achievable in general without substantive research.

However, just because a problem is hard does not mean it is not worthwhile. The term "fake news" has captured the imagination and sound bites of politicians, citizens, and media outlets. All sides point the finger at their opponents and fear misinformation could or has already influenced political results. This is not an exclusively American problem either; deceptive news has been used throughout the world and some theorize that governments are attempting to create "information weapons" to use against their international opponents [30]. As such, many are looking for a way to mitigate the effects of maliciously deceptive stories. For instance, major tech companies such as Google and Facebook are trying to discourage "fake news" websites by refusing to sell them ads [38]. This sort of action naturally raises a host of questions in the realm of censorship and the control such major corporations have over online activity. For the moment, this paper will side-step these questions and instead consider the primary question these companies and anyone studying the field must consider: how do we identify deceptive news articles?

As alluded to previously, this is a complex problem. To be practical, any answers must be achievable quickly, reliably, and (as much as possible) objectively. If reasoned about properly, these specifications could be an appropriate use case for machine learning. That is exactly the idea of this thesis - with one caveat. This paper will consider a topic related to so-called "fake news" and compare it against generally assumed to be reliable sources. The topic of primary concern is satire. Satire is a natural topic to consider for three main reasons. Firstly, the data is easy to label. This paper considers only websites that self-label as satirical and assumes that all articles published by such sites should be considered satire. All other news sources will be considered legitimate, excepting sections in serious publications devoted to satire (for instance *The Borowitz Report* in *The New Yorker*). By choosing a data set with natural labels, it is possible to use supervised learning techniques without the laborious and potentially error-prone process of having experts label the data by hand. This also reduces the potential for disputes over labels, which one might expect to arise for the reasons outlined above.

The second reason why satire is an attractive subset to study is that most of what receives the label "fake news" claims to be satirical. Browsing some popular lists of "fake news" sites [19] [31] reveals that many, if not most, of the sites listed contain a disclaimer that they are satirical or meant for entertainment purposes only. The intent here is unclear. There are inarguably certain sites which intend to entertain and challenge, for which the "fake news" label should not be applied. Others may be using the guise of satire as a defense from criticism or legal action as a result of their publications.

The final reason for considering satire is that "fake news" may more closely resemble satire than any serious news [15]. Intuitively, both are meant more to persuade and inspire action than to simply recount events. They may both tend to use simpler, less equivocal language than their serious counterparts. Alternatively, they may try to convey convoluted concepts in a dense manner and so could become even more confusing than serious news. This similarity between "fake news" and satire also ties into the previous concept that some sites skirt the line between the two categories. These intuitions have been backed up empirically by [15]. Clearly these two topics, while distinct, have a shared border which may be open to individual delineation.

This is not to say that *all* satirical articles should be classified as "fake news." Satire has an important role in calling to attention inconsistencies, absurdities, and failings in the world as well as entertaining. This thesis does not attempt to disparage the value of satire by this comparison, but merely hopes to gain some insight into the nature of "fake news." The very fact that "fake news" sources claim to be satire points to a relationship between the two. Broadly, when discussing satire when compared to serious news, this paper will include all satirical articles including those which could be labeled "fake news." When comparing satire to "fake news," this paper will exclude from the satirical corpus those articles which could be identified as "fake news."

In summary, the problem of identifying the accuracy of news articles is becoming increasingly large. This paper aims to approach the subproblem of classifying satirical articles, separating them from serious or legitimate news articles. Hopefully such a project will also yield insights into the larger problem of misinformation and "fake news," serving as a stepping stone to a more complete solution to the problem. This thesis also hopes to consider how the relationship between news and satire has changed over time by evaluating the performance of classifiers when trained and tested with different time periods.

1.2 Goals

The goals of this paper are two-fold. The first goal is to build a classifier that will separate satirical articles from serious ones. This classifier will also be used to classify satirical articles against "fake news" articles and "fake news" articles against serious articles. The classifier will be partially modeled on spam filters (see Chapter 2.2), and will be primarily tasked with separating serious and satirical articles using the entire corpus.

There is the hope that inherent in satirical and "fake news" article is a structure or set of defining characteristics. Intuitively, both types of article are written with a clear agenda. Satire tries to entertain and challenge with its narrative, described in [29] as working with "opposing scripts." In the "fake news" context, the intent is to deceive the reader into believing a fabricated story. What's more, both of these must be effective in their messages. This implies that these articles should tend to converge toward styles that achieve their goals. By comparison, serious news should be about reporting the truth, facts which are limited by the reality of the events. There may be a bias or intent, but one that is constrained by the ground truth. This thesis expects to find and hopes to characterize a common structure in "fake news" and satire that makes it separable from serious news articles because those two categories are limited by their goals while legitimate reporting is not.

The secondary goal of this paper is to briefly consider the change in the relationship between news and satire over time. This will be achieved by training the chosen classifier on the data from one time period and testing on data from another time period. These divisions are meant to reflect the changing nature of news as Americans turn increasingly to online news [13] and news found on social media [14]. If the classifier performs significantly worse when tested on a different time period, that would be evidence that serious news and satire are changing in relation to each other. It would also suggest that classifiers are only useful if trained with recent data. If the classifier does not perform significantly worse, it might indicate that there has not been a large shift in the distinction between news and satire. It might also indicate that the difference between serious news and satire is robust, in that the style and quality of serious news is distinct enough to separate it from satire and "fake news" regardless of the topic. Of course, testing using only classifiers will not allow us to make strong claims about the nature of satire and serious news. At best, it may indicate a direction for further study or some points to be considered in application.

Chapter 2

Literature Review

> "The crisis we face about 'truth' and reliable facts is predicated less on the ability to get people to believe the *wrong* thing as it is on the ability to get people to *doubt* the right thing."
>
> Jamais Casico, 2017

Despite the importance of being able to reliably identify "fake news" and satirical articles, as established in the introduction, there appears to be relatively little literature focused on this specific problem. After addressing the existing literature this chapter will consider works related to spam filters since they serve as a partial basis for this project. The final section covers some related topics, including those that involve categories similar to satire or focus on broader text classification.

2.1 Detecting Satire and Misinformation

Laurel Felt predicted that "there will be mechanisms for flagging suspicious content and providers" [35]. As early as 2015, the need for an automatic and reliable filter for misinformation was being proposed [6]. Citing changes in both the way people are accessing news and how the news itself is being structured, the authors of [6] see an obvious need for "an automated news verification system" to assist journalists in creating content, readers in discerning the truth, and educators in teaching ways to recognize questionable stories. The conversation never turns to censorship; the goal is not to remove provocative stories or misinformation altogether. Rather, the intent is to alert people to the nature of the content they are reading so that they can make an informed decision about how to treat it. The presence or nature of misinformation may itself provide valuable information.

That is not to say that all experts are convinced that machine learning or automated systems can handle this problem. Some believe the root cause is societal and that, in the words of Mike Devito in [35], "we can't machine-learn our way out of this disaster, which is actually a perfect storm of poor civics knowledge and poor information literacy." One may argue that identifying "fake news" is treating the symptom rather than the cause, and that the real effort should be going towards better educating the public. While it does seem clear that teaching people to recognize dubious content on their own would be preferable, it is also much slower, more expensive, and more difficult than simply flagging the content for them. The best approach is likely to work from both ends, by creating a better-informed public and implementing accurate classifiers. Additionally, even if machine learning does not *solve* the problem, it may help to alleviate it.

2.1.1 Defining Satire

To focus specifically on satire, we must first establish roughly what the term means. This is not a trivial task [7]. [1] provides a definition of satirical articles as ones "which tend to deliberately expose organizations, real world individuals and events to ridicule." This definition, while valid, is a bit broad. When relating satire to the issue of "fake news" and misinformation, it is important to note that satire generally contains mistruths, through a distortion or exaggeration of true events or complete fabrication. This hints that satirical articles may have more complex motivations than other, more direct, forms of humor. By focusing on exposing things from the real world to ridicule, satire may help its readers to expose these things to scrutiny as well. Satire may have an intent beyond entertainment, that is to make social or political commentary [7]. Satire may also be very difficult to recognize for readers who lack the proper "contextual or cultural background" [39]. Some true events may be so unbelievable as to appear almost satirical or the satire might be so subtle that those without a high degree of prior knowledge accept it as truth. This paper will focus on sources that self-identify as satire or are widely accepted as such. Admittedly, this may result in unexpected interactions between differing definitions, but represents accurately the mix of articles likely to be publicly shared. Such differences are unlikely to produce significant changes in the results of this work.

[29] also addresses the definition of satire, including in their definition the assertion that satire "must also serve a purpose beyond simple spectacle... some form of critique or call to action is required." Satire needs to be complex according to this definition, combining entertainment with critiques and calls to action. Satire often carefully emulates a serious article but eventually requires a recognition of its true nature to have its intended effect [29]. This distinction is one of the major sep-

arators between satire and misinformation, which is spread without the intention of being recognized as untrue. "Fake news" may attempt to drive action, but of a different kind. Where satire tries to encourage further thought and scrutiny, "fake news" desires knee-jerk reactions. The difference here is that satire is meant to be known as untrue. This requires a reader to consider how the narrative of a satirical piece fits with real world events. "Fake news" is meant to be accepted as fact and further study would reveal it to be false. For this reason, "fake news" pieces must expect immediate reactions without careful consideration.

"Fake news," as one category of misinformation, is divided into three subcategories: "serious fabrications," "large-scale hoaxes," and "humorous fakes" by [28]. This paper will be primarily concerned with the third category as it is mostly closely related to legitimate satire. These subcategories are not necessarily mutually exclusive. If other, more credible sources pick up on humorous fakes and treat them as truth, it may lead to a large-scale hoax. Similarly, since [28] includes tabloids in the category of "serious fabrications," it may be hard for readers to distinguish which claims are satirical and which are simply malicious.

2.1.2 Analytic Approaches

[8] proceeds to divide the detection of deceptive articles into two approaches: linguistic and network. The paper also introduces the bag of words and classifier methods as part of the linguistic approaches, both of which will be used in this project. The bag of words method is a common one in text classification [32] and spam filters [40]. Simply put, this method takes text as input. Each word of the text is then tokenized and the input is represented as a vector that indicates whether or not a given word was present in the text. One could also choose to represent each word in the vector by its frequency or relative frequency in the document. This representation of the text relies on commonalities in diction among input of a certain label. For instance, one would expect spam emails to frequently contain words like "FREE" or "EXCLUSIVE," but not contain words such as "ostensibly" or "stochastic." The bag of words approach is simple, powerful, and frequently used [1] [5] [27]. For the deep learning analysis a related but more complex concept is used. Each word is added to a dictionary and assigned an index. Sentences are then turned into vectors of indices and fed as input to the neural network. This allows the classifier to observe sequential relationships, perhaps distinguishing the use of a common spam phrase "FREE LIMITED TIME OFFER" from a genuine email containing "I'd like to accept your offer of lunch tomorrow but my free time is limited."

The second relevant linguistic approach addressed by [8] is classifiers, which are the main focus of this project. Features extracted from the text, including the vectors from the bag of words approach, can be used to train a machine learning classifier to predict if future input is satirical or serious. The two classifiers most commonly associated with satire detection are Naïve Bayes Classifiers (NBCs) and Support Vector Machines (SVMs) [5][8] [27] [25]. This is likely the case due to the prevalence of these classifiers in spam detection (see Ch 2.2). Classifiers are a necessary step in automating the detection of satire. Classifiers that have enough (high quality) training data are more robust than prescriptive approaches such as filtering based on keyword or site of origin. For a more in-depth look at the classifiers used for this project and why they were chosen, see Ch 4.3. [8] also addresses network approaches to deception detection, which involve using data about how the article is spread to identify its veracity. This project does not use network data because it is harder to obtain, especially due to privacy concerns, and often limits the classifier to a single social network. See Chapter 7 for some suggestions as to how network analysis might be incorporated into an expansion of this project.

In what appears to be the seminal work on automatic satire detection [5], Burfoot & Baldwin recognize that "satirical news articles tend to mimic true newswire articles" and that this mimicry can be "surprisingly subtle in nature" [5]. The pair used 4000 serious articles and only 233 satirical ones, arguing that this split accurately reflected the distribution of articles at the time (2009).[1] has a slightly higher ratio of 2624 serious articles to 171 satirical ones. Another study, focusing on tweets in Spanish, gathered 5,000 each of serious and satirical [25]. In a Stanford study from 2017, researchers estimated that around the 2016 US presidential election there were 159 million impressions on "fake news" sites compared to roughly 3 billion impressions on top news websites [2]. While these numbers do not capture satirical sources directly, they do serve to illustrate that serious news articles from recognizable sources are much more likely to be viewed than other articles. However, [2] also estimated that about 42% of traffic from "fake news" sites came from social media, compared to about 10% from social media for serious news sources. This suggests that "fake news" and related satirical articles have a disproportionate presence on social media. While serious news outlets appear to garner traffic from users who frequently check their sites for updates, "fake news" could rely more on a story being spread virally. Note that this study considers only traffic and impressions, which would not necessarily include people reading headlines on social media.

Burfoot & Baldwin focus on the bag-of-words approach using an SVM. They also employ binary feature weights and bi-normal separation feature scaling (BNS) [5]. For a further discussion of such approaches, see Ch 4.1. The pair also focused on headlines as an important feature, claiming that humans were able to immediately recognize most satirical articles from the headline [5]. Headlines are often all people read before reacting. According to [12], 73% of Reddit posts were voted on (positively or negatively rated) without being read. In polls cited by [29], 40% of Americans admitted to reading only the headline. Thus, from both the standpoint of writers who create the articles and readers who decide whether or not to view an article, the headline deserves special consideration. [29] also observed the importance of headlines, noting that satirical articles are likely to rephrase their headlines in the first line.

It is also possible that headlines in serious news articles are more closely tied to the rest of the text. Serious news articles tend to make one argument, while satirical ones tend to have opposing messages or unexpected developments. Some have attempted to capture this by assigning a binary value to the topical similarity of of the first and last sentences of articles (the Hum feature in [29]), although this did not appear particularly informative for their classifier [29]. Interestingly, the importance of headlines may translate directly to "fake news," as [15] found that "fake news" titles tend to convey much more than their serious counterparts. This likely comes from the above discussion of how frequently people read only the title of articles. "Fake news" outlets seem to rely on people only reading their title without even considering the arguments contained in the body. For all of these reasons, titles and headlines are especially important for the purposes of this paper.

[5] also considers profanity and slang as two lexical features worthy of special notice, with the observation that they are far less likely to occur in serious news articles. Even when serious articles use profanity or slang, the words often appear as part of quotations or in quotation marks. Satirical pieces do not necessarily do this, and may employ profanity frequently for comedic effect. Neither of these features are intended to be especially precise, as it is impossible to compile a current and exhaustive list of all slang or profanities. In [5], profanities are determined using a `Perl` library and slang as a relative count of how often words have slang meanings on Wiktionary. It might also be possible to look at potential misspellings or grammatical errors. It seems likely that satirical articles, especially those that come from less prominent sources, have less exposure to copy-editing and may therefore have more spelling and grammar errors. This was not found in any of the discussed literature; in fact in [25], spelling errors were corrected as part of the preprocessing.

A number of papers consider a measure of sentence complexity [29] [39]. One approach is to consider the effect of punctuation to determine the number of clauses and a broader grammar feature to determine which sort of words are being used. The grammar feature appeared particularly effective in [29] and used somewhat less effectively in [25]. Another approach is to estimate a grade-level style readability index, such as the Flesch Reading Ease, Flesch-Kincaid Grade Level, Coleman-Liau index, Automated Readability Index, and Gunning Fog score. These provide an estimate about the amount of education required to read a given body of text. Many are focused around the number of syllables in each word of the text. This method was employed effectively in [39].

Burfoot & Baldwin also acknowledge that "good satirical news articles tend to emulate real news in tone, style, and content," and so looking exclusively at lexical differences may prove insufficient [5]. To counter this somewhat, the pair measures what they call "semantic validity" by identifying named entities in the articles and counting the number of documents returned by a Google search on the given group of entities. The assumption is that legitimate organizations and entities (such as the Food and Drug Administration) should appear relatively more frequently than fabricated ones or strange combinations of entities[5]. A similar approach termed "absurdity" is used in [29]. This approach is highly subject to online trends and may be quickly disrupted by developing stories. It may also fall prey to the popularity of a satirical story. Some articles may gain enough traction or see enough response articles correcting them that they add validity to themselves in this approach.

Using all of these features, the SVM achieves over 95% accuracy with a recall of around 68% [5]. The researchers remark that the low recall is likely due to the subtle satirical articles but note that the classifier achieves an impressive accuracy given the relative simplicity of the features [5]. Other papers have achieved similar or slightly lower accuracy with increased recall [1] [29]. [27] did better on different datasets than Burfoot & Baldwin, but performed worse on the dataset used by the pair.

2.2 Spam Detection

For this project, the decision was made to model satire detection on common approaches to spam detection. This model, one addressed briefly in [5], is useful for a number of reasons. Firstly, it is a binary classification based on text features in much the same way as spam filtering. Secondly, one can consider the parallels almost directly. If we equate a newsfeed with an inbox, each incoming email is analogous to an article entering the newsfeed. A proportion of these are legitimate and a portion are designed to appear legitimate. As with spam emails, the hardest to identify satirical articles are those that are most effective at emulating legitimate articles. The most damage is done when humans are unable to correctly label the incoming document. For a spam email, this may result in a virus or loss of information. For satire, this causes confusion and misunderstanding. For "fake news," this could change an individual's mind about a topic or convince them to take a particular course of action. It seems clear then that these problems are closely related. Thirdly, the spam filtering approach may prove fruitful because spam filtering has long been a focus of study due to its practical applications. Relative to satire detection, a much larger body of work exists for spam filtering. However, to many spam filtering has been "solved" for a number of years and so has seen a decline in focus. Thus, much of the state-of-the-art in spam filtering may rely on outdated machine learning techniques and could benefit from more modern methods. Even acknowledging this limitation, modeling satire detection as spam filtering seems an effective approach *prima facie*.

The most commonly used approach to spam classification appears to be Support Vector Machines (SVMs). Used frequently in the satire detection literature and notably in [5], SVMs appear to be the most effective for sufficiently large datasets. A comparison in [40] between a number of classifiers for spam filters showed that SVMs consistently outperformed other methods. Notably, the classifiers tested in [40] include a Naïve Bayes (NBC) but not a neural network. A broader survey of spam filtering techniques provided in [3] seems to support the predominance of SVMs in this application. In both of these papers, the main feature is the bag of words as discussed.

2.3 Other Related Topics

There are a few other relevant topics when covering satire detection. A related field to satire is parody, in that a parody is usually an imitation of a specific work with broader, more comedic goals than satire [37]. In a study on parody in [37], the authors conducted an experiment which attempted to label parody videos on YouTube. However, since the material itself (in this case the video) was not textual, the experiment relied heavily on metadata. It did not use a SVM, focusing instead on an instance-based classifier. Even so, the results were more or less consistent with the other satire detection studies, achieving an average F-measure of a bit over 90% [37].

Another related topic to satire is sarcasm. Like satire, sarcasm attempts to convey a message very different from the one it is explicitly sending and may be similarly difficult to detect or understand. It also can be almost entirely contextual; a sarcastic sentence without context may appear as a perfectly valid complement [9]. In general, sarcasm tends to be shorter than satire, usually limited to a few sentences at most. Some experiments have used data from Twitter and Amazon product reviews to build a seeded K-Nearest Neighbors classifier [9]. Although some satirical articles are likely to be short, most will be longer than a tweet or typical product review. Satirical articles also tend to be generated professionally, rather than the user-created content on Twitter and Amazon.

There is also the more general consideration of automated text categorization using machine learning. This is a field with a substantial amount of literature, as text categorization has a multitude of practical applications. Satire detection as considered in this paper will be a single-label categorization problem, meaning that no article can have more than one label [32]. Specifically, it will be the binary case. This decision was made for a number of reasons. Firstly, for simplicity. It is easier and clearer to label whether or not something is satire than to attempt to identify a number of possible labels.

Secondly, it creates a broader and more useful classifier. If news were divided into more labels based on topic (such as politics, sports, local, national, international, etc), this could lead to confusion as to which label is most representative of the article. A news article could be satirical, but also relevant to a number of other topics or possible labels, which would imply a multi-label problem. Such considerations are beyond the scope of this project and would serve only to muddy the waters in the analysis of satire.

Thirdly, the "degree" of satire is not especially important for this work. It would not be difficult to argue that some articles are more clearly or ostentatiously satirical than others and so should be considered differently. There is an analogous problem in spam: spam emails that are requests of money from Nigerian princes sent from a dubious address should be easier to detect than those that mimic a colleague sharing a link or attachment. However, we do not see a separation of these in spam literature [3] [40]. Similarly, the literature on satire as discussed in Ch 2.1.1 does not generally make these distinctions, nor do the majority of the previously discussed satire detection projects. Those that mention this concern "leave degree estimation for the future" [39].

Lastly, this approach allows for consideration of an entire article rather than specific sections, paragraphs, or sentences. Although [39] does acknowledge and consider paragraphs without satirical content in satirical documents, the authors still assign an overall binary classification. This reflects the complexity of satire addressed earlier. Satire is meant to challenge, with multiple levels of meaning and complex ideas. Considering only parts of the document at a time might lead to confusion with related topics such as sarcasm and irony. For all of these reasons, a single binary label makes the most sense for the goals of this paper.

Since the classifier will be using a binary label, this text classification must be by construction document-pivoted categorization and "hard" classification, as described in [32]. Document-pivoted categorization means that the focus will be to assign labels to documents individually rather than finding all of the documents in the corpus that best fit a particular label. This method allows for documents to be added to the corpus, and so is essential for any expansions or further applications of this project. Hard classification means that only one label will be offered, which is mandated by the single-label categorization problem with binary labels [32].

Chapter 3

Corpus

"But that joke isn't funny anymore
It's too close to home
And it's too near the bone"

That Joke Isn't Funny Anymore
The Smiths, 1985

Even the cleverest classifiers are defined by the data on which they are built. Careful consideration must be given to every aspect of data collection to ensure that the corpus used for research is representative of the area of study and well-suited to the techniques that will be used to analyze it. This chapter outlines several aspects of data collection, focusing on the decisions made to include or exclude certain sources and the resulting corpus.

3.1 Sources

Satire is a broad topic and not relegated to any specific format, time period, or location. Examples exist from classical Greece and Rome to modern Japan [7]. Serious news articles are even more ubiquitous. Clearly a subset of this data must be considered in order to make sense of the problem at hand.

The first and most obvious restriction is to limit our analysis only to articles written in English. Analyzing across multiple languages would be immensely difficult, as the majority of features will be based on grammar or word usage that will vary by language (see Ch 4). Additionally, different languages likely imply different cultures, which may have different norms with respect to news and satire. The decision to focus on English comes from my familiarity with the language and many of the media outlets that use it, the availability of articles, and the use of English in existing scholarship. Although "fake news" may be prevalent or growing in other languages, its primary current usage is in the United States and so is predominantly in English. With this decision made, it is likely that many of the same techniques could be applied to various other languages with only minor adjustments. There does not appear to be a significant barrier in implementing similar projects in other languages or special peculiarities in the English language ripe for exploitation, assuming there is enough data for analysis in the language under consideration. This paper will focus on articles originally written in English. While translations from other languages may be available, they are likely to feature translation errors or losses of information. The nuance of satire may be especially hard to convey through translation. Similarly, articles originally written in other languages may come from different cultural standards or contexts and so would require special care to be compared with documents originally in English.

A decision must also be made on which sort of news outlets to consider. In general, the decision was made to focus on outlets that cover a variety of topics, rather than just one. For serious sources, this would mean focusing on *The New York Times* instead of *ESPN* or considering broadly satirical sites such as *The Onion* rather than *The Duffel Blog*, which focuses only on articles pertaining to the military. The intent is to create a broader and more robust classifier. It also prevents classifiers from focusing on technical words specific to a given field rather than words related to satirical or serious articles. Lastly, it provides for the broadest range of sources on both sides, as there are more general news and satire outlets than specific ones. With this in mind, articles are likely to be related to politics, as political events tend to dominate serious news cycles and are often ripe for satire. However, this is in line with the goal of using satire detection as a step towards identifying "fake news," since "fake news" is predominantly political in nature.

Another consideration is the target audience of articles in terms of geography, i.e. national or international news compared to local news. This pertains mostly to serious news articles, as satirical ones tend to have a broad audience with the exception of college humor magazines such as *The Princeton Tiger*. As such, it makes sense to focus on major news outlets that have national circulation, such as *The New York Times*, *The Washington Post*, *CNN*, etc. since they are closest to the style imitated by satire. Additionally, this avoids regional differences in terms or local interest stories. This decision does not consider that local news outlets may have as much credibility and circulation as national ones in certain communities. It is also possible that local news outlets do not have the same degree of copy-editing or the same styles as national ones. It may be the case that local news share more characteristics with satirical articles. Due to the sheer scale of considering local news, these potential considerations have been left to further research.

3.1.1 Considerations from Literature

In [28], nine specific requirements are given for a corpus that aims to detect "fake news." As the paper includes satire as one of the categories of "fake news," these criterion seem a reasonable place to start. Of these, five (mix of satire and serious, text as a medium, reliable labeling, a consistent timeframe, and "pragmatic concerns") are covered explicitly by the construction of this project [28]. Let us now consider the other four in turn. The first two are concerned with similarities between articles, specifically in length and writing style. In other words, *New York Times* articles that span several pages should not be compared to one sentence headlines from *The Onion*, nor should scholarly articles be compared to blog posts. In general, both article length and style can be clear delineators for satire. The size of the corpus and breadth of sources should account for some of these variations without eliminating potentially valuable differences.

Another requirement is to consider how the news is delivered. This project lacks much of this feature, as the data will be gathered directly from the source. A project that looked solely at links posted on social media would be able to associate further data about the user and the text of the post providing the link to the article to place it in context. Here the project will implicitly assume that articles are displayed without context or prior priming. The last requirement to consider is language and culture. This project controls for English so language is not a concern. The question of culture relates to the decision above. A focus on American sources of satire and serious articles may provide a more accurate classifier. Consider also cultural differences: one feature considered may be use of profanity [5]. As is the case with British television, it is possible that the British press is more prone to profanity than the American press. Cultural differences over the meaning of words may also be important here. These may vary across the United States, but sources with a national or international audience should account for this in their publications.

With these cultural considerations in mind, the decision was also made to limit the corpus to articles originating in the United States. This purposefully excludes other nations in the Anglosphere, such as Canada, the United Kingdom, Ireland, Australia, and New Zealand as well as India, South Africa, and other countries with large English-speaking populations. The decision to focus on America comes again from my personal knowledge, the availability of data, and the immediate political concerns. An approach that used data from other nations is certainly feasible, and is in fact studied [29] [39]. On one hand, this may create a more robust classifier which is better suited to the international nature of the Internet. On the other, it may lead to decreased accuracy as different nations may have different spellings of shared words, unique grammar differences, and specific cultural idioms or phrases.

Each of these differences could influence a classifier, especially if the distribution of satirical articles was not consistent across nations of origin. Since there are enough articles from the United States to use for our analysis, the decision was made to focus on those alone.

3.1.2 Different Time Periods

Satire, as a highly contextual style, may not be static in nature. As cultural norms shift, so too must satire. Similarly, as different political and cultural figures fluctuate in popularity, satire may respond with changes in sentiment. It seems intuitive that the relationship between news and satire has evolved over time.

One reason why this might be the case is that the Internet has become an increasingly important part of how individuals get their news and so an increasingly important source of revenue for news outlets [13]. Additionally, more news than ever is being shared via social media [14] [18] [34]. As such, news outlets may be tempted to style articles in a way so that they are more likely to "go viral." This could manifest as more sensationalist headlines, in a number of ways such as shorter articles and more casual language. Some news outlets such as *Buzzfeed* do serious journalism but often rely on "clickbait" titles or "listicles" to garner attention online. The success of these sources may lead to other sources emulating their style or increased research into which type of articles are likely to receive the most clicks.

It could also be the case that events themselves are evolving rapidly. To give an example, the recent populist movement in the United States and the election of "political outsider" Donald Trump has resulted in the abandonment of a number of norms [11]. As norms are cast aside, topics that were formerly the domain of satire may actually be covered in the news. Political norms, as well as rapid technological shifts, may make previously unbelievable events a reality.

Recognizing this complexity presents a number of challenges for building a machine learning classifier. It becomes a trade-off of performance and robustness. Training on a large dataset that includes older articles is more likely to produce a robust classifier that performs consistently across years. Training on a smaller subset of dates will likely increase performance in that time frame, but may drastically reduce performance in other time periods. The hypothesis that the relationship between news and satire is changing could be false. It may be that the style of satire is sufficiently different from the style of serious reporting and that the changes described above are too trivial to disrupt the classifier in a statistically significant way.

To attempt to address these concerns, this thesis will use different training and testing sets. Let us consider three possible divisions of the data. The first will use all available data (2010-2017). The second will consider data from 2010-2013 as "old" and 2014-2017 as "new." This division contains a presidential election in both segments, and each has enough data and variation to be interesting. The final division will consider 2014 and 2015 as "old" and 2016 and 2017 as "new." This will be the narrowest classifier and will be most useful for studying the changing norms associated with the 2016 US presidential election. Note that not all sources, especially satirical ones, existed for the entire 2010-2017 time period.

There are three ways possible ways to consider handling the date field. For training and testing across the entire corpus, dates could be ignored altogether by excluding a date feature. Another approach is to create a boolean indicator for each year, known as one-hot encoding. This explicitly indicates that articles are from different years without explicitly encoding how far apart they are. The third approach would be to consider each year as an integer, explicitly encoding the magnitude of time separation. Integer encoding will be omitted because not all sources have an available date. One-hot encoding allows for this, with simply all of the date fields being zero. Integer encoding does not. Additionally, it is standard practice to use one-hot encoding on such a field because any difference inherent between gaps of years should be learned by the classifier implicitly.

The possible change over time will be evaluated by training and testing on one data set as a benchmark for that classifier and then training on that set and testing on the other. A major change in classifier performance between the two sets would indicate that the articles from each era are distinct. However, if there is not a substantial decline when testing on the other data set, this implies that the relationship between satire and serious news is somewhat robust and changing relatively little over time. When comparing across time the date field will be omitted because testing sets will contain only previously unseen dates and so not provide useful information to the classifier.

3.2 Sources Used

3.2.1 Existing Datasets

A few existing papers provided their data sets, which can be used to compare to the data collected by this experiment and the results it provides. The most recent and comprehensive of these is found in Yang et al [39] at `https://github.com/fYYw/satire`. This data set has over 16,000 satirical articles and more than 160,000 serious articles. However, this dataset does not provide sources or dates for articles. As [39] was published in September 2017, it would be safe to assume that these articles are largely within the range of 2010-2017 that this paper considers. It will not be possible to perform an analysis of change over time with this data, but it will still be a useful benchmark for the classifiers developed in this paper.

3.2.2 Satirical Sources

All satirical sources were gathered using ScrapingHub (`https://scrapinghub.com`). This these considers only sources that are widely known to be satirical or (in almost all cases) contained a satirical disclaimer somewhere on the website. The sources were chosen both from personal knowledge and exploration as well as the list of "fake news" sources provided by Snope's [19]. Many of these sites have been taken down and currently have their domain for sale. Search results for those sites included in [19] tend to be dominated by articles discussing whether or not the site in question is trustworthy. These two factors suggest there is some policing by individuals or corporations in an effort to prevent the spread of misinformation. Of course, some sites may be identified only after a false article has been spread, and even then individuals may be unlikely to fact check on their own. Some sources included in [19] were excluded for the purposes of this paper because they did not contain a significant number of articles or were riddled with formatting inconsistencies that made them hard to parse. We also omit a few satirical sources that were too narrow in focus, such as *Duffel Blog* which provides satirical articles focused exclusively on military life. These are unlikely to have a sufficient number of similar articles in the serious data set. Table 3.1 contains some basic data on the data from satirical sources.

Site Name	# of Articles	% of Corpus	Years	Snopes
The Faux Report	18	0.1%	2017	N
National Report	55	0.4%	2013-2017	Y
Satire Wire	289	2.2%	2010-2017	N
World News Daily Report	316	2.4%	-	Y
Huzlers	547	4.2%	-	Y
The Borowitz Report	832	6.3%	2012-2017	N
The Babylon Bee	1524	11.6%	2016-2017	N
Empire News	2119	16.1%	2014-2017	Y
The Onion	3369	25.6%	2010-2017	N
The Spoof	4063	30.9%	2010-2017	N
Total Article Count:	13,132			

Table 3.1: Satirical Sources

It is now worth discussing a few of the sources provided here. First, there are not dates for *World News Daily Report* or *Huzlers*. This is because neither website shows the date an article was published. Both sites are relatively new, and so all of their articles will fall within the 2010-2017 window considered, but more precise data than that is not available. Another important point is that some of these satire sites are professionally curated, meaning that full time writers contribute articles. These include *The Onion* and *The Borowitz Report* and can be expected to have a higher degree of editorial quality in general. Some other sites are clearly user-generated, in that anyone can publish articles to them. *The Spoof* is an example of this. On one hand, these sites have many articles spanning all of the years under consideration. They are also able to provide a broad range of different writers on a broad range of topics. On the other, they are more likely to be dissimilar from professionally written news pieces, both stylistically and in terms of copy-editing. However, due to the quantity available that spans the entire timeline, they will still be considered for this paper.

Lastly, note that four sources are on Snopes' list of "fake news" websites according to [19]. The largest of these is *Empire News* with over 2000 articles and the smallest *National Report* with only 55. While gathering the data, it became clear that the sites listed in [19] tended to have more vulgar and outlandish articles than those not listed. This may make them easier to classify in general as they do not as closely mimic serious articles. These sources make up 23% of the satirical corpus.

3.2.3 Serious Sources

The sources chosen as "serious" are from a study by the PEW Research Center which provides a list of the top twenty-five most frequently shared online news sources [23]. All sources used by this project are included on that list. These sources all have a strong online presence and a good degree of name recognition. This paper avoided news aggregator sites such as *Google News* and focused on sites that primarily do their own journalism. A detailed breakdown is provided in Table 3.2 below.

The majority of these articles were gathered using LexisNexis Academic (`http://www.lexisnexis.com/hottopics/lnacademic/`), limiting to a few major sources and the appropriate timeframe. Articles from *The Wall Street Journal* consist mostly of abstracts, and so tend to be much shorter than the rest of the corpus.

A few observations can be made from a brief survey of the serious news data. The first is that serious news articles tend to be much longer than satirical ones. This seems intuitive, as a serious topic may easily be worn out, whereas a heavily researched report may be more in depth and comprehensive. It is also the case that many news organizations embed Tweets or other posts from social media in their articles. This is not common in satire, as the Tweets useful for satire would likely have to be fabricated, and so could not be linked directly without revealing that they are false. Serious news sources are more likely to reference their source, in

Site Name	# of Articles	% of Serious Corpus	Years
The Los Angeles Times	429	0.2%	2017
ABCNews	1694	1.0%	2010-2017
The Wall Street Journal	1761	1.0%	2010-2017
CNN	2654	1.5%	2013-2017
Fox News	3780	2.1%	2010-2017
USA Today	11,060	6.2%	2012-2017
The New York Daily News	12,936	7.3%	2010-2017
The New York Post	13,427	7.6%	2010-2017
NBC News	15,035	8.5%	2010-2017
The Washington Post	15,214	8.6%	2010-2017
The New York Times	20,683	11.7%	2010-2017
CBS News	35,692	20.1%	2010-2017
Reuters	43,019	24.3%	2010-2017
Total Article Count:	177,384		

Table 3.2: Serious Sources

that many articles from *CNN* (to give an example) contain *CNN* in the text of the article in phrases such as "the senator, speaking to *CNN*, revealed" or "*CNN* correspondent." It would likely be possible to remove such references, but it is unclear if that is desirable. While these mentions do reveal their source, so too might satirical sources. It may also be the case that satire simply does not reference serious news sources, and so this is a strong indicator of a serious news piece. For this thesis, structural references were removed (i.e. references that appeared in every article regardless of content) but those included as part of the reporting were generally retained. This is a difference from some other satire detectors, which remove all reference to any source in the corpus [39].

It is also worth noting that many of these sites are largely focused online or as a companion to a cable program, while others rely more on print. This may result in different styles. Additionally, a number of these sources are widely regarded as somewhat biased. This is an unavoidable fact of news. Hopefully by considering sources on both sides of the political spectrum this will balance out. It is also reflective of the sources shared as given by [23]. The sources here may be of varying quality, but those known for frequently sharing fabricated stories were intentionally omitted.

The entire corpus is now 190,532 articles and about 7.4% satire. This percentage is comparable to other sources: Yang et al[39] used about 9% satirical articles, while [1] and [5] used around 6%. Because [39] is both the most recent of these (September 2017) and has the closest corpus size (about 176,000 articles), it seems logical to have a similar percentage of satirical articles. It is also clear that for the above dataset, some sources tend to dominate. A few source make up a significant part of the corpus. For satirical sources, this is difficult to avoid because the data is harder to gather and there are simply fewer articles in existence. It may be desirable to reduce some of the top serious sources in order to get a more balanced and diverse set of serious articles. Reducing the number of these articles will also increase the percentage of the entire corpus that is satirical. By randomly selecting 20,000 articles from those sources that significantly exceed this number (namely *Reuters* and *CBS News*) and removing some articles with formatting problems, we get the following arrangement for the corpus as a whole:

Type of Article	# of Articles	% of Corpus
Serious	138,281	91.3%
Satire	13,127	8.7%
Total Article Count:	151,408	

Table 3.3: Corpus

Chapter 4

Method

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

xkcd 1838: Machine Learning
Randall Munroe, 2017

4.1 Features

Although having appropriate data is perhaps the most important part of any machine learning project, extracting features from that data is also a crucial part of the process. Features are simply the aspects of the data which will be given as input to the classifier. The following section will outline the features considered in this project, including why they were chosen and how they were implemented.

4.1.1 Bag Of Words

Perhaps the most naïve and versatile approach to text classification is to consider that different categories of text use different words. One way to measure this is to simply provide a count of the number of times each word appeared in a given text. This is called the bag of words approach and has been shown to be roughly as effective as more complex methods for considering the terms used in a text [32].

The bag of words approach is also well-established in the existing literature, both for detecting spam [3] and satire [1] [5][8]. This paper will consider two typical weighting schemes. The first is binary weighting, in that all words will be given a weight of 1 if they appear in a given text and 0 if they do not. This method is used in [1] and [5]. The second weighting scheme is called term frequency-inverse document frequency (TF-IDF), used by [1] and [29]. TF-IDF is the product of term frequency in a given document and the inverse of the frequency of that term across the corpus. This weighting prioritizes words that are infrequent in the corpus but appear often in a particular document. Both of these weighting methods are implemented in the `Python` package `scikit-learn`.

For the bag of words feature, this project uses a built-in list of "stop words," which are words that occur so frequently in English as to not be noteworthy. These words, mostly articles and pronouns, would only to add noise to the model [1].

4.1.2 Profanity

One additional feature often considered for satirical articles is profanity. In the United States especially, serious news publications almost never include profanity. When they do, the profanity is usually part of a quote and is often censored. Satire has no such restrictions, and often uses profanity to enhance the humorous or mocking nature of an article. [5] used a binary feature to indicate profanity. This variable would take the value one if at least one profane word appeared in an article.

This approach can be problematic for two reasons. The first is that profanity can be subjective. Words some cultures or groups consider acceptable may be offensive to others. This concern is somewhat mitigated by the decision to consider articles from the United States. A second reason is that profanity can be used legitimately if necessary for a particular event. If a politician or celebrity uses a generally considered profane word in a newsworthy way, it may be necessary to include the word in a serious article for the sake of clarity. However, events of this sort should be relatively infrequent and so it seems clear that the majority of articles that contain profanity will be satirical in nature.

To identify profanity, this project will use a list of words partially adapted from the `Python` library `profanity`. This library identifies a number of frequently used terms generally considered profane. However, it also contains many words that are frequently used in normal discussion (such as "sex"). The project will use a list of words based largely on those included in `profanity` but without those that have legitimate uses in a serious news setting. It is important to note that this list includes biological terms, which are not necessarily profane in nature. It does seem likely that even these words would be less frequent in serious news reporting. To compensate for this and the aforementioned concerns, this paper will use a count of the number of profane words rather than a binary indicator. Intuitively, serious articles should use profane words more sparingly than satire even in articles which contain profanity.

4.1.3 Complexity

One possible difference between serious news articles and satirical ones is the level of complexity in the writing. In [15], it is suggested that the body of "fake news" articles tend to employ simpler language than more serious news articles. This effect may be somewhat mitigated by the wider array of news sources used, as one might expect published articles to use a larger vocabulary than those exclusively online. Nonetheless, it seems a worthwhile field to consider, and is in fact addressed in [39].

There are a few ways one might measure text complexity. The first is a simple word count. Upon observing the raw data from serious and satirical articles, it became apparent that serious articles tend to be longer than satirical ones. Specifically, satirical news articles averaged 332 words with a standard deviation of 201 and serious news articles averaged 850 words with a standard deviation of 992. Thus, word count may be a significant factor in predicting satirical articles. The number of unique words will be captured fully by the bag of words approach and so will not be considered individually.

Another approach would be to measure the average number of syllables per word. More complex texts are likely to use longer words, and so should have a higher average syllable count. Syllables were counted using the `textstat Python` library, which has a function that returns the total number of syllables in a document. This was then divided by the word count, as described above. It was found that satirical articles tend to average 1.25 syllables per word, with a standard deviation of around 0.11 syllables per word. Serious news articles average 1.29 syllables per word with a standard deviation of 0.11. Because these are so close, average syllable count may not provide much differentiation. This feature was used in [39].

The final approach this paper will use to consider text complexity is that of readability indices. The majority of these analyze a text to return an appropriate level of reading difficulty, usually corresponding to the grade level necessary to comprehend the text. These grade levels are primarily based on the US education system [4]. A number of different readability indices have been used for satire detection, notably Gunning Fog, SMOG, Flesh-Kincaid [15], Flesh reading ease, Automated Readability, and ColemanLiau [39]. Each of these indices and two more (Linsear Write and Dale-Chall) are implemented in the `textstat Python` library. This thesis will make use of Flesh Reading Ease and Gunning Fog (both used by [15] and [39]) as well as the Automated Readibility Index. The Flesch Reading Ease index is based on the number of words, the number of sentences, and the number of syllables. Gunning Fog uses sentence length and the number of multi-syllable words [4]. The Automated Readability Index (ARI) is based on the number of characters per word and the number of words per sentence. ARI was chosen as a third index because it does not depend on the number of syllables in a word. Determining the number of syllables can be difficult, especially for proper nouns or words that do not appear in the dictionaries used by `textstat`.

4.1.4 Links

One feature not found in the literature is the number of links to outside sources. During formatting, links to internal pages were removed, as were generic links to social media (i.e. "Share this story on Facebook" style links). However, it became clear from a scan of serious news articles that many contain links to other news sources, company or government websites, or embedded tweets. Intuitively, satirical articles would be unable to link to external sources because they would reveal the satirical nature of the article. If a satirical article references a study, the study is probably fabricated for the article. If a serious article mentions a study, it is likely to include a link for readers to access the original document.

By identifying strings beginning with `http://` or `https://`, some estimate of the number of links can be established. Because embedded Tweets do not always contain this, we can search explicitly for Twitter image URLs and uses of the '@' and `'#'` characters. These will be separate features: one will count the number of outside links and the other the number of Twitter indicators found.

4.1.5 Headline

As pointed out in the Chapter 2, headlines may be of particular importance when considering satire. [15] found them as a strong differentiator between satire and serious news, resulting in an accuracy of 75% without considering the body of the article. This paper will consider a few measures of complexity listed above, specifically word count, average syllable count, and the readability indices (Flesch Reading Ease, Gunning Fog, and ARI).

4.1.6 Other Features

One possible feature to consider is the use of slang in articles. As mentioned in [5], one would expect satirical articles to contain much more slang and appear more informal. However, slang is a hard category to consider. It is largely time and context dependent. There are not well-established methods of handling it. Additionally, direct quotations or embedded documents are likely to introduce a number of slang terms to serious news articles. For these reasons, no slang feature will be considered for this thesis.

For a more complete discussion of possible features, see Ch 7.1.4.

4.2 Classifier

Measuring Classifier Performance

This project will use a number of well established metrics to evaluate a classifier's performance. First, consider the following four terms: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Here a label is considered TP if the article's true label and predicted label are both satire and a label FP if the article's true label is not satire but its predicted label is satire. Similarly, the term TN will correspond to articles who have a true label of not satire and a predicted label of not satire, while FN designates articles that are truly satirical that have been predicted to be not satirical by the classifier.

Providing a confusion matrix which contains each of these values for a given testing set yields more insight than the other metrics alone and so will be included when possible. This project will provide four metrics: accuracy, precision, recall, and F-measure. Accuracy is defined as the number of correctly labeled articles over the total number of articles, or

$$\text{Acc} = \frac{TP + TN}{N}$$

where $N = TP{+}TN{+}FP{+}FN$ is the size of the corpus. Precision will correspond to the number of correctly labeled satirical articles over the total number of articles labeled satirical, or

$$\text{Pre} = \frac{TP}{TP + FP}$$

Recall is the proportion of articles satirical articles that were correctly labeled, so

$$\text{Rec} = \frac{TP}{TP + FN}$$

Lastly, the F-measure is a combination of precision and recall:

$$\text{F} = \frac{2 * \text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}}$$

These are the measures found throughout the majority of the literature, including in [1], [5], [25], [29], and [39]. Precision and recall are often a trade-off. High precision means that there are few false positives, while high recall means that there are few false negatives. An increase in accuracy can boost both but often an increase in precision means a decrease in recall and vice-versa. The F measure tries to compensate for this trade-off by providing a metric combining both measures. In general, the relative importance of precision and recall depend on the application. If we planned to censor "fake news" or satirical articles, we might want to boost precision so that fewer serious articles are censored. If the intent is simply to provide a warning, it might make more sense to boost recall - otherwise users might stop thinking critically about "fake news" articles without a warning [24].

4.2.1 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are the type of classifier most frequently used in the satire detection literature. This likely arises from the prevalence of SVMs in spam filters and other established text classification fields as suggested by [5]. Briefly, SVMs use features to map data points to high dimensional spaces. The SVM then finds an equation for a boundary between points of different classes, maximizing the size of the gap between classes. The SVM predicts the class of new data by mapping the data point to the same space and seeing which side of the boundary it falls on. Basic SVMs implemented with `scikit-learn` are used by [5] [29] [39] while [1] [15] use other implementations. Among these, [15] [29] [39] use a linear kernel.

Learning Hyperparameters

A support vector machine with a linear kernel has two hyperparameters. The first is called C in `scikit-learn` and corresponds to the coefficient of the regularization term. This is necessary for any type of SVM. For the values of C, powers of 10 from 10^{-5} to 10^{3} were tested. The other hyperparameter to learn is the class weights parameter. When this is set to `None`, the classifier does not change the weights of class labels. If this is set as `balanced`, classes are given weight proportional to their frequency in the corpus. For this thesis, balanced classes would result in the loss function being penalized more for missing a satirical article than for missing a serious article in the training set. Yang et al [39] found that switching to `balanced` improved performance.

These results are evaluated using `GridSearchCV` provided as part of the `scikit -learn` library. See Appendix A for the code used for training these hyperparameters. Note that all data has been normalized using the `scikit-learn` library. The results using a random 85% of the data are below. For each table, the columns indicate the value of `class_weight` and the rows the value of C. Displayed is the average accuracy of the classifier on the test data using a three-fold cross validation model. Since the bag of words is the largest and most comprehensive feature, hyperparameters were trained on just the bag of words and then the bag of words with the other features. This was done for both of bag of words implementations.

	Bag of Words Only		All Features	
	None	Balanced	None	Balanced
10^{-5}	0.9133	0.9288	0.9194	0.9176
10^{-4}	0.9561	0.9501	0.9674	0.9533
10^{-3}	0.9850	0.9762	0.9879	0.9808
10^{-2}	**0.9883**	0.9862	**0.9907**	0.9893
10^{-1}	0.9870	0.9871	0.9895	0.9895
1	0.9862	0.9862	0.9888	0.9889
10	0.9860	0.9860	0.9887	0.9887
10^{2}	0.9860	0.9860	0.9887	0.9887
10^{3}	0.9860	0.9860	0.9887	0.9887

Table 4.1: SVM Hyperparameters: Binary Bag of Words Accuracy

The best results come from $C = 0.01$ and `class_weight` = `None` for both feature sets, with accuracies of 98.826% and 99.073%. The additional features seem to provide a relatively modest improvement over just the binary bag of words feature. The hyperparameters agree on both feature sets using the binary weighting.

	Bag of Words Only		All Features	
10^{-5}	0.9133	0.9501	0.9165	0.7477
10^{-4}	0.9133	0.8915	0.9196	0.8323
10^{-3}	0.9133	0.8970	0.9341	0.8770
10^{-2}	0.9232	0.9510	0.9628	0.9508
10^{-1}	0.9832	0.9829	0.9874	0.9852
1	0.9898	**0.9901**	0.9921	**0.9922**
10	0.9898	0.9900	0.9920	0.9921
10^{2}	0.9897	0.9897	0.9919	0.9920
10^{3}	0.9897	0.9897	0.9919	0.9919

Table 4.2: SVM Hyperparameters: TF-IDF Bag of Words

The best result is $C = 1$ and `class_weight` = `Balanced` with accuracies of 99.015% and 99.220%. The TF-IDF weighting appears to generally result in greater accuracy when compared to the BIN bag of words weighting.

4.2.2 Deep Learning Approach (C-LSTM)

Neural networks have been among the most popular classifiers in recent years for general classification problems. Although rarely implemented in spam applications, there is no reason to believe that a neural network will not be an accurate classifier. Specifically, Recurrent Neural Networks (RNNs) have had great success in Natural Language Processing (NLP) applications. In the research, only [39] use a neural network based model, implementing a Gated Recurrent Unit (GRU) instead of a Long Short-Term Memory (LSTM) unit.

This project will utilize a C-LSTM, which uses a Convolutional Neural Network (CNN) on the text data to generate an input to an LSTM, as outlined in [41]. This implementation was chosen as an expansion on the standard LSTM, as it was shown to generally outperform both a CNN and an LSTM on their own.

This model requires different input, as it is attempting to capture information about phrases rather than about the frequency of words as a whole. As such, this paper will use methods provided in `keras.preprocessing` to transform the text into vectors of numbers. Each vector represents an article, and each number in the vector a word in that article. By adding the title to the front of the article, the title will be is included in the analysis. Note that this information is the *only* feature given to the C-LSTM. This is a deep learning model and so it will hopefully learn what is important about these different texts without being provided with additional features.

There are 501,279 unique words in the entire corpus. Of these, 222,169 were used only once and 401,906 were used 10 or fewer times. Only words with a substantial number of occurrences can provide useful information to the C-LSTM. Two different vocabulary sizes (20,000 words and 100,000 words) were used for testing. The vocabulary size refers to the number of words in the training set which are tracked in the network. If the vocabulary size is 100,000, the most 100,000 commonly used words in the training set are used as input. A vocabulary size of 100,000 would include every word used more than 10 times. Increasing vocabulary size causes a corresponding increase in the number of computations and may introduce noise. Here, stop words were not omitted as they may be useful for the deep learning analysis. Any word that does not appear in the vocabulary was replaced with a 0 to indicate that it is "outside of the vocabulary." Only the training set is used to build the vocabulary.

The C-LSTM requires all input to be of the same length. This involves padding the text of articles which are shorter than the longest. However, dealing with over 150,000 articles padded to over 30,000 words causes memory problems for Princeton's Adroit cluster and is generally inefficient as 99% of the corpus has fewer than

4000 words. In order to circumvent this problem, a limit is imposed on the number of words kept per article. The smallest number used here is 4,000, as this captures 99% of the data and the largest is 10,000. Again, this reduces dramatically the amount of computational resources required without excluding entirely outlying articles. In application, one could also imagine considering article length separately as an indicator of satirical articles.

Lastly, the C-LSTM model allowed for class weighting. Both a balanced weighting (where satirical articles are weighted 10 times as much as serious articles) and no weighting were used for testing. This weighting only alters the loss function.

Learning Hyperparameters

Due to the sheer number of hyperparameters for the C-LSTM model and the amount of computational time required to evaluate each, this project was only able to fit some of them from the data. Using two subsets of the training data, one for training and the other for validation, different values for the dropout parameter and L_2 kernel regularization were tested. Here, classes are balanced, step size is set to 512, embedding size is 100, kernel size is 3, maximum length article is 4,000, the sigmoid function for the final dense layer, and both the CNN and LSTM layers have 64 hidden units. The results are displayed below:

Parameters		Training				Validation			
Dropout	L_2 Parameter	Acc	Pre	Rec	F	Acc	Pre	Rec	F
0.1	0	0.9692	0.7991	0.9775	0.8690	0.9449	0.5910	0.6726	0.6292
0.1	0.001	0.9551	0.7388	0.9727	0.8250	0.9438	0.5739	0.8879	0.6971
0.1	0.01	0.3662	0.2613	0.9268	0.4077	0.9230	0.0000	0.0000	0.0000
0.1	0.1	0.8613	0.6471	0.9700	0.7400	0.0770	0.0770	**0.9258**	0.1422
0.2	0	0.8852	0.5740	0.9474	0.6725	0.7536	0.2140	0.7790	0.3358
0.2	0.001	0.9713	0.7869	0.9779	0.8647	0.9590	0.6711	0.7916	0.7264
0.2	0.01	0.9404	0.7376	0.9835	0.8181	0.7978	0.2498	0.7861	0.3791
0.2	0.1	0.3387	0.1482	0.9919	0.2360	0.5328	0.1390	0.9091	0.2411
0.3	0	**0.9931**	**0.9392**	**0.9967**	**0.9639**	**0.9736**	0.7484	0.7887	0.7680
0.3	0.001	0.9870	0.8899	0.9929	0.9347	0.9700	**0.7554**	0.7583	0.7568
0.3	0.01	0.9718	0.7971	0.9579	0.8621	0.9684	0.7346	0.8205	**0.7752**
0.3	0.1	0.8594	0.5100	0.8082	0.6254	0.8397	0.3097	0.8172	0.4492

Table 4.3: C-LSTM Hyperparameters

Note that early stopping was employed here so not all feature combinations went through the same number (maximum of 10) of epochs. Additionally, there is a degree of randomness in initialization of the C-LSTM which appears to have some impact on the results. In general, it appears that increasing the regularization parameter results in a decrease in accuracy and precision but an increase in recall. Based on these results, future classifiers will be trained with no regularization and a dropout of 0.3. This seems to perform the best in general.

Chapter 5

Data Analysis

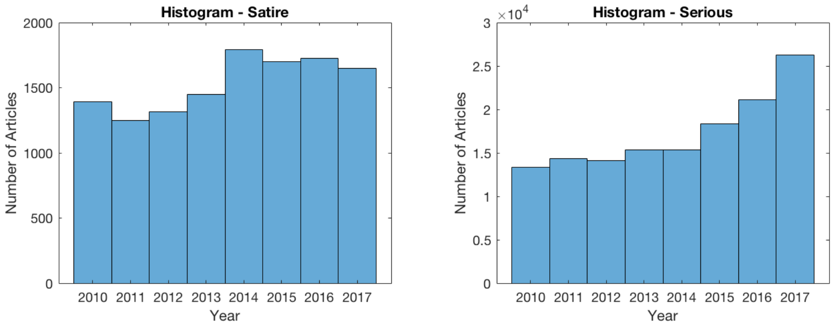
"There is an old cliché
Under your Monet"

Believe Me Natalie
The Killers, 2003

This chapter is dedicated to exploring characteristics of the corpus in more detail. The idea is to understand the characteristics of a few of the major sources in depth and look for patterns that may provide insight about the nature of serious or satirical articles. This process also helped to identify problematic articles which were removed from the dataset. Lastly, data analysis provided an explanation for a number of outliers. The following plots were created using `MATLAB`.

Looking first at the distribution of years for the two groups:

Figure 5.1: Year Histograms



Note that the number of satirical articles remains fairly stable throughout the years, while the number of serious articles gradually increases. There are also satirical articles without dates, which were excluded from this plot.

The next step is to look at the distribution across sources for the two classes.

Figure 5.2: Source Distribution - Satire

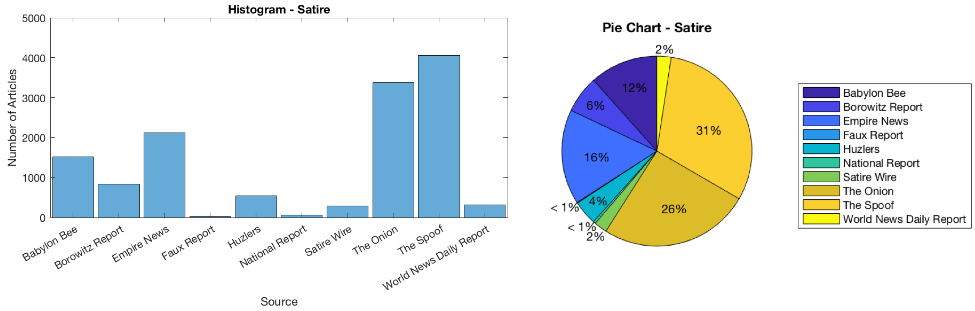
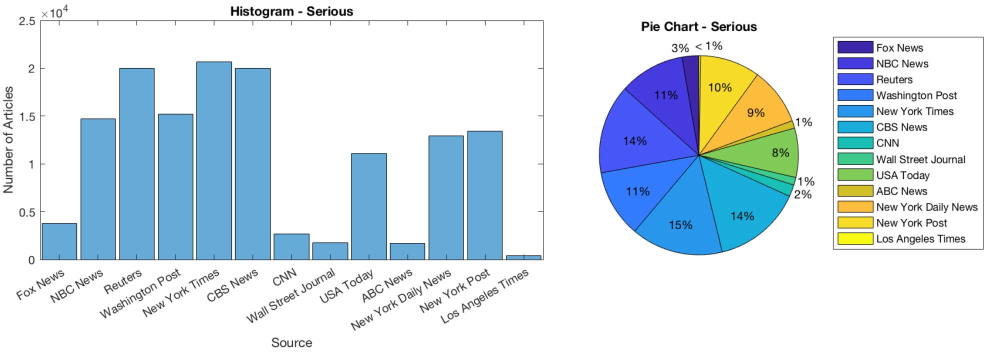


Figure 5.3: Source Distribution - Serious

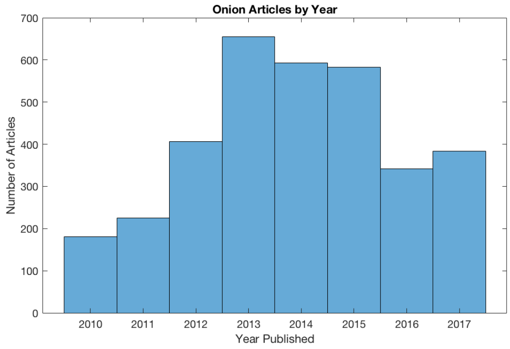


Clearly, satirical sources are dominated by *The Onion* and *The Spoof* which combined contribute 57% of the satirical articles. The smallest five satirical sources combined contribute less than 10% of the data. For serious news articles, the results are more balanced as the two largest sources (*The New York Times* and *CBS News*) make up less than 30% of the corpus. The smallest five sources constitute less than 6% of the data but the rest are closer in proportion to the largest sources. These results suggest *The Onion* and *The New York Times* may be of particular importance for this thesis, and so will be studied in detail below.

5.1 *The Onion*

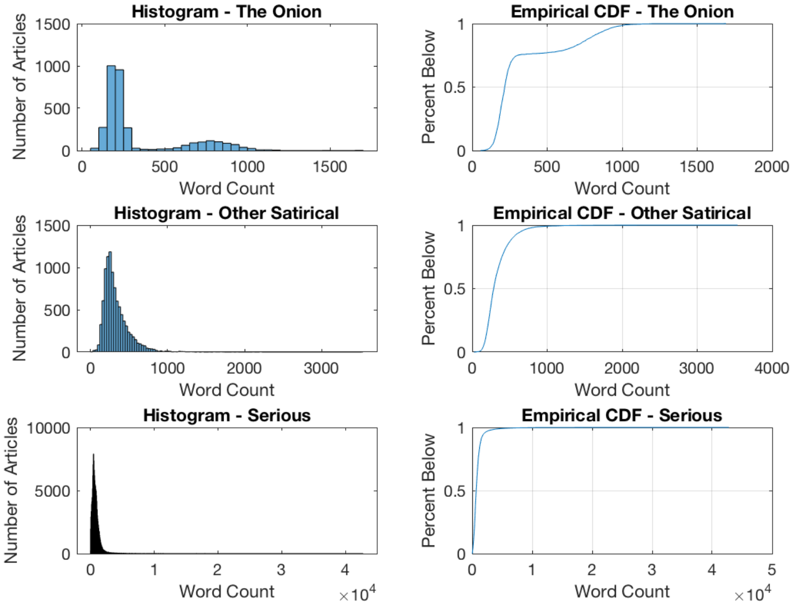
We now consider articles originating from *The Onion*. There are 3,369 such articles. When comparing against other satirical articles, articles from *The Onion* are excluded from the analysis, as the goal is to find ways in which *The Onion* may differ from both other satirical articles and serious news articles as a whole. The outliers and trends in serious articles will be considered in the next section, but serious data is included here for comparison.

Figure 5.4: Onion Year Histogram



Looking at this histogram of years, it is apparent that slightly more articles tend to come from the middle years considered, with the fewest coming from the earliest years. This is likely due to the method of data collection and possibly publishing trends at *The Onion*.

Figure 5.5: Onion Comparison: Word Count



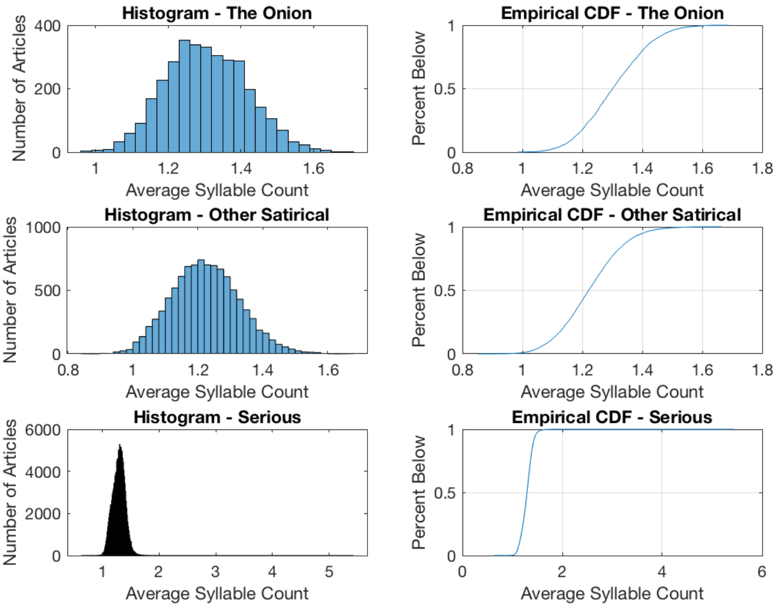
	The Onion	Other Satirical	Serious
Mean	338	331	850
Standard Deviation	260	175	992
Median	217	286	665
Mode	187	249	442
Minimum	55	23	21
5th Percentile	137	152	132
95th Percentile	891	650	1929
Maximum	1692	3539	42786

Table 5.1: Onion Comparison: Word Count

Most of the *Onion* data seems to be around the median with a second cluster of articles around 750 words. There is also a bit of a heavy tail, as a few articles are over 1000 words. The mean is close to that of other satirical articles but well below that of serious articles. The satirical outliers do not appear to be related, just articles with a greater amount of content than usual.

The next feature is the average syllable count, which simply sums the number of syllables in an article and divides by the number of words.

Figure 5.6: Onion Comparison: Average Syllable Count



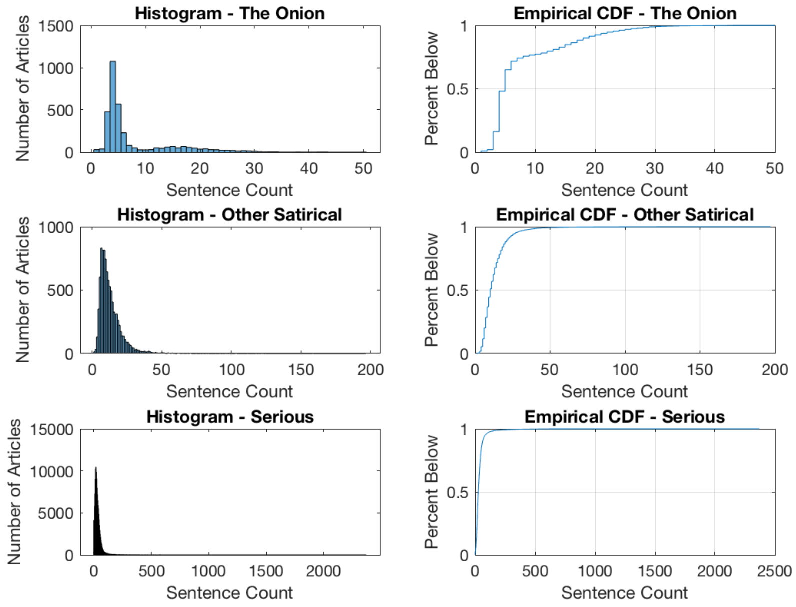
	The Onion	Other Satirical	Serious
Mean	1.31	1.22	1.29
Standard Deviation	0.11	0.10	0.11
Median	1.30	1.22	1.29
Mode	1.4	1.25	1.33
Minimum	0.98	0.85	0.63
5th Percentile	1.13	1.06	1.11
95th Percentile	1.49	1.40	1.46
Maximum	1.69	1.66	5.43

Table 5.2: Onion Comparison: Average Syllable Count

It appears that the average syllable count has a roughly normal shape for satire. Other satirical articles have a lower mean and variance than *Onion* articles. The mean is also slightly higher than serious news articles.

The next feature is sentence count, the number of sentences found in each article.

Figure 5.7: Onion Comparison: Sentence Count



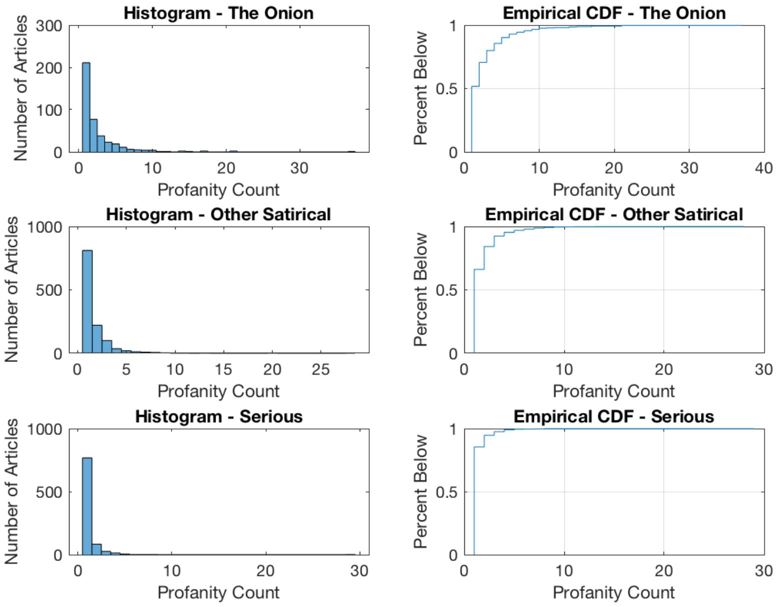
	The Onion	Other Satirical	Serious
Mean	7.73	12.55	35.29
Standard Deviation	6.82	9.06	44.85
Median	5	10	26
Mode	4	6	1
Minimum	1	1	1
5th Percentile	3	4	5
95th Percentile	23	27	83
Maximum	50	197	2370

Table 5.3: Onion Comparison: Sentence Count

Most *Onion* articles are fewer than 10 sentences, with a heavy tail extending to 50 sentences. The average and standard deviation are lower than that of the rest of the satirical corpus. As the number of words used in *The Onion* is relatively consistent, this implies that sentences in *The Onion* tend to be longer than those in other satirical articles. *Onion* articles are still much shorter than serious articles.

The next category is a profanity count, the number of profane words per article. Here, 408 of 3369 or 12.11% of *Onion* articles contained profanity, compared to 1226 of 9758 or 12.57% of all satirical articles. Only 899 of 138,281 or 0.82% of all serious articles contained profanity. The results below are only for articles which include at least one profane word.

Figure 5.8: Onion Comparison: Profanity Count



	The Onion	Other Satirical	Serious
Mean	2.62	1.72	1.27
Standard Deviation	3.27	1.63	1.16
Median	1	1	1
Mode	1	1	1
Minimum	1	1	1
5th Percentile	1	1	1
95th Percentile	8	4	3
Maximum	37	28	29

Table 5.4: Onion Comparison: Profanity Count

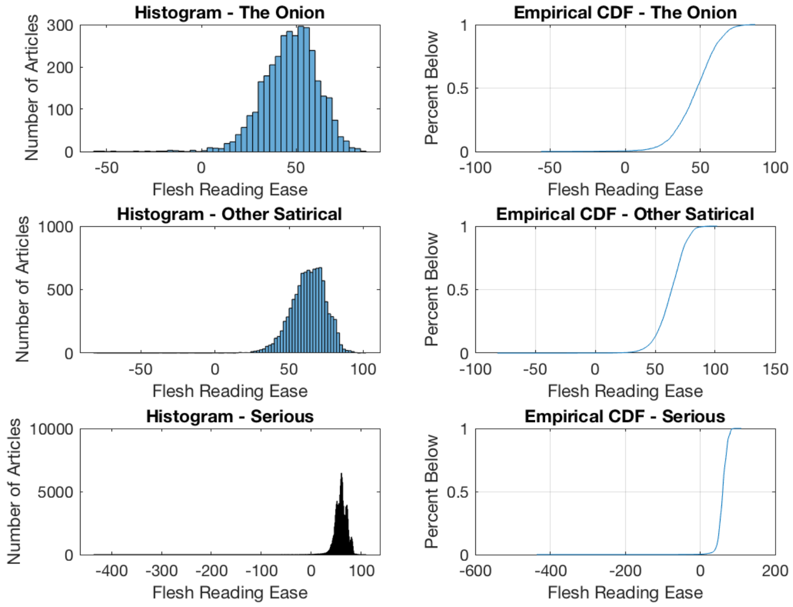
These results indicate that *The Onion*, when using profanity, tends to use more profane words than other satirical sources. Still, most articles with profanity contain only a single instance.

The next two features consider links to other sites and indicators of Twitter characters. These should be very infrequent in satire. Five *Onion* articles contained links to outside sources, all of which come from sponsored articles. Additionally, 30 *Onion* articles contained Twitter characters. There are only two other links in the other satirical data, and 200 other satirical articles with Twitter characters. This indicates that of the satirical data, articles from *The Onion* are far more likely to contain links to outsides sites (all of which are sponsored) and less likely to contain Twitter characters. By comparison, 30,648 serious articles contained Twitter characters (over 22% of the data) and 4,172 had links (about 3%).

This paper considers three reading indices: Flesh Reading Ease (FR), Gunning

Fog Index (GF), and Automated Readability Index (ARI). Note that a greater difficulty of reading according to these metrics may not necessarily imply better grammar, diction, or wit. The indices only point towards more complicated writing according to their formulas outlined in [4].

Figure 5.9: Onion Comparison: Flesh Reading Ease

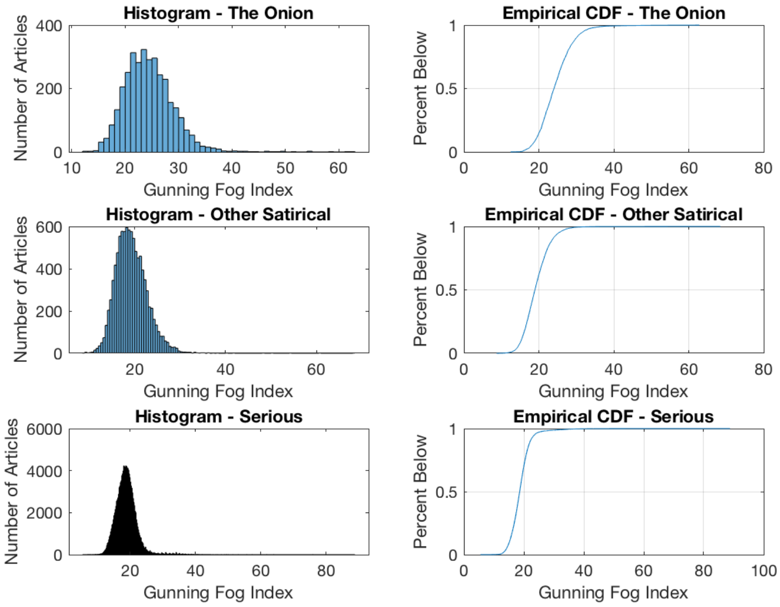


	The Onion	Other Satirical	Serious
Mean	46.99	62.92	59.61
Standard Deviation	14.41	11.79	12.32
Median	47.86	63.73	60.04
Mode	52.53	66.27	60.35
Minimum	-56.09	-81.46	-435.78
5th Percentile	23.29	42.72	41.4
95th Percentile	68.5	80.62	78.48
Maximum	86.6	101.6	109.6

Table 5.5: Onion Comparison: Flesh Reading Ease

Flesh Reading Ease assigns higher scores to easier to read texts. The mean for *The Onion* is lower than that of other satirical articles and serious news articles, suggesting that *The Onion* writes at a higher reading level than its satirical peers and serious news articles as a whole. Those with highly negative scores tend to be articles that contain lists or tables, as those are not easily interpreted by the Flesh Reading Ease algorithm.

Figure 5.10: Onion Comparison: Gunning Fog Index

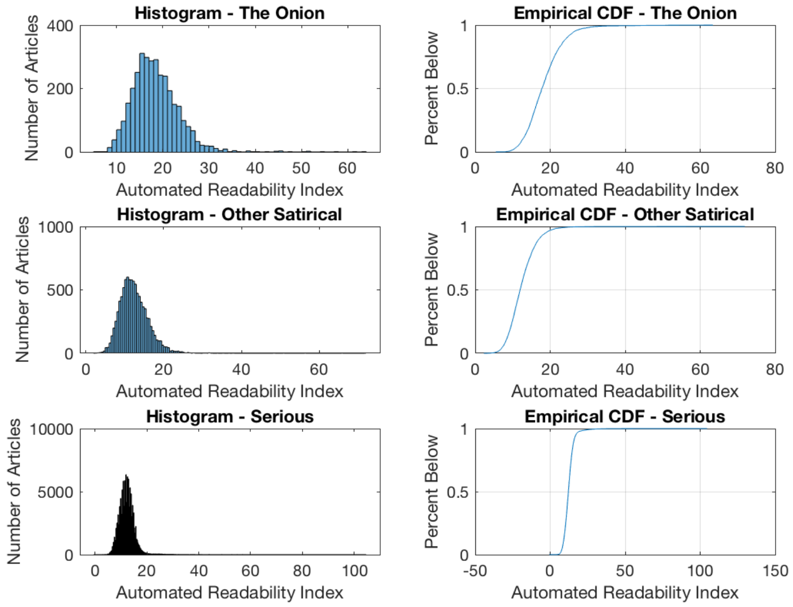


	The Onion	Other Satirical	Serious
Mean	24.45	19.29	18.80
Standard Deviation	4.59	3.51	3.36
Median	24.03	18.96	18.60
Mode	22.68	20	20
Minimum	12.46	8.73	5.56
5th Percentile	18.00	14.39	14.24
95th Percentile	31.94	25.29	23.50
Maximum	62.85	68.44	88.77

Table 5.6: Onion Comparison: Gunning Fog Index

For the Gunning Fog Index, a lower score indicates a more readable text. The mean for *The Onion* was higher than that of other satirical articles and serious news articles, with similar results for the median. This agrees with the Flesh Readability Index in indicating that *The Onion* has a higher reading difficulty than other sources.

Figure 5.11: Onion Comparison: Automated Readability Index



	The Onion	Other Satirical	Serious
Mean	18.37	12.41	12.38
Standard Deviation	5.07	3.75	3.39
Median	17.8	12	12.1
Mode	15.8	10.6	11.8
Minimum	5.5	2.3	-0.6
5th Percentile	11.5	7.3	8
95th Percentile	26.4	18.8	17.1
Maximum	63.4	71.9	104.6

Table 5.7: Onion Comparison: Automated Readability Index

For ARI, higher scores correspond to harder to read texts. The average for *The Onion* was significantly greater than that of all satirical articles and of all serious articles, again with similar results for the medians. These three metrics all agree that *The Onion* is more difficult to read than both satire as a whole and serious news articles. This may indicate that articles in *The Onion* have long clauses or employ longer, more complicated words. This follows from the fact that *The Onion* articles tend to be longer than other serious articles but contain fewer sentences and a higher average syllable count. Note that for each of these reading indices, the results seem to have a roughly normal shape with a few extreme outliers.

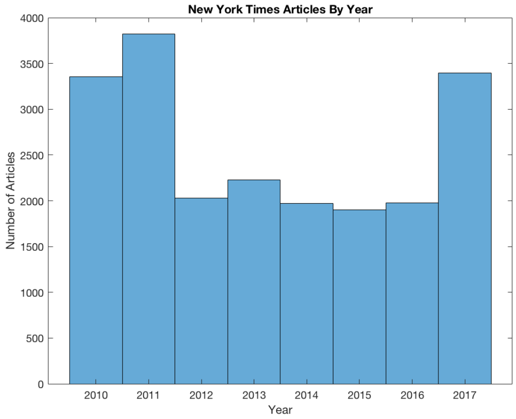
The final group of features all relate to an article's title: title word count, the average number of syllables in the title, and all three reading scores. From these features, we find that the title word count is roughly the same as that of satire in general and that the mean average syllable count is significantly higher than that of satire and of serious news. For the reading indices, the results were consistent with those on the body, in that titles from *The Onion* consistently scored significantly more difficult to read than those of satire as a whole and serious news articles. See Appendix B for a more complete look at these features.

Lastly, we analyze which words appear in the body of articles from *The Onion* more frequently than usual for satirical articles. This is accomplished by testing which of the 500 most frequently used words from the *Onion* data do not appear in the 1000 most frequently used words for the other satirical sources. Words such as "data," "researchers," "adding," "percent," "Thursday," and "nation's" appear more frequently in *The Onion*.

5.2 *The New York Times*

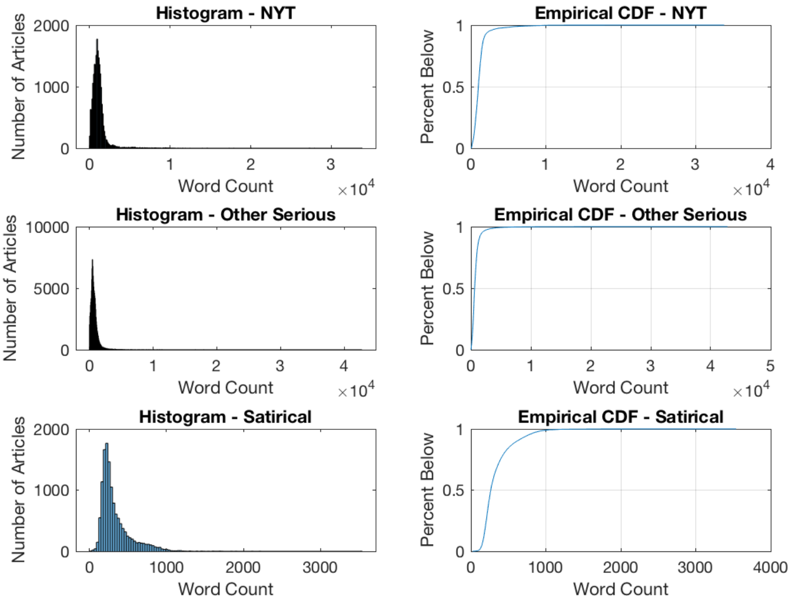
We now compare data from *The New York Times* (abbreviated NYT) to satirical data and data from all other serious news sources. There are 20,672 NYT articles in the corpus.

Figure 5.12: NYT Year Histogram



Most articles from 2010-2011 and 2017. This is simply a product of how the articles were gathered, using *LexisNexis Academic*. Recall that this does not follow the trend of serious news articles as a whole, which tend to somewhat newer.

Figure 5.13: NYT Comparison: Word Count

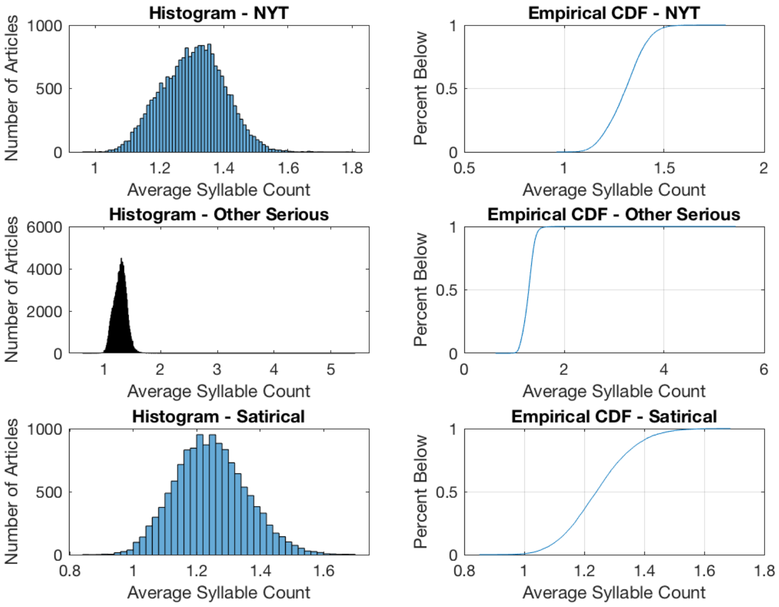


	The New York Times	Other Serious	Satirical
Mean	1167	795	333
Standard Deviation	1134.86	953.97	200.67
Median	992	619	265
Mode	1002	442	237
Minimum	25	21	23
5th Percentile	239	121	146
95th Percentile	2408	1835	765
Maximum	33790	42786	3539

Table 5.8: NYT Comparison: Word Count

Data from *The New York Times* tends to have a significantly higher word count, in both median and mean, as well as a much higher standard deviation. Short NYT articles tend to be updates or corrections, whereas very long articles tend to be full transcripts. Some shorter articles from other sources include summaries or brief market updates.

Figure 5.14: NYT Comparison: Average Syllable Count

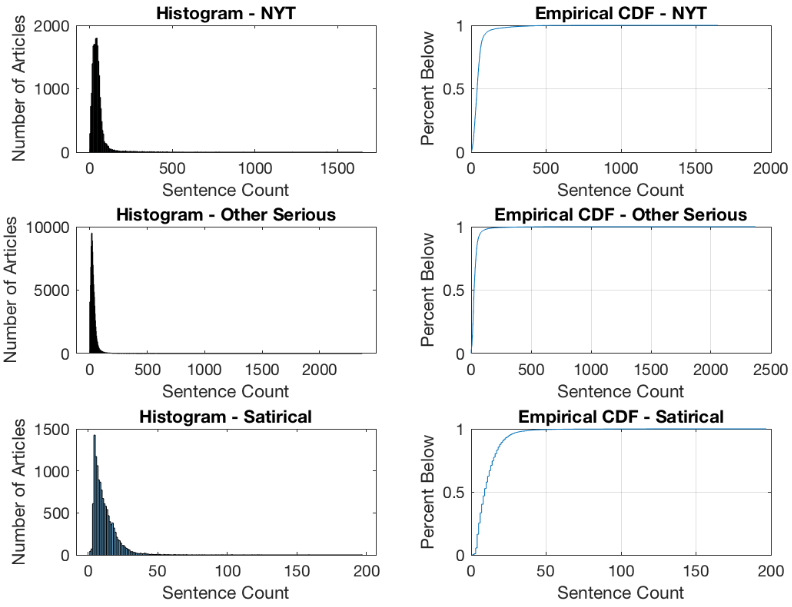


	The New York Times	Other Serious	Satirical
Mean	1.31	1.29	1.25
Standard Deviation	0.10	0.11	0.11
Median	1.31	1.29	1.24
Mode	1.33	1.33	1.25
Minimum	0.96	0.63	0.85
5th Percentile	1.15	1.10	1.07
95th Percentile	1.46	1.46	1.44
Maximum	1.81	5.43	1.69

Table 5.9: NYT Comparison: Average Syllable Count

NYT articles tend to have a significantly higher average syllable count than other serious and satirical sources. It also has a smaller standard deviation, possibly indicating that this difference is somewhat robust. For all sets, other than a few outliers, the average syllable distribution appears to be somewhat normal. Note that those with low syllable counts (less than one) tend to contain tables or many numbers, as do those with unusually high syllable counts. This formatting appears to confuse the syllable counting process. Of course, it is worth noting that serious news articles are more likely to contain tables or large amounts of data compared to satirical articles. Although these outliers make the syllable count behave strangely, they do capture something interesting about the data and so have been included.

Figure 5.15: NYT Comparison: Sentence Count



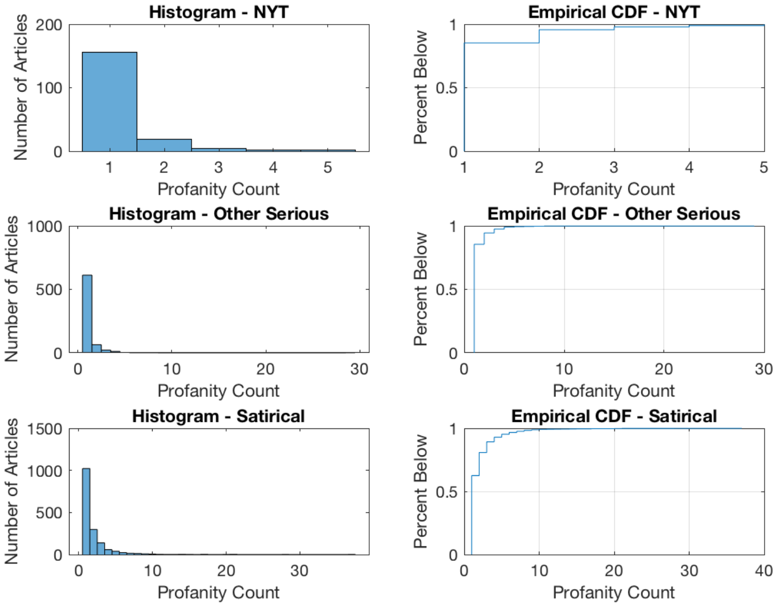
	The New York Times	Other Serious	Satirical
Mean	49.29	32.83	11.32
Standard Deviation	55.54	42.21	8.80
Median	40	25	9
Mode	45	1	4
Minimum	1	1	1
5th Percentile	10	4	3
95th Percentile	109	78	26
Maximum	1646	2370	197

Table 5.10: NYT Comparison: Sentence Count

Clearly *The New York Times* tends to have many more sentences on average than both other serious sources and satirical sources. In fact, less than five percent of articles from *The New York Times* are have fewer sentences than the median satirical article. It is also notable that the NYT contains many more sentences than other serious sources. Number of sentences seems to be a strong indicator for NYT articles. As with word count, articles with many sentences tend to be transcripts of political debates or testimonies. This is the case for all five NYT articles with over 1000 sentences.

Looking next at profanity count, 183 NYT articles with a flagged profanity out of 20,672 or about 0.89%. By comparison, 716 other serious articles out of 117,609 for 0.61%. For satirical articles, 1634 articles with profanity out of 13,127 or about 12.45%. Thus, there are slightly more instances of profanity in NYT articles than in other serious articles but still far fewer than in satirical articles. The results below include only articles with at least one word that is marked as profane. Note that profanity is subjective and may be context-dependent so these results serve as a rough estimate rather than a precise metric. Additionally, no distinction is drawn between words used in quotations and those used as part of the rest of the article.

Figure 5.16: NYT Comparison: Profanity Count



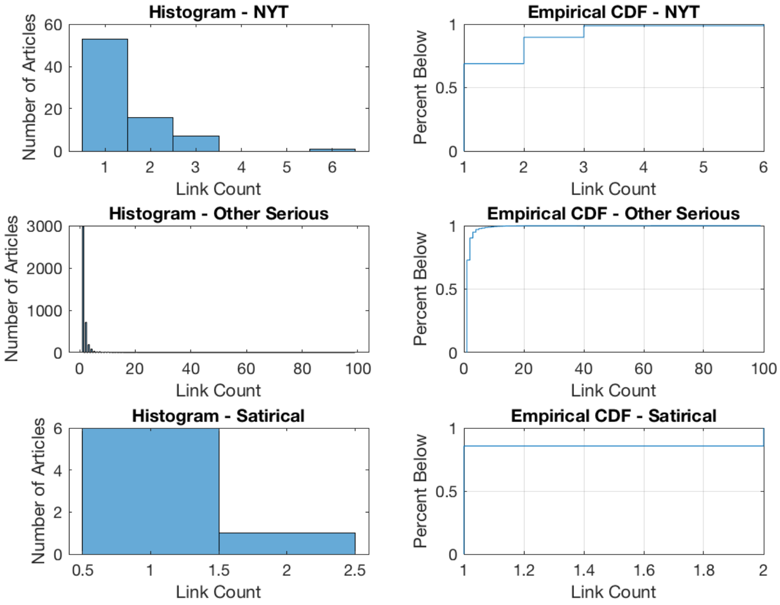
	The New York Times	Other Serious	Satirical
Mean	1.22	1.28	1.94
Standard Deviation	0.65	1.26	2.19
Median	1	1	1
Mode	1	1	1
Minimum	1	1	1
5th Percentile	1	1	1
95th Percentile	2	3	5
Maximum	5	29	37

Table 5.11: NYT Comparison: Profanity Count

Note that when profanity appears in a NYT article, it only does so once or twice. This may be the result of quotations or terminology necessary to convey events. Even so, this is slightly lower than the average for other serious sources and substantially lower than the average for satire. The majority of NYT articles with profanity seem to contain only a single instance, which again may be from a quotation. The other serious articles with high profanity counts tend to focus on Trump's comments regarding women. It is also worth noting that no weight was assigned to different profanities. Some may be considered more acceptable than others and may have more of a place in serious reporting. Again, this distinction would be culturally subjective and is unlikely to significantly impact the effectiveness of this feature.

Examining the number of links to other sites and the number of indicators of Twitter characters, 77 NYT articles contained links (about 0.37%). This is much less than the 4095 links found in other serious articles, accounting for 3.48% of other serious articles. The plots below represent only articles with at least one outside link.

Figure 5.17: NYT Comparison: Link Count



	The New York Times	Other Serious	Satirical
Mean	1.45	1.60	1.14
Standard Deviation	0.84	2.51	0.38
Median	1	1	1
Mode	1	1	1
Minimum	1	1	1
5th Percentile	1	1	1
95th Percentile	3	4	2
Maximum	6	99	2

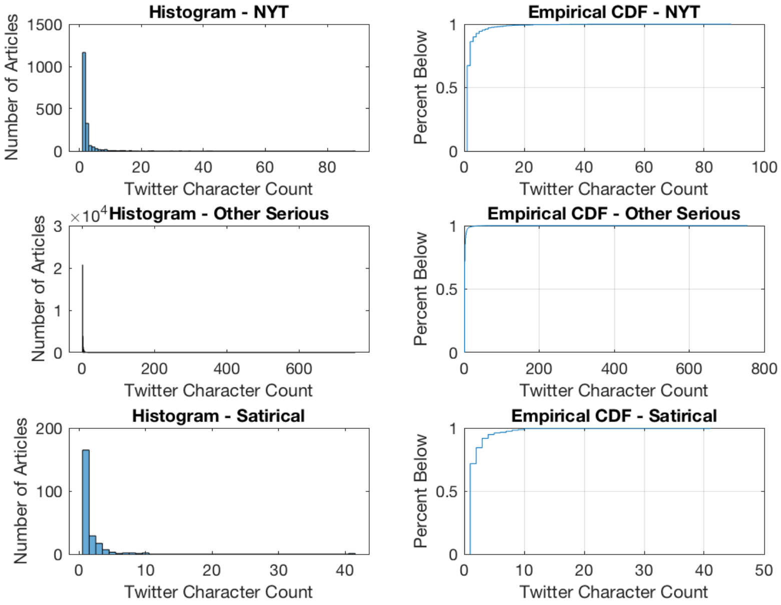
Table 5.12: NYT Comparison: Link Count

It is clear that the majority of NYT articles contain only a single link, as opposed to several other sources which frequently have multiple. Those articles with a high number of links tend to include hyperlinks to key terms, information about upcoming events, or product recommendations.

For indicators of Twitter links, 1731 NYT articles have Twitter characters, accounting for 8.4% of articles. Of the other serious sources, 28,917 or 24.6% contain

Twitter characters. For satirical articles, only 230 contain a Twitter link, accounting for 1.75% of the data. Again, the information below only shows articles with at least one Twitter link.

Figure 5.18: NYT Comparison: Twitter Character Count

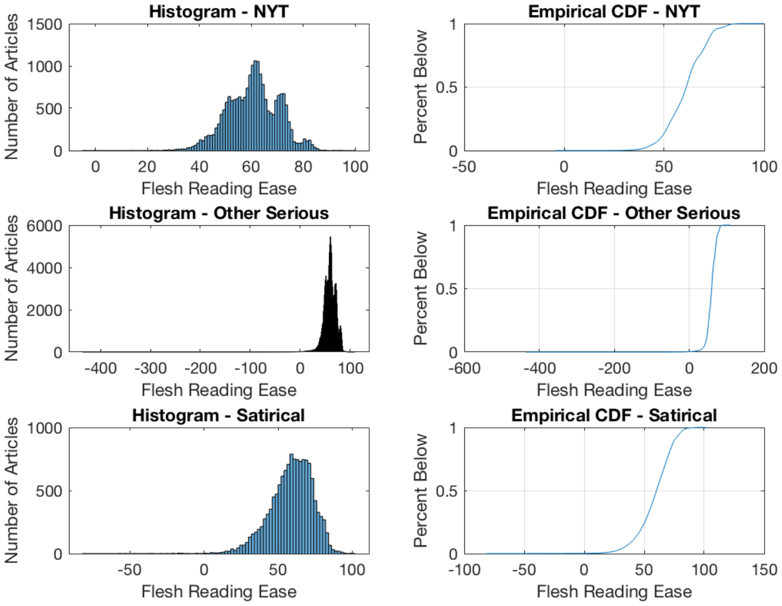


	The New York Times	Other Serious	Satirical
Mean	2.07	2.03	1.84
Standard Deviation	3.72	5.84	3.00
Median	1	1	1
Mode	1	1	1
Minimum	1	1	1
5th Percentile	1	1	1
95th Percentile	6	6	5
Maximum	89	756	41

Table 5.13: NYT Comparison: Twitter Character Count

Many of the articles with very high counts are in fact about Twitter or focus on important tweets. Some consider President Trump's Twitter usage or the impact of Twitter movements such as #MeToo. While fewer NYT articles contain Twitter symbols, those that do tend to have about the same amount as other serious sources.

Figure 5.19: NYT Comparison: Flesh Reading Ease

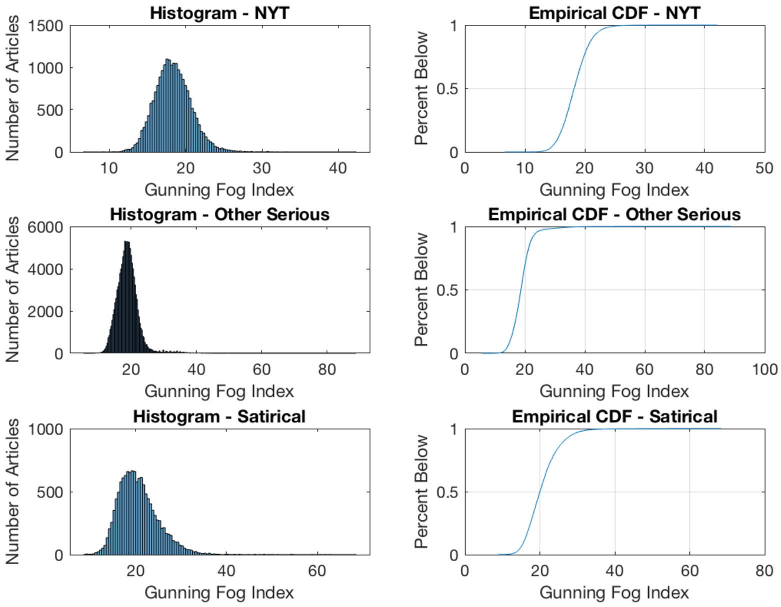


	The New York Times	Other Serious	Satirical
Mean	60.69	59.42	58.83
Standard Deviation	9.75	12.71	14.32
Median	61.06	59.90	60.24
Mode	60.45	60.04	66.27
Minimum	-4.49	-435.78	-81.46
5th Percentile	44.34	40.72	33.17
95th Percentile	75.40	78.79	79.50
Maximum	99.94	109.55	101.60

Table 5.14: NYT Comparison: Flesh Reading Ease

The Flesh Reading Ease indicates that *The New York Times* tends to be marginally easier to read than other serious sources and satirical sources in general. It also seems to suggest that *The New York Times* has a more consistent reading level than the other sources, based on both the lower standard deviation and smaller range. The articles with a highly negative Flesh Reading Ease are generally those that contain tables or other data.

Figure 5.20: NYT Comparison: Gunning Fog Index

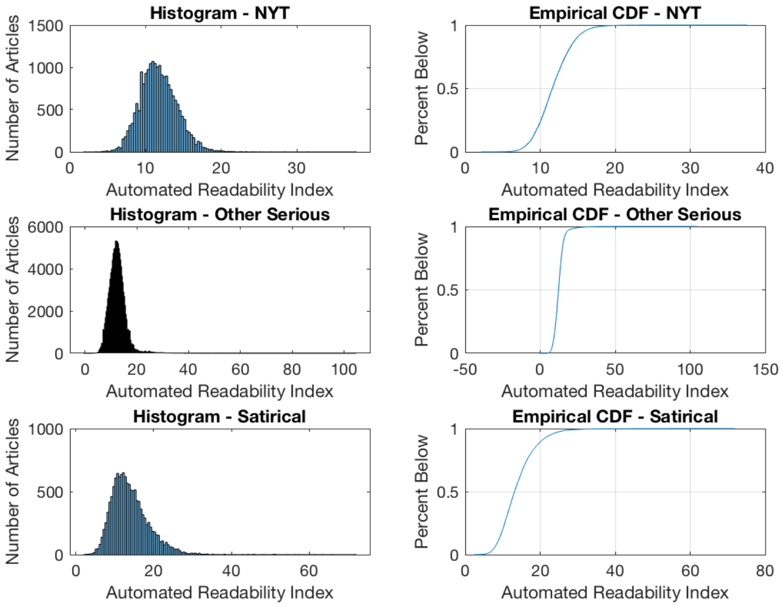


	The New York Times	Other Serious	Satirical
Mean	18.34	18.88	20.62
Standard Deviation	2.37	3.50	4.43
Median	18.22	18.68	20.00
Mode	18	20	20
Minimum	6.63	5.56	8.73
5th Percentile	14.68	14.16	14.71
95th Percentile	22.32	23.73	28.56
Maximum	42.13	88.77	68.44

Table 5.15: NYT Comparison: Gunning Fog Index

These results agree with the Flesh Reading Ease, indicating that NYT articles are generally easier to read and somewhat more consistent than other serious and satirical sources.

Figure 5.21: NYT Comparison: Automated Readability Index



	The New York Times	Other Serious	Satirical
Mean	11.77	12.48	13.94
Standard Deviation	2.42	3.52	4.88
Median	11.6	12.2	13.1
Mode	10.8	11.8	10.6
Minimum	2.1	-0.6	2.3
5th Percentile	8.1	7.9	7.7
95th Percentile	15.9	17.3	22.8
Maximum	37.6	104.6	71.9

Table 5.16: NYT Comparison: Automated Readability Index

The ARI data agrees with the other two readability indices, in that *The New York Times* is slightly easier to read than other serious sources and significantly easier than satirical sources. Again, NYT data also appears to be the most consistent, with the smallest variance and range. Sources with outlying scores tend to have many numbers or a formatted table, both of which can make an article difficult to read in terms of sentences.

Overall for Gunning Fog and Flesh Reading Ease, NYT articles do not seem to be significantly different from other serious sources, especially considering the number of outliers in other serious articles. For the Automated Readability Index, NYT articles appear to be more separated from the other serious articles. This is because although NYT articles tend to have a similar average syllable count to other serious news, NYT articles tend to be longer in terms of word and sentence count. ARI does not consider syllables, so the difference in word and sentence count is reflected more strongly.

The fact that NYT articles tend to be rated as easier to read than satirical articles was unexpected. This may be a feature of how the readability indices work and the nature of satirical articles. Satirical articles frequently have to communicate confusing or absurd ideas succinctly, which could decrease readability. Satirical articles may also have an incentive to be short to hold a reader's interest, which could result in sentences with many buzzwords or complex logic. Serious news articles are able to be more direct in communication, leading to more readable sentences. These results for *The New York Times* likely reflect the higher sentence count found in NYT articles, which generally results in each sentence being easier to read.

The final set of features relate to the title. For complete results, see Appendix B. In general, titles from *The New York Times* to be somewhat shorter than those from other sources, with a slightly higher average syllable count. Titles tended to be somewhat easier to read than other serious sources.

Finally, we consider words which appear in the body of articles from *The New York Times* more frequently than in articles from other serious news sources using the same technique as described in the section on *The Onion*. Of the 500 most commonly used words in *The New York Times*, the words "senator," "book," "seemed," "politics," "lawyer," and "appeared" were not in the 1000 most commonly used for other serious news sources.

5.3 General Observations

As observed previously, the greatest difference between serious and satirical articles seems to be word count. Serious articles tend to be substantially longer than satirical articles, although there is overlap. As a result of this, serious articles tend to have more sentences than satirical articles. Average syllable count seemed to be a slight indicator, in that higher average syllable counts tended to correspond to serious articles. The difference is relatively small, and in fact *The Onion* and *The New York Times* have the same average syllable account on average. As expected, high link count and high Twitter character count are strong indicators of serious articles.

Articles from *The Onion* tend to have a higher word count, higher average syllable count, and lower sentence count than their satirical peers. As a result of this, *Onion* articles are consistently scored as more difficult to read than other satirical sources. *New York Times* articles are generally longer in terms of words and sentences than other serious sources, with a slightly higher average syllable count. NYT articles also tend to have fewer outside links and fewer Twitter characters than other serious sources. Lastly, NYT articles are usually scored as easier to read than other serious sources due to having substantially more sentences.

Chapter 6

Results

> "Misinformation is not like a
> plumbing problem you fix.
> It's a social condition, like crime,
> that you must continually
> monitor and adjust to."
>
> Tom Rosenstiel, 2017

Included below are the results of training and testing each of the classifiers. Each section will include information about the data used. Each result comes from using the hyperparameters learned above. For the SVM, this is $C = 0.01$ and `class_weight` = `None` for a binary word weighting and $C = 1$ and `class_weight` = `Balanced` for TF-IDF weighting.

Here, BIN will represent a binary weighting, TF the TF-IDF weighting, T features that come from the title, PR the profanity feature, TW the twitter feature, L the link feature, RI the readability indices, TC the text complexity features, and DO the dates with one-hot encoding. ALL will represent all non-bag of words features. For confusion results, TP refers to "true positive," FP to "false positive," FN to "false negative," and TN to "true negative." These confusion numbers provide a more complete look at the performance of the classifier.

6.1 Yang et al Corpus

Here, a benchmark for the SVM is established using the data provided in Yang et al [39]. This data lacks a number of features employed by this paper. Namely, the Yang dataset does not include the date or title. It also lacks URLs and title indicators. We will test only the binary and TF-IDF classifiers with all available features, specifically word count, profanity count, average syllable count, sentence count, and the reading indices. All available training and testing sets provided are used at once, for 12,632 satirical articles and 133,043 serious articles in the training set with 3,601 satirical and 33,619 serious articles in the test set.

Using the same method as for this paper's corpus, it was found that the optimal hyperparameters for the Yang data agree with those found for this corpus. Those hyperparameters produce just over 98% accuracy on the training set.

The results are below:

Features	Acc	Pre	Rec	F
BIN + ALL	0.9675	**0.9349**	0.7140	0.8096
TF + ALL	**0.9714**	0.9101	**0.7814**	**0.8409**

Table 6.1: SVM Results: Yang et al Corpus

	BIN + ALL		TF + ALL	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	6.91%	2.77%	7.56%	2.11%
True Ser	0.48%	89.84%	0.75%	89.58%

Table 6.2: Confusion Matrix: Yang et al Corpus

The best result here comes from using TF-IDF weighting, which achieves over 97% accuracy. This is relatively comparable with the SVM models used in [39], which range in accruacy from 97% to just over 98% on the test data. Precision is also comparable, while recall (and so F-score) for this implementation is much lower than that found in [39]. It is interesting to note that in this corpus, fewer sources are used in serious news (only six, two of which are British). Additionally, satirical sources are split between sets, meaning that satirical sources used in the training data set do not appear in the testing data set and vice-versa. These differences, along with a smaller number of available features, suggest that this classifier will perform worse on the dataset from Yang et al than it will on the corpus gathered for this project.

If we combine the training and testing sets and draw from each 85% of the data for training and 15% for testing, we get the following results:

Features	Acc	Pre	Rec	F
BIN + ALL	0.9806	0.9193	0.8567	0.8869
TF + ALL	**0.9834**	**0.8897**	**0.9277**	**0.9083**

Table 6.3: SVM Results: Yang et al Corpus - Randomized

We see signifincant gains in accuracy and recall, at the cost of precision. These results are closer to those found in [39] for an SVM.

6.2 Entire Corpus: Serious vs Satire

6.2.1 SVM

The following are results from using all serious and satirical articles with satire as the positive class. Satire includes both those identified as "fake news" with a satire disclaimer and sources only classified as satire. 85% of the data is used for training and 15% for testing.

The following results come from using the training and test data in `train.csv` and `test.csv` respectively.

Features	Acc	Pre	Rec	F
BIN	0.9889	0.9493	0.9213	0.9351
BIN + T	0.9900	0.9585	0.9254	0.9416
BIN + T + PR	0.9902	0.9605	0.9249	0.9423
BIN + T + PR + TW	0.9903	0.9596	0.9274	0.9432
BIN + T + PR + TW + L	0.9904	0.9596	0.9289	0.9440
BIN + T + PR + TW + RI	0.9904	0.9567	0.9310	0.9437
BIN + T + PR + TW + RI + TC	0.9908	0.9612	0.9310	0.9458
BIN + T + PR + TW + RI + TC + DO	0.9914	**0.9664**	0.9335	0.9497
BIN + ALL	**0.9915**	0.9654	**0.9355**	**0.9502**

Table 6.4: BIN, Fixed Set of Entire Corpus - Metrics

There is a general increase in accuracy as additional features are added. Using all features provided the best results in terms of accuracy and recall, achieving 99.15% accuracy, 96.54% precision, 93.55% recall, and 95.02% F measure. The confusion numbers are below, reported as percentages of the test set.

Features	% TP	% FP	% FN	% TN
BIN	7.99	0.43	0.68	90.90
BIN + T	8.03	0.35	0.65	90.98
BIN + T + PR	8.02	0.33	0.65	91.00
BIN + T + PR + TW	8.04	0.34	0.63	90.99
BIN + T + PR + TW + L	8.06	0.34	0.62	90.99
BIN + T + PR + TW + RI	8.07	0.37	0.60	90.96
BIN + T + PR + TW + RI + TC	80.7	0.33	0.60	91.00
BIN + T + PR + TW + RI + TC + DO	8.10	**0.28**	0.58	91.04
BIN + ALL	**8.11**	0.29	**0.56**	**91.04**

Table 6.5: BIN, Fixed Set of Entire Corpus - Confusion

As expected, the number of correct predictions increases with more features and the number of false predictions decreases. As with the metrics, it appears that adding the DO feature, corresponding to dates with one-hot encoding, boosted performance the most.

Features	Acc	Pre	Rec	F
TF	0.9903	0.9357	0.9533	0.9444
TF + T	0.9914	0.9449	0.9569	0.9508
TF + T + PR	0.9915	0.9462	0.9558	0.9510
TF + T + PR + TW	0.9913	0.9439	0.9569	0.9503
TF + T + PR + L	0.9914	0.9448	0.9563	0.9506
TF + T + PR + RI	0.9919	0.9430	0.9655	0.9541
TF + T + PR + TC	**0.9924**	0.9473	**0.9665**	**0.9568**
TF + T + PR + DO	0.9919	**0.9537**	0.9523	0.9530
TF + ALL	**0.9924**	0.9509	0.9624	0.9566

Table 6.6: TF, Fixed Set of Entire Corpus - Metrics

These results are a bit more complex. While generally including more features boosted performance, the best result overall came from only the bag of words, profanity, title information, and information on text complexity. It should be noted that this feature set performed only moderately better than all features and that including all features performed better on precision. It is also worth noting that using the TF-IDF weighting resulted in somewhat better accuracy and much better recall for all feature sets when compared with binary weighting, at the cost of some precision.

Features	% TP	% FP	% FN	% TN
TF	8.27	0.57	0.41	90.76
TF + T	8.30	0.48	0.37	90.84
TF + T + PR	8.29	0.47	0.38	90.86
TF + T + PR + TW	8.30	0.49	0.37	90.83
TF + T + PR + L	8.29	0.48	0.38	90.84
TF + T + PR + RI	8.37	0.51	0.30	90.82
TF + T + PR + TC	**8.38**	0.47	**0.29**	90.86
TF + T + PR + DO	8.26	**0.40**	0.41	**90.93**
TF + ALL	8.35	0.43	0.33	90.90

Table 6.7: TF, Fixed Set of Entire Corpus - Confusion

These results seem to be in line with the metrics discussed above. It is interesting to note that the number of false positives tends to be greater than for feature sets with the binary bag of words, but the false negatives tend to be much fewer. The number of true negatives also tends to decline slightly, balanced out by a significant increase in the number of true positives. This is likely a result of the `balanced` parameter as this causes the loss function to penalize mis-classified satirical articles more heavily. We expect and see that this results in fewer false negatives on the test set.

The above results come from a single shuffle of the corpus, from selecting randomly one training set and one test set. Below are the average results from testing the same feature sets on ten random training and testing sets. This should provide a better picture of the general performance of the classifier.

Features	Acc	Pre	Rec	F
BIN	0.9894	0.9545	0.9216	0.9378
BIN + T	0.9901	0.9566	0.9286	0.9423
BIN + T + PR	0.9903	0.9595	0.9277	0.9433
BIN + T + PR + TW	0.9905	0.9580	0.9313	0.9444
BIN + T + PR + TW + L	0.9902	0.9555	0.9301	0.9426
BIN + T + PR + TW + RI	0.9907	0.9576	0.9346	0.9460
BIN + T + PR + TW + RI + TC	0.9909	0.9610	0.9330	0.9468
BIN + T + PR + TW + RI + TC + DO	0.9917	0.9647	0.9385	0.9514
BIN + ALL	**0.9922**	**0.9675**	**0.9421**	**0.9546**

Table 6.8: SVM Average Results: BIN Weighting

Features	Acc	Pre	Rec	F
TF	0.9909	0.9426	0.9531	0.9478
TF + T	0.9913	0.9468	0.9534	0.9501
TF + T + PR	0.9920	0.9488	0.9597	0.9542
TF + T + PR + TW	0.9916	0.9449	0.9595	0.9521
TF + T + PR + L	0.9912	0.9439	0.9556	0.9497
TF + T + PR + RI	0.9917	0.9484	0.9564	0.9524
TF + T + PR + TC	0.9918	0.9478	0.9587	0.9532
TF + T + PR + DO	0.9925	0.9543	0.9590	0.9566
TF + ALL	**0.9929**	**0.9553**	**0.9629**	**0.9591**

Table 6.9: SVM Average Results: TF Weighting

The best scores for all four metrics with both weightings come from utilizing all features. Note that adding features during some steps appears to have decreased accuracy, but as a whole all of the features together maximize accuracy. It is also worth noting that TF-IDF provides significantly better accuracy, recall, and F score, but worse precision than the binary weighting. This is not surprising, as TF-IDF tended to produce more false positives than binary weighting. This is consistent with the fixed training and testing set and the `balanced` hyperparameter as discussed earlier.

6.2.2 C-LSTM

The C-LSTM was also applied to comparing serious and satirical articles for the entire corpus. This set was chosen as the primary application of the C-LSTM because deep learning methods tend to fare best with large data sets. Although the C-LSTM does not require different feature sets, a number of parameters were tweaked. The results from training with `train.csv` and testing with `test.csv` are included below. Here, Vocab Size is the size of the vocabulary considered, Embedding is the size of the word embedding, and Balanced refers to whether or not the classes were balanced. The maximum article length was fixed to 4,000 words, meaning that all shorter articles were padded with "out of index" symbols and all longer articles were truncated. This length retains about 99% of the articles in their entirety while avoiding the computational overhead of padding each article to the length of the longest (over 42,000 words).

Vocab Size	Embedding	Balanced	Acc	Pre	Rec	F
20,000	100	N	**0.9882**	0.8764	0.8992	**0.8877**
20,000	100	Y	0.9636	0.7123	0.9056	0.7974
20,000	300	N	0.9871	**0.8841**	0.8606	0.8722
20,000	300	Y	0.9761	0.7835	**0.9169**	0.8450
100,000	100	N	0.9850	0.8624	0.8876	0.8748
100,000	100	Y	0.9777	0.8076	0.8684	0.8369
100,000	300	N	0.9133	0.0000	0.0000	0.0000
100,000	300	Y	0.9689	0.7454	0.8972	0.8143

Table 6.10: C-LSTM Results

Increasing the maximum article length to 10,000 words produced the following:

Vocab Size	Embedding	Balanced	Acc	Pre	Rec	F
20,000	100	N	0.9754	0.8180	0.8088	0.8134
20,000	300	N	**0.9874**	**0.8726**	**0.8939**	**0.8831**

Table 6.11: C-LSTM Results, Maximum Length 10,000 words

This increase in maximum length more than doubles the computation time of each epoch, without performance gains. In general, it appears that the vocabulary size of 20,000 with embedding size of 100, maximum article length of 4,000, and unbalanced classes performed the best. Unbalanced classes tended to boost accuracy and precision, except in one case where no articles were labeled satirical. Class balancing generally improved recall. It seems that neither larger vocabulary size nor greater embedding was a major factor in performance. Overall, the C-LSTM performed worse than the SVM and was much more computationally intensive. The C-LSTM results were roughly on par or better than other attempts at satire detection [5][39]. The biggest benefit of the C-LSTM seems to be its deep learning nature, in that preprocessing was much simpler and more generalizable. Further adjustment of hyperparameters and repeated experiments to overcome randomness might improve C-LSTM performance.

6.3 Serious vs Fake News

For this category, we will only consider the two best feature groups for each bag of words weighting. We will also select a subset of the serious news data approximately equal to ten times the size of the "fake news" dataset. This subset will be chosen randomly in each iteration from the entire serious news corpus. The intent of this is to make the results more comparable to those from the previous part and to increase the relative frequency of "fake news" articles.

Features	Acc	Pre	Rec	F
BIN + T + PR + RI + TC + DO	0.9947	**0.9831**	0.9585	0.9706
BIN + ALL	0.9944	0.9805	0.9580	0.9691
TF + T + PR + DO	0.9948	0.9780	0.9644	0.9711
TF + ALL	**0.9948**	0.9704	**0.9725**	**0.9714**

Table 6.12: SVM Results: Serious vs Fake

Here we found that TF + ALL was the most effective across three metrics, with both binary feature sets outperforming it in precision. All metrics saw an improvement over the SVM test comparing serious news and satire. While accuracy only increased by about 0.20%, the other metrics each increased by over 1%. This suggests that "fake news" may be easier to identify than satire in general.

6.4 Satire vs Fake News

Here we compare the subset of satirical news articles labeled "fake news" with those which have only been identified as satire. Note that those articles which have been labeled "fake news" also carry a satire disclaimer. As such, this process is likely to be noisier than the previous comparisons.

Features	Acc	Pre	Rec	F
BIN + T + PR + RI + TC + DO	0.9709	0.9500	0.9224	0.9360
BIN + ALL	0.9714	**0.9526**	0.9222	0.9371
TF + T + PR + RI + TC + DO	**0.9754**	0.9382	**0.9567**	**0.9473**
TF + ALL	0.9747	0.9379	0.9538	0.9458

Table 6.13: SVM Results: Satire vs Fake

The classifier performed significantly worse on this dataset, as expected. The TF-IDF weighting without the Twitter or link features performed the best overall, with recall and F-score slightly below those found previously, and accuracy and precision significantly worse. This is likely due to a few reasons. Firstly, as has been established, satire and "fake news" share many similarities which can make distinguishing them difficult [15]. Second, this dataset is smaller than the previous two. With only about 13,000 articles total, there may simply not have been enough data. Overall, 97% accuracy (especially when "fake news" is about a quarter of the data) is good for most applications and exceeds some other papers on all metrics [5]. With more data and more of a focus on the differences between "fake news" and satire, these results could likely be improved.

6.5 2010 to 2013 vs 2014 to 2017

The following analysis will attempt to capture the changing relationship between news and satire by studying the impact of the date of publication on the accuracy of the classifier. First, all data sources without years (*World News Daily Report* and *Huzlers*) will be dropped. Next, the data will be partitioned with one set containing both serious and satirical articles dated 2010-2013 and the other containing articles dated 2014-2017. Some sources, particularly satirical ones, do not span the entire 2010-2017 window and most of these skew towards 2013 and later. This effect will not be accounted for, in that articles from all sources with dates will be included. The arrival of new sources could indicate something about a change in style or content that this part of the project is trying to capture. Each set will be trained and tested on itself as a benchmark for performance for that time period. Then, the classifier will be trained on one set and tested on the other. For these tests, the classifier will use the entirety of one set for training and the entirety of the other for testing.

We will use two possible feature sets. We will not include the date feature because when comparing across time periods none of the dates in the training set will appear in the testing set. This could cause our benchmarks to return a higher accuracy by making use of this feature that will not be relevant when comparing across sets. Additionally, because TF-IDF weighting is performing better in general (except in precision), we will limit the testing to only feature sets with that weighting. Specifically, we will use TF + T + PR + RI + TC and TF + ALL - DO (hereafter TF_1 and TF_2) as they seem to have performed the best in general of feature sets without the date feature.

6.5.1 Train on 2010-2013, Test on 2014-2017

Here, the benchmark will refer to training on a subset of the 2010-2013 data and validating on another subset of the 2010-2013 data. For testing, the SVM will be trained on all of the 2010-2013 data and tested on all of the 2014-2017 data.

Features	2010-2013 Benchmark				Test on 2014-2017			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
TF_1	0.9920	0.9565	0.9506	0.9535	0.9876	**0.9223**	0.9185	0.9204
TF_2	**0.9922**	**0.9565**	**0.9531**	**0.9548**	**0.9876**	0.9172	**0.9238**	**0.9205**

Table 6.14: SVM Results: Train 2010-2013, Test 2014-2017

The benchmark seems relatively consistent with the corpus as a whole. A slight decline may be attributed to the lacking date feature. When testing on the 2014-2017 data, accuracy dropped by about half a percentage point with precision and recall each decreasing by about 3%.

	TF_1		TF_2	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	7.17%	0.64%	7.21%	0.59%
True Ser	0.60%	91.60%	0.65%	91.55%

Table 6.15: Confusion Matrix: Train 2010-2013, Test 2014-2017

False positives and false negatives seem to be roughly balanced for this classifier. Overall, there appears to be a modest decline in classifier performance when split by years in this way.

6.5.2 Train on 2014-2017, Test on 2010-2013

Features	2014-2017 Benchmark				Test on 2010-2013			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
TF_1	**0.9942**	**0.9665**	**0.9592**	**0.9628**	0.9838	0.9012	**0.9125**	0.9068
TF_2	0.9937	0.9615	0.9580	0.9597	**0.9842**	**0.9063**	0.9107	**0.9085**

Table 6.16: SVM Results: Train 2014-2017, Test 2010-2013

For the 2014-2017 benchmark, accuracy and precision appear to have exceeded the average for the entire corpus, while recall declined slightly. This suggests a slight increase in the number of articles predicted to be satirical. When tested on data from 2010-2013, accuracy fell by about 1%, while precision and recall both fell by around 5%.

	TF_1		TF_2	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	7.88%	0.76%	7.87%	0.77%
True Ser	0.86%	90.50%	0.81%	90.55%

Table 6.17: Confusion Matrix: Train 2014-2017, Test 2010-2013

These results suggest that there has been change over time in the way satirical and serious articles are written. Accuracy improves significantly when training and testing are limited to the 2014-2017 dataset, possibly indicating that satirical and serious articles are in general more different in this time period. There does not seem to be a clear cause for the change other than different data, as false positives and false negatives appear to be relatively even. Granted, the decline experienced across years does not seem insurmountable. Even at its worst, the SVM still achieved over 98% accuracy and reasonably high precision and recall. Further tweaks and regular updates of the classifier could yield significant performance gains in practice.

6.6 2014 and 2015 vs 2016 and 2017

For an even closer look at change over time, the data is divided in to two groups, one with all articles from 2014-2015 and the other with all articles from 2016-2017. This division will hopefully provide further insight about recent developments in the relationship between serious news and satire.

6.6.1 Train on 2014 and 2015, Test on 2016 and 2017

Features	2014-2015 Benchmark				Test on 2016-2017			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
TF_1	**0.9922**	**0.9547**	**0.9578**	**0.9562**	0.9854	0.8516	**0.9455**	0.8961
TF_2	0.9919	0.9537	0.9560	0.9548	**0.9858**	**0.8559**	0.9452	**0.8983**

Table 6.18: SVM Results: Train 2014-2015, Test 2016-2017

	TF_1		TF_2	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	6.28%	0.36%	6.27%	0.36%
True Ser	1.09%	92.27%	1.06%	92.30%

Table 6.19: Confusion Matrix: Train 2014-2015, Test 2016-2017

Again, the benchmark metrics seem consistent with those of the corpus in general. The classifier declined when tested on the 2016-2017 dataset as expected. Most interestingly, precision declined by about 10%. This is reflected by the relatively high percentage of false positives. This indicates that there is more similarity between the serious articles of the 2016-2017 data and the satirical articles of the 2014-2015 data than between the serious articles of the 2014-2015 data and the satirical articles of the 2014-2015 data. This would support a hypothesis that serious news has become more like satire of previous years.

6.6.2 Train on 2016 and 2017, Test on 2014 and 2015

Features	2016-2017 Benchmark				Test on 2014-2015			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
TF_1	0.9938	0.9607	0.9460	0.9531	0.9840	**0.9620**	0.8632	0.9099
TF_2	**0.9943**	**0.9633**	**0.9505**	**0.9568**	**0.9840**	0.9602	**0.8653**	**0.9103**

Table 6.20: SVM Results: Train 2016-2017, Test 2014-2015

	TF_1		TF_2	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	8.11%	1.28%	8.13%	1.27%
True Ser	0.32%	90.29%	0.34%	90.27%

Table 6.21: Confusion Matrix: Train 2016-2017, Test 2016-2017

The benchmark results indicate that satirical and serious articles from 2016-2017 may be more easily separable than those from the corpus as a whole. Testing on

2014-2015 supports the conclusion drawn previously, that serious articles from 2016-2017 are relatively similar to satirical articles from 2014-2015. This is reflected by the decline in recall of almost 10% and the increase in false negatives. However, these results do not seem to indicate that serious and satirical articles from 2016-2017 are harder to separate. This suggests that although serious news is becoming more similar to satirical news of the past, recent satire has changed to remain distinct from current serious reporting. It may be the case that serious news is adjusting to new cultural or political norms and that satire must do the same to keep a distinct voice. This could result in serious news articles that seem closer to satirical articles of the past but remain distinct from contemporary satire.

6.6.3 Balanced Data Sets

When considering the separations by year previously, all articles from each time period were used. The 2014-2015 set had 3,488 satirical articles and 33,649 serious ones compared to 3,374 satirical articles and 47,450 serious ones in the 2016-2017 set. These extra 14,000 serious articles could have affected the classifier trained on 2016-2017 data by causing it to overfit when compared to the classifier trained on the 2014-2015 data. Unlike for the previous time split, limiting each set to about 3000 satirical articles and 30,000 serious articles had a significant impact on performance.

Features	2014-2015 Benchmark				Test on 2016-2017			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
TF_1	**0.9919**	**0.9516**	0.9567	**0.9541**	0.9841	0.8898	**0.9417**	0.9150
TF_2	0.9913	0.9410	**0.9620**	0.9511	**0.9860**	**0.9123**	0.9360	**0.9240**

Table 6.22: SVM Results: Train 2014-2015, Test 2016-2017, Balanced

	TF_1		TF_2	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	8.56%	0.53%	8.56%	0.58%
True Ser	1.06%	89.85%	0.82%	90.09%

Table 6.23: Confusion Matrix: Train 2014-2015, Test 2016-2017

The benchmark appears comparable to the unbalanced set. Again there is an increase in false positives and a significant decline in precision, but it is not as dramatic as the one found with the larger data sets. Part of this might reflect the randomness of the selection.

We next train on the 2016-2017 data and test on the 2014-2015 data.

Features	2016-2017 Benchmark				Test on 2014-2015			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F
TF_1	**0.9908**	**0.9617**	0.9632	**0.9624**	0.9842	**0.9435**	0.8787	0.9099
TF_2	0.9898	0.9489	**0.9704**	0.9593	**0.9846**	0.9400	**0.8873**	**0.9129**

Table 6.24: SVM Results: Train 2016-2017, Test 2014-2015, Balanced

	TF_1		TF_2	
	Predicted Sat	Predicted Ser	Predicted Sat	Predicted Ser
True Sat	7.99%	1.10%	8.07%	1.02%
True Ser	0.48%	90.43%	0.52%	90.39%

Table 6.25: Confusion Matrix: Train 2016-2017, Test 2014-2015, Balanced

Most notable here is the drop in accuracy of the benchmark set compared with the unbalanced data set, while precision and recall appear to have improved. On the testing set, there is a large decline in recall and a corresponding rise in false negatives. These results still support the hypothesis of the previous section but suggest that the larger 2016-2017 data set did have an effect on the classifier.

Chapter 7

Conclusion

> "Information is only as reliable as the people who are receiving it. If readers do not change or improve their ability to seek out and identify reliable information sources, the information environment will not improve."
>
> Julia Koller, 2017

This thesis has explored the relationship between serious news, satire, and "fake news" and how it has changed over time. We found that Support Vector Machines tend to perform best with all available features, regularly achieving an accuracy of over 99% and precision, recall, and F score around 95% for this data set. It seems that the TF-IDF weighting with balanced classes tends to increase accuracy and recall at the cost of precision by significantly increasing the number of false positives while decreasing the number of false negatives. We also find that a C-LSTM approach achieved almost 99% accuracy, with precision, recall, and F measure around 90%. Although the C-LSTM performed worse than the SVM and required more computational resources, it still outperformed other classifiers found in the literature [5] [39]. The advantage of the C-LSTM approach is flexibility, as no features need to be explicitly determined. No additional preprocessing would be required for the C-LSTM model to classify different categories of text. Overall, both methods seem effective in identifying satirical articles

When comparing satire and serious news across all years, the results for this corpus were substantially better than those achieved by using the corpus from Yang et al [39]. The number of false negatives increased greatly, resulting in a much lower recall than for this project's corpus. The trend of the binary weighting for bag of words with no class weights resulting in a higher precision but lower accuracy and recall persisted. There are a few reasons why the classifier performed worse on the Yang et al dataset than this project's corpus. The first is that the Yang et al dataset lacked certain information, such as date and title, that eliminated some of the more effective features. This explanation is not sufficient because the SVM performed better on this project's corpus with only the bag of words. A second explanation points to the decision of Yang et al to include different sources and separate those sources for the testing and training set. The corpus described in [39] contained both serious and satirical articles from other English-speaking countries such as the UK and Canada. It also provides only articles from *The Onion* and *The Spoof* for the satirical part of the training set and tests and validates on many more satirical sources. This thesis included samples from all available satirical sources in the training set. By keeping some sources out of the training set, Yang et al were able to simulate how a classifier would behave with a never-before-seen source. This decision does limit the ability of the classifier to get a broader view of satire and was not done in other sources [1] [5]. This is likely to have a much larger impact on classifier performance than the inclusion of sources from other nations. By randomly shuffling the data, this project saw a modest gain in accuracy and a large one in recall. This suggests that the inclusion of more satirical sources in the training set did have a substantial impact on classifier performance. Yang et al also employed different preprocessing techniques, such as removing all references to specific sources. This thesis removed structural references but retained those unique to each article. Lastly, the data gathered for this corpus could contain topical similarities due to the way it was collected. Articles were discovered by following links on the pages of other articles. This could result in a focus on similar topics linked on a given article. However, due to the variety of sources, size of the data set, and content of the articles it is unlikely that this played a major role in the performance of this classifier.

This project also considered the problem of identifying serious news against "fake news." The classifier with a TF-IDF bag of words weighting and all features generally performed the best. This classifier saw a boost in all metrics over the problem of separating satire from serious news. This suggests that "fake news" may mimic real news to a lesser extent than satire in general. It may also reflect the fact that more outrageous articles that claim to be satirical are likely to be classified as "fake news." When comparing "fake news" and satire, with "fake news" being the positive class, the performance of the classifier declined. Accuracy dropped to around 97.5%, while the other metrics achieved around 95%. On one hand, the smaller amount of data may have contributed to this. With only 13,000 articles overall, the SVM had much less data to train on when compared to the case of serious news against satire. However, it also points to satire being more closely related to "fake news" than serious news, as was expected by this paper and supported by the literature [15].

In addition to considering "fake news" as a separate category, this thesis examined how the difference between satire and serious news changed over time. In all cases, this paper found a decline in performance when an SVM was trained on data from one time period and tested on data from another time period. Specifically, training on data from 2014-2017 and testing on data from 2010-2013 saw accuracy drop by about 1%, with the other measures falling between five and six percent when compared with training and testing on data from 2014-2017. This was a more drastic drop than training on 2010-2014 and testing on 2014-2017, suggesting that the features of older articles are somewhat distinct from those of newer articles.

There were similar declines when comparing data from 2014 and 2015 to data from 2016 and 2017. For these eras, false positives increase when trained on the 2014-2015 data and tested on the 2016-2017 data while false negatives increased when the reverse was tested even when controling for the size of the data sets. This suggests that serious articles from 2016-2017 are similar to satirical articles from 2014-2015. This may imply that serious news articles have become more similar to satirical news articles over time. This could come as a result of changes in the way people view and share news [13][14][18]. It could also reflect a change in norms, so that ideas that seemed ludicrous in the past have become subjects of serious reporting today [11]. The benchmark performance did not decline as dramatically, which indicates that serious and satirical articles from 2016-2017 are still distinct from each other. This could reflect the change in serious news. If current serious news is becoming more like the satire of the past, current satire must go further to distinguish itself. If serious news begins to push the limits of what was previously satire, satire must set new limits.

In summary, this paper suggests that it is possible to accurately classify satirical and "fake news" articles using some relatively simplistic tools. It also suggests that the year an article was published may be important in identifying its nature. By expanding on this project by adding more features and more data, it seems likely that such classifiers will only improve. Some such possible expansions will be discussed below, along with the the impact an integration of such technology into social media might have.

Chapter 8

Future Work & Applications

"Several hours or several weeks
I'd have the cheek to say
they're equally as bleak"

Do Me A Favour
Arctic Monkeys, 2007

As discussed in Chapter 1, this thesis is meant to be a step in the direction of counteracting misinformation on the Internet specifically with the intent of handling "fake news." The problem of "fake news" is not one that seems likely to go away in the near future. As humans become more connected and spend more of their lives online, spreading disinformation becomes easier and more effective. Without better education on issues and a reliable way to identify disinformation, little progress is likely. This thesis suggests that machine learning can be effectively applied to the task of identifying "fake news" and that this is a topic that deserves future study.

This project can be expanded by adding data and computational resources or changing focus. It also has a number of potential applications, both in its current state and with some adjustments. These expansions and applications are discussed in detail below.

8.1 Expansions

8.1.1 False Positives

In the words of famous satirist Tom Lehrer, "political satire became obselete when Henry Kissinger was awarded the Nobel peace prize" [26]. The real complexity of detecting satirical articles can arise from serious articles that seem so absurd as to be satirical. In this vein, subreddit r/nottheonion is a community that shares serious news articles that contain titles or stories so absurd as to seem at first glance satirical. Similarly, some of President Trump's comments regarding foreign nations have made use of profanity, often a hallmark of satire. This has resulted in many serious news outlets reporting quotes that contain profanity in a way that is not typical. A good test for the robustness of this classifier would be to introduce a sufficient number of these articles, which would be strong candidates for false positives, and seeing how the classifier handles them. This approach might be somewhat laborious, as it would require humans to identify articles that appear satirical but are meant to be taken seriously and verify that the claims reported are true. The subreddit would be a good start, but it often contains international sources or ones known for sensationalist articles. If these likely false positives were to be considered in training, it would also be necessary to find enough so as to be significant in building the classifier.

8.1.2 Other Categories

This paper makes focuses on two main categories (serious and satirical) as well as the third category of "fake news." These three categories are of course insufficient to adequately describe the complex world of news and text, so a few more possible demarcations are considered below.

Parody/Comedy

Satire is not the only category of news meant for entertainment. For instance, to use an example from television, one might classify *The Daily Show* or *Last Week Tonight* as merely humorous compared to *The Colbert Report* as overt satire. Further divisions could be introduced; [37] considers irony, parody, and satire as distinct but overlapping classifications. Such distinctions were avoided in this paper for two principle reasons. The first is that it would complicate the labeling process immensely, especially as the lines between categories are blurred and individual works may contain both satire and parody. The second reason is that the majority of the works that fit firmly in one of the other categories are easier for humans to recognize. Satire aims explicitly mimic reality in a believable way but other categories such as comedy tend to be more open and apparent. Since these are easier for humans to identify, they are less frequently confused for the truth or labeled as "fake news." For these two reasons, this paper focused specifically on satire and attempted to avoid any ambiguous labeling. Research into more specific labeling for such work could prove useful in some applications for increased performance.

Bias & Reliability

Not all serious news sources are created equally. It would not be hard to argue that many popular and credible publications come with a striking bias in one political direction or another. It is also clear that some sources rely on more investigation and diligent reporting than others. It would undoubtedly be useful to be able to use a classifier to label bias in news sources or give them a reliability score. A similar approach to the one taken in this paper might be useful in these cases. The problem with these labels is that they are more subjective and harder to label than the two simple ones used in this paper. Individuals may disagree on the overall bias of a source, or the source may be biased in one way on certain issues and in another on other issues. Reliability would be an easier thing to measure objectively, but could take extensive research to create a sizable corpus. A reliability check would again have to confront the notion of truth, likely by building consensus among established publications. Classifiers for both of these topics may be particularly useful as bias and reliability are easily overlooked by many readers.

A General Classifier

The implications of this classifier may extend beyond the realm of online news and satire. Theoretically, it seems likely that any successes of this classifier stem from an ability to determine the *style* associated with satire rather than evaluating the truth value of the article. This approach could be generalized to a number of similar problems. For instance, such a classifier might prove effective in predicting whether or not college admissions essays were written by applicants. Some admissions essays could be written by parents or professional coaches in an attempt to create a stronger application. Fundamentally, this classification task is the same as the one studied in this thesis, in that a classifier would attempt to find stylistic differences between the writing of high school seniors and adults. As needed, different features explored in this paper could be adjusted. Since college admissions essays are unlikely to contain profanity, a more nuanced look at slang and generational differences in word usage could be explored. This data may be difficult to gather, as there are confidentiality concerns about college essays (since they often contain personal information). Additionally, it there may be a relatively small number of essays which can be guaranteed to be from someone other than a student. If these challenges can be overcome, there is no reason this project could not be adapted to serve a similar function with this dataset.

Note that the deep learning approach would require little modification. Provided that the data set was appropriately cleaned and labeled, it would only be a matter of vectorizing the articles in the same way as before. This reflects a strength of deep learning, in that few assumptions about or operations on the input are necessary.

8.1.3 Other Classifiers

The classifiers used in this project were chosen based on their use in the satire detection literature. To continue with supervised learning, Naïve Bayes Classifiers or other forms of neural networks could be tried. For instance, a feed forward neural network could be evaluated on all of the features used in the SVM analysis. One could also try unsupervised or semi-supervised learning in an attempt to cluster satirical articles and serious articles separately.

8.1.4 Other Features

Although adding more features comes with the trade-off of increasing dimensionality, there may be some that could improve the results of this project. Sentiment analysis is an obvious expansion (used in [37]). It is possible that satirical or "fake news" articles have a recognizable emotional pattern arising from their goal of challenging or manipulating readers. Another expansion would be an analysis of comments posted on the articles, similar to the way *YouTube* comments were considered in [37]. Responses to satirical articles might be different than those to serious articles and may even explicitly state that the article is satire. Depending on how the comments are gathered, it might also be possible to leverage some metadata about the poster of the article or commenters. This might be difficult for a third party due to privacy restrictions, but social networks (such as Facebook) could utilize an understanding of the network to aid classification.

Another possible feature would be the advertisements shown on a page. It is possible that advertisers would be willing to pay more to be on reputable sites than satirical ones. "Fake news" sites may lack advertisement altogether [38] or have advertisements from less reputable brands. It might also be the case that different products or services tend to be advertised on each type of site. One would expect to see ads for comedy shows or humorous products on satirical pages, while serious sites may have more automotive or fashion advertisements. Similarly, one could gather a PageRank measure of how important each site is by analyzing its ordering in a number of searches with several keywords. This approach has a number of limitations, but it could help identify potentially untrustworthy sites with little traffic.

Utilizing a psycholinguistic library such Linguistic Inquiry and Word Count (LIWC) as in [39] may yield some useful features. Such a dictionary may reveal more impassioned words in satire and clearer ones in serious news. The use of this library was omitted from this project due to cost constraints. Similarly, some information might be gathered from comparing the relative frequency of parts of speech. [39] hypothesized that satirical news articles would contain more imaginative language and so identifying parts of speech would aid in detecting satire. This could be accomplished using the `Python` library `nltk`.

8.1.5 Other Types of Media

One limitation of this project is that it only considers textual sources. This decision was motivated by two primary factors: the spam filter model and the prevalence of both satire and "fake news" in a text form. Spam detection deals almost exclusively with textual analysis, as do many methods in Natural Language Processing (NLP). Similarly, most researchers ([1] [5] [29] etc.) emphasize textual data due to its prevalence and the amount of study it receives. For these reasons and the fact that many of the articles shared online are textual, text was the natural medium to consider for this paper.

Text is of course not the only way information is shared. The following sections will briefly consider other possible forms of satire and serious news.

Video

There are dozens of televised news networks and many articles on major news sites include a video component. There are also shows that straddle the line between satire, comedy, and news (*Last Week Tonight*, *The Late Show*, *The Daily Show*, etc.). This is definitely a field that deserves study, similar to the work done in [37]. However, in [37] the authors relied on human annotaters to classify videos. This is at best time-consuming and expensive and at worst unreliable and prone to disagreement. Even with the data labeled, decisions must be reached about what features to use. [37] focused heavily on a transcript of the audio and metadata about the video. Another approach might use some sort of computer vision method (such as a convolutional neural network)to analyze the video itself. This by nature will likely be computationally expensive if possibly more accurate. The methods described in this project could work similarly using a transcript of the video and some associated metadata. This might not be perfect though - sometimes the joke may be purely visual and so not encapsulated by the transcript. Additionally, transcripts are likely to introduce more errors, especially if they have been translated or done automatically. The problem of categorizing satire in video is one that requires more analysis than is possible here.

Audio & Pictorial

Satire may also be presented purely through audio or a pictorial representation. The first of these does not seem especially common, but pictures are frequently used in satire. *The Onion*, to give an example, frequently shows pictures with only a title as an entire article. Again, the approach would likely involve computer vision and a thorough use of metadata. It might also be possible to search for evidence of alterations in satirical pictures, especially those concerning famous people or locations. If no suitable picture exists to fit an article, one might be drawn or generated by digitally editing a genuine picture. Thus, evidence of alteration may indicate a satirical piece.

Images may have particular importance in the realm of advertisements, where social media companies have seen misleading ads designed to influence or deceive those reading them [10]. Similar methods to those used for satire detection could be employed to help identify fake advertisements.

Taken together, these approaches could be applied to accounts to identify bots. Metadata from the account, in addition to an analysis of text, image, and link posts, could provide information about the likely authenticity of the account.

8.1.6 Satire in a Given Network

One limitation of this project is that it takes articles directly from the source, without gathering information about how the article may have been received or shared. This disregards a large number of possibly valuable features. If we were to pick a social media network such as Facebook and try to detect links to satirical articles versus serious articles, we could use information about the poster's profile including number of friends, page likes, number and frequency of posts, stated political affiliation, age of the account, demographics, and more. From the post itself, we could get the associated status, the number of reactions and shares, and the number and content of comments. In some cases, the poster may acknowledge the nature of the shared article and in others they may be duped or attempting to deceive friends. Either way, the nature of the post and the responses to it would provide valuable information to a classifier.

This approach may run in to privacy concerns, as the data would need to be gathered and anonymized. Similarly, some features may simply add noise and detract from the task of making a judgment about the article itself. This may be an expansion worth doing, but would require significantly more time and resources than were available for this thesis.

8.2 Application: Integration into Social Media

Once satirical articles have been identified and labeled, the question emerges of what to do with that information. This is not a trivial issue. Clearly, censorship should be avoided and all satirical articles should be available to those who would like to read them. It is less clear what standards should be applied to articles considered "fake news" or suspected of being entirely fabricated to deceive.

In [34], research found that the person who shared a post on social media was more important to others believing it than the content or source of the publication. This implies that providing more information could have a serious effect on how articles are interpreted and shared. To avoid censorship, such sites could simply provide a warning every time someone goes to share or open a link to a satirical or "fake news" article. The intent of this would be to alert the reader/sharer to the veracity of the article without preventing them from accessing it. Similarly, it might be useful for a social network to attempt to identify articles frequently shared by bot accounts and provide a warning.

However, this approach is not without its problems. On one hand, this may be opposed by different groups. Well-established, legitimate satirical sites may not wish for all of their articles to come with a satire tag, especially if it is grouped with sources considered "fake news." On the other hand, research suggests that labels for satire or "fake news" may create a level of complacency, in that people may become less critical of articles that do not have a "fake news" label. In [24], researchers found that although tags do reduce trust in labeled "fake news" articles, they also increase trust in articles that are unlabeled. Intuitively, users may assume that all untrustworthy articles are labeled, and so they cease to be suspicious of those that are unlabeled. This may be a difficult problem to overcome. One solution would be to provide a warning that contains a probability for the article being satirical or fake and allowing user feedback to update that probability. Instead of relying on a label, all articles could get a confidence score. This would allow for more levels of nuance, but would require additional research for how best to provide this information.

Recently, Facebook has included an information button in all news posts. Clicking this provides a description of the website of origin and some additional information. This will allow users to learn some basic information about the source they are reading. It is unclear what this displays for new websites or those not tracked by Facebook. Additionally, this requires users to click to get information and so is an "opt-in" service. A warning such as the one outlined above could be made to appear without user input, which would make it more difficult to miss or ignore.

This algorithm may be used in conjunction with others to identify bots or other accounts that spread "fake news" or misinformation. Being able to realize which articles are problematic may make it easier to flag potentially malicious accounts.

Of course, deploying such a classifier might cause "fake news" sources to change their articles to avoid flagging. This could be accomplished through machine learning ways to avoid the classifier or by simply testing the integrated classifier on social media. It is possible this would lead to an arms race between creators of "fake news" and those aiming to identify it. Scott Spangler of IBM even went so far as to suggest that machine learning could be used to "make fake information almost indistinguishable from the real thing." [35] Regardless, this implementation would still catch many articles, and impose a cost on making effective "fake news."

Appendix A

Code Excerpts

For the full code, visit `https://github.com/ahare63/Thesis`.

A.1 Learning SVM Hyperparameters

Excerpt from `trainHPNobel.py`

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
from sklearn.model_selection import GridSearchCV
from scipy.sparse import hstack
from sklearn import preprocessing

allSatire = pd.read_csv('./combinedSets/allSatire.csv')
allSatire = allSatire.drop(['Unnamed: 0'], axis=1).sample(frac=.85)
allSerious = pd.read_csv(./combinedSets/allSerious.csv')
allSerious = allSerious.drop(['Unnamed: 0'], axis=1).sample(frac=.85)
train = pd.concat([allSatire, allSerious],
    ignore_index=True).dropna(how='any',
    subset={'Body'}).sample(frac=1)

vectorizer = CountVectorizer(stop_words='english', binary=True)
binaryBog = vectorizer.fit_transform(train.Body)
print('Vectorized train')

labels = train["isSatire"].values
columns = ['ARI', 'FR', 'GF', 'avgSyl', 'linkCount',
    'profanityCount', 'senCount', 'titleARI',
          'titleAvgSyl', 'titleFR', 'titleGF', 'titleWordCount',
             'twitChar', 'wordCount', '2010', '2011', '2012', '2013',
             '2014', '2015', '2016', '2017']
features = preprocessing.scale(train[list(columns)].values)
features = hstack([binaryBog, features])
```

```
param = {'class_weight': [None, 'balanced'], 'C': [10**-5, 10**-4,
    10**-3, 10**-2, 10**-1, 1, 10, 10**2, 10**3]}

svc = LinearSVC()
clf = GridSearchCV(svc, param, scoring='accuracy')
clf.fit(features, labels)
print(clf.cv_results_)
print(clf.best_estimator_)
print(clf.best_params_)
print(clf.best_score_)
print('All features done')
```

A.2 Testing SVM

Excerpt from SVMTestNobel.py

```
train = pd.read_csv('./combinedSets/train.csv')
# sample to shuffle
train = train.drop(['Unnamed: 0'], axis=1).sample(frac=1)
test = pd.read_csv(./combinedSets/test.csv')
test = test.drop(['Unnamed: 0'], axis=1).sample(frac=1)

# Get binary bag of words
vectorizer = CountVectorizer(stop_words='english', binary=True)
bogTrain = vectorizer.fit_transform(train.Body)
bogTest = vectorizer.transform(test.Body)
svc = LinearSVC(C=0.01, class_weight=None)
labels = train["isSatire"].values

# Get other features
columns = ['titleAvgSyl', 'titleFR', 'titleGF', 'titleWordCount',
    'titleARI']

# Normalize features
features = preprocessing.scale(train[list(columns)].values)
features = hstack([bogTrain, features])
svc.fit(features, labels) # Fit SVM

Xtest = preprocessing.scale(test[list(columns)].values)
Xtest = hstack([bogTest, Xtest])
yT = svc.predict(Xtest)
printMeasures(yT, Ytest, scoreList, 1) # Record performance measures
```

A.3 Testing C-LSTM

Excerpt from NNTest.py

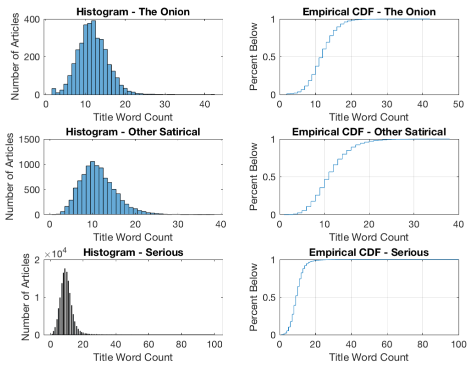
```
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.models import Sequential
from keras.layers.embeddings import Embedding
from keras.layers import Dropout, Conv1D, LSTM, Dense
import keras.backend as K
from keras.callbacks import EarlyStopping

# fix out of memory errors
config = tf.ConfigProto()
config.gpu_options.allow_growth = True
K.set_session(tf.Session(config=config))

train = pd.read_csv('./combinedSets/train.csv')
test = pd.read_csv('./combinedSets/test.csv')

# tokenize and pad
maxWords = 20000
tokenizer = Tokenizer(num_words=maxWords)
tokenizer.fit_on_texts(train.rawText)
asSequence = tokenizer.texts_to_sequences(train.rawText)
maxlen = 10000
xTrain = np.array(pad_sequences(asSequence, maxlen=maxlen))
testAsSequence = tokenizer.texts_to_sequences(test.rawText)
xTest = np.array(pad_sequences(testAsSequence, maxlen=maxlen))
# remove words not seen in training set
# 0 reserved for unknown string in preprocessing
xTest[xTest > maxWords] = 0

# specify model
model_CLSTM = Sequential()
model_CLSTM.add(Embedding(maxWords + 1, 100, input_length=maxlen))
model_CLSTM.add(Dropout(0.3))
model_CLSTM.add(Conv1D(filters=64, kernel_size=3, activation='relu'))
model_CLSTM.add(LSTM(units=64))
model_CLSTM.add(Dense(1, activation='sigmoid'))

model_CLSTM.compile(loss='binary_crossentropy', optimizer='Adam',
   metrics=['accuracy', get_prec, get_rec, get_f])
callbacks = [EarlyStopping(monitor='loss')]

model_CLSTM.fit(xTrain, y=np.array(train.isSatire),
    callbacks=callbacks, batch_size=512, epochs=10, shuffle=True),
    class_weight={0: 1, 1: 10})
a = model_CLSTM.evaluate(xTest, np.array(test.isSatire))
print(" %.4f & %.4f & %.4f & %.4f \\\\" % (a[1], a[2], a[3], a[4]))
```

Appendix B

Additional Results

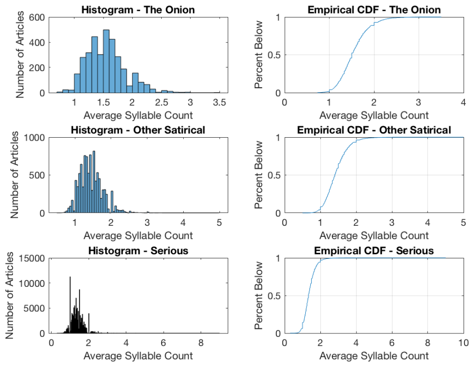
B.1 The Onion: Additional Feature Data

Figure B.1: Onion Title Word Count



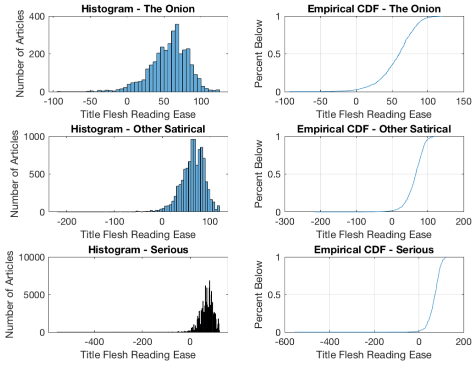
	Onion	Other Sat	Serious
Mean	11.7967	11.3174	9.6677
Std Dev	3.7966	4.1875	3.7362
Median	12	11	9
Mode	12	10	9
Minimum	2	1	1
5th Percentile	6	5	4
95th Percentile	18	19	16
Maximum	42	38	100

Figure B.2: Onion Title Average Syllable Count



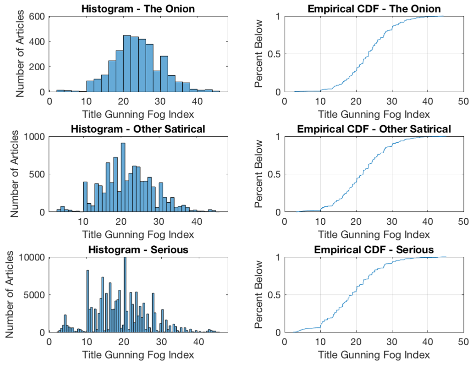
	Onion	Other Sat	Serious
Mean	1.5428	1.4387	1.3763
Std Dev	0.32796	0.32112	0.32907
Median	1.5	1.4	1.3333
Mode	1.5	1.5	1
Minimum	0.73333	0.5	0.30769
5th Percentile	1.0769	1	0.91667
95th Percentile	2.1429	2	2
Maximum	3.5	5	9

Figure B.3: Onion Title Flesh Reading Ease



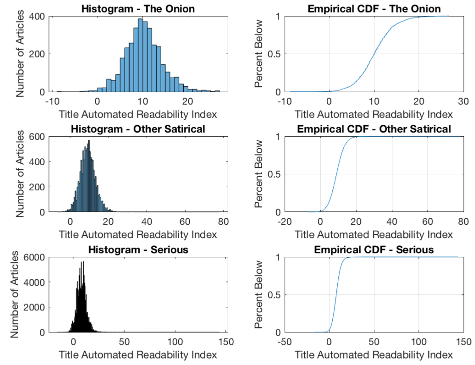
	Onion	Other Sat	Serious
Mean	56.9164	63.8444	69.1616
Std Dev	25.5416	24.7039	26.7231
Median	60.31	66.74	71.82
Mode	68.77	68.77	71.82
Minimum	-93.33	-217.19	-555.59
5th Percentile	9.55	20.176	24.61
95th Percentile	93.14	99.23	106.67
Maximum	120.21	119.19	121.22

Figure B.4: Onion Title Gunning Fog Index



	Onion	Other Sat	Serious
Mean	23.5832	21.8011	19.0927
Std Dev	6.5111	6.8617	7.3668
Median	23.4667	21.8857	18.9333
Mode	23.4667	22	20.2
Minimum	2.8	3.2	2.4
5th Percentile	13.4667	10.5143	7.2
95th Percentile	34.02	33.4667	31.0667
Maximum	44.4	45.2	45.2

Figure B.5: Onion Title Automated Readability Index

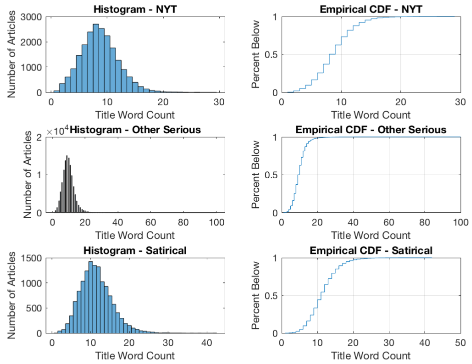


	Onion	Other Sat	Serious
Mean	10.1847	9.1288	7.9441
Std Dev	4.0406	4.0701	4.6948
Median	10.1	9	7.7
Mode	9.3	9.7	5.6
Minimum	-8.8	-6.8	-16.3
5th Percentile	3.795	2.8	0.9
95th Percentile	17.1	15.9	15.6
Maximum	26.7	78	143.9

B.2 The New York Times: Additional Feature Data

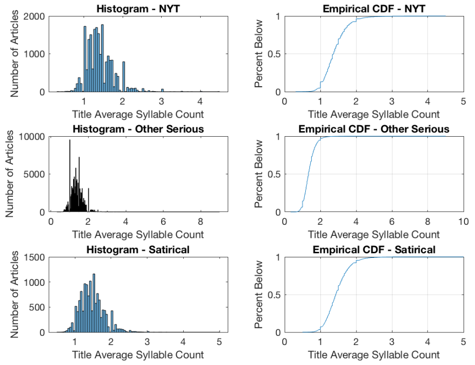
Note there may be a slight problem with spacing in some *Washington Post* articles, resulting in improperly long or short titles.

Figure B.6: New York Times Title Word Count



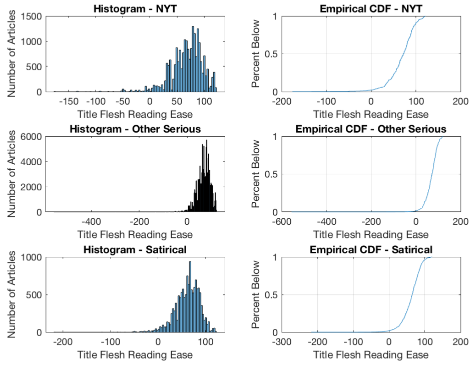
	NYT	Other Ser	Satirical
Mean	8.6808	9.8411	11.4404
Std Dev	3.3241	3.7775	4.096
Median	8	9	11
Mode	8	9	10
Minimum	1	1	1
5th Percentile	4	5	6
95th Percentile	14	16	19
Maximum	29	100	42

Figure B.7: New York Times Title Average Syllable Count



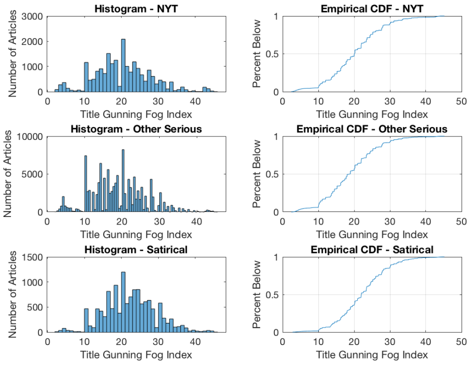
	NYT	Other Ser	Satirical
Mean	1.4079	1.3707	1.4654
Std Dev	0.34886	0.32516	0.32607
Median	1.3636	1.3333	1.4286
Mode	1	1	1.5
Minimum	0.30769	0.33333	0.5
5th Percentile	1	0.91667	1
95th Percentile	2	2	2
Maximum	4.5	9	5

Figure B.8: New York Times Title Flesh Reading Ease



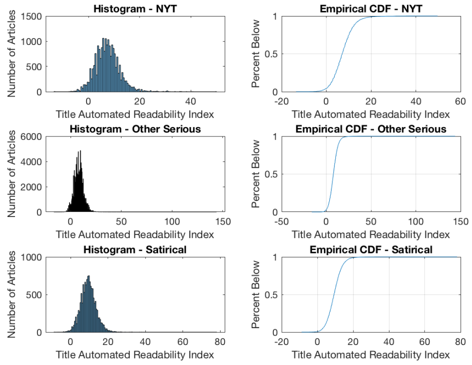
	NYT	Other Ser	Satirical
Mean	68.406	69.2944	62.0663
Std Dev	28.4771	26.4006	25.1037
Median	71.82	71.82	64.71
Mode	81.29	68.77	68.77
Minimum	-175.9	-555.59	-217.19
5th Percentile	19.03	26.47	18.01
95th Percentile	106.67	106.67	96.18
Maximum	121.22	121.22	120.21

Figure B.9: New York Times Title Gunning Fog Index



	NYT	Other Ser	Satirical
Mean	20.2196	18.8946	22.2584
Std Dev	7.658	7.2965	6.8178
Median	20.1333	18.6667	22
Mode	20.2	20.2	23.3778
Minimum	2.4	2.4	2.8
5th Percentile	9.5943	6.8	10.6667
95th Percentile	33.3714	31.0667	33.6
Maximum	45.2	45.2	45.2

Figure B.10: New York Times Title Automated Readability Index



	NYT	Other Ser	Satirical
Mean	7.3181	8.0541	9.3998
Std Dev	4.7259	4.6807	4.0885
Median	7.1	7.9	9.3
Mode	5.6	5.6	9.3
Minimum	-13.5	-16.3	-8.8
5th Percentile	0.4	0.9	2.9
95th Percentile	15	15.7	16.2
Maximum	49.7	143.9	78

Bibliography

[1] T. Ahmad et al. ``Satire Detection from Web Documents Using Machine Learning Methods''. In: *2014 International Conference on Soft Computing and Machine Intelligence*. 2014, pp. 102–105. DOI: `10.1109/ISCMI.2014.34`.

[2] Hunt Allcott and Matthew Gentzkow. ``Social Media and Fake News in the 2016 Election''. In: *Journal of Economic Perspectives* 31.2 (2017), pp. 211–236. URL: `https://web.stanford.edu/~gentzkow/research/fakenews.pdf`.

[3] Enrico Blanzieri and Anton Bryl. ``A Survey of Learning-based Techniques of Email Spam Filtering''. In: *Artif. Intell. Rev.* 29.1 (Mar. 2008), pp. 63–92. ISSN: 0269-2821. DOI: `10.1007/s10462-009-9109-6`. URL: `https://doi.org/10.1007/s10462-009-9109-6`.

[4] N. Boztas et al. ``Readability of internet-sourced patient education material related to "labour analgesia"''. English. In: *MEDICINE* 96.45 (2017).

[5] Clint Burfoot and Timothy Baldwin. ``Automatic Satire Detection: Are You Having a Laugh?'' In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. ACLShort '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 161–164. URL: `http://dl.acm.org/citation.cfm?id=1667583.1667633`.

[6] Yimin Chen, Niall Conroy, and Victoria Rubin. ``News in an Online World: The Need for an " Automatic Crap Detector "''. In: *The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015), Nov.* Vol. 6.10. Oct. 2015.

[7] Conal Condren. ``Satire and definition''. In: *Humor* 25.4 (Nov. 2012), pp. 375–399. DOI: `https://doi.org/10.1515/humor-2012-0019`.

[8] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. ``Automatic Deception Detection: Methods for Finding Fake News''. In: University of Western Ontario, London, Canada, 2015. URL: `https://www.asist.org/files/meetings/am15/proceedings/submissions/posters/193poster.pdf`.

[9] Dmitry Davidov, Oren Tsur, and Ari Rappoport. ``Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon''. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. CoNLL '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 107–116. ISBN: 978-1-932432-83-1. URL: `http://dl.acm.org/citation.cfm?id=1870568.1870582`.

[10] Matt Drange. ``In Fake News Era, Facebook Still Struggling With Bogus Ads''. In: *Forbes* (2017). URL: https://www.forbes.com/sites/mattdrange/2017/03/02/a-basic-design-feature-makes-it-easy-to-create-fake-advertisements-on-facebook.

[11] Daniel W Drezer. ``Donald Trump's three types of norm violations''. In: *Washington Post* (2016). URL: https://www.washingtonpost.com/posteverything/wp/2016/12/19/donald-trumps-three-types-of-norm-violations/.

[12] M. Glenski, C. Pennycuff, and T. Weninger. ``Consumers and Curators: Browsing and Voting Patterns on Reddit''. In: *IEEE Transactions on Computational Social Systems* 4.4 (2017), pp. 196–206. DOI: 10.1109/TCSS.2017.2742242.

[13] Jeffrey Gottfried and Elisa Shearer. *Americans' online news use is closing in on TV news use*. 2017. URL: http://www.pewresearch.org/fact-tank/2017/09/07/americans-online-news-use-vs-tv-news-use/.

[14] Jeffrey Gottfried and Elisa Shearer. *News Use Across Social Media Platforms 2016*. 2016. URL: http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

[15] Benjamin D. Horne and Sibel Adali. ``This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News''. In: *CoRR* abs/1703.09398 (2017). arXiv:1703.09398. URL: http://arxiv.org/abs/1703.09398.

[16] Charles F. Bond Jr. and Bella M. DePaulo. ``Accuracy of Deception Judgments''. In: *Personality and Social Psychology Review* 10.3 (2006). PMID: 16859438, pp. 214–234. DOI: 10.1207/s15327957pspr1003_2. eprint: https://doi.org/10.1207/s15327957pspr1003_2. URL: https://doi.org/10.1207/s15327957pspr1003_2.

[17] The Killers. *Believe Me Natalie*. From album *Hot Fuss*, written by Flowers and Vannucci. Las Vegas, 2004.

[18] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. ``News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks''. In: *Social Media + Society* 1.2 (2015), p. 2056305115610141. DOI: 10.1177/2056305115610141. eprint: https://doi.org/10.1177/2056305115610141. URL: https://doi.org/10.1177/2056305115610141.

[19] Kim LaCapria. *Snopes' Field Guide to Fake News Sites and Hoax Purveyors*. 2017. URL: http://www.snopes.com/2016/01/14/fake-news-sites/.

[20] Garrett Mattingly. ``Machiavelli's "Prince": Political Science or Political Satire?'' In: *The American Scholar* 27.4 (1958), pp. 482–491. ISSN: 00030937, 21622892. URL: http://www.jstor.org/stable/41208453.

[21] Arctic Monkeys. *Do Me A Favour*. From album *Favourite Worst Nightmare*, written by Turner. London, 2007.

[22] Randall Munroe. ``Machine Learning''. In: *xkcd* (). URL: https://xkcd.com/1838/.

[23] Kenneth Olmstead, Amy Mitchell, and Tom Rosenstiel. ``Navigating News Online: The Top 25''. In: *Pew Research Center* (2011). URL: http://www.journalism.org/2011/05/09/top-25/.

[24] Gordon Pennycook and David G. Rand. ``Assessing the Effect of 'Disputed' Warnings and Source Salience on Perceptions of Fake News Accuracy''. In: *SSRN* (2017). URL: https://ssrn.com/abstract=3035384.

[25] María del Pilar Salas-Zárate et al. ``Automatic detection of satire in Twitter: A psycholinguistic-based approach''. In: *Knowledge-Based Systems* 128.Supplement C (2017), pp. 20 –33. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2017.04.009. URL: http://www.sciencedirect.com/science/article/pii/S0950705117301855.

[26] Todd S Purdom. ```'When Kissinger won the Nobel peace prize, satire died'''. In: *The Guardian* (2000). URL: https://www.theguardian.com/culture/2000/jul/31/artsfeatures1.

[27] Kumar Ravi and Vadlamani Ravi. ``A novel automatic satire and irony detection using ensembled feature selection and data mining''. In: *Knowledge-Based Systems* 120.Supplement C (2017), pp. 15 –33. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2016.12.018. URL: http://www.sciencedirect.com/science/article/pii/S0950705116305226.

[28] Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. ``Deception Detection for News: Three Types of Fakes''. In: *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. ASIST '15. St. Louis, Missouri: American Society for Information Science, 2015, 83:1–83:4. ISBN: 0-87715-547-X. URL: http://dl.acm.org/citation.cfm?id=2857070.2857153.

[29] Victoria L. Rubin et al. ``Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News." In: NAACL-CADD2016. 2016. URL: https://www.researchgate.net/profile/Victoria_Rubin/publication/301650504_Fake_News_or_Truth_Using_Satirical_Cues_to_Detect_Potentially_Misleading_News/links/571fed4c08aeaced788acd8e/Fake-News-or-Truth-Using-Satirical-Cues-to-Detect-Potentially-Misleading-News.pdf.

[30] Jim Rutenberg. ``RT, Sputnik and Russia's New Theory of War''. In: *The New York Times Magazine* (2017), p. MM44. URL: https://www.nytimes.com/2017/09/13/magazine/rt-sputnik-and-russias-new-theory-of-war.html?emc=edit_ta_20170913&nl=top-stories&nlid=62144678&ref=cta.

[31] Sydney Schaedel. *Websites that Post Fake and Satirical Stories*. 2017. URL: http://www.factcheck.org/2017/07/websites-post-fake-satirical-stories/.

[32] Fabrizio Sebastiani. ``Machine Learning in Automated Text Categorization''. In: *ACM Comput. Surv.* 34.1 (Mar. 2002), pp. 1–47. ISSN: 0360-0300. DOI: 10.1145/505282.505283. URL: http://doi.acm.org/10.1145/505282.505283.

[33] The Smiths. *That Joke Isn't Funny Anymore*. From album *Meat Is Murder*, written by Morrissey and Marr. London, 1985.

[34] Media Insight Project Team. ``'Who shared it?': How Americans decide what news to trust on social media'''. In: *American Press Institute* (2017). URL: `https://www.americanpressinstitute.org/publications/reports/survey-research/trust-social-media/`.

[35] ``The Future of Truth and Misinformation Online''. In: *Pew Research Center* (2017). Used as source for various quotes throughout paper and in chapter epigrams. URL: `http://www.pewinternet.org/2017/10/19/shareable-quotes-from-experts-on-the-future-of-truth-and-misinformation-online/#`.

[36] *Truth is Incontrovertible*. 2017. URL: `https://www.winstonchurchill.org/resources/quotes/truth-is-incontrovertible/`.

[37] Joshua L Weese et al. ``Parody Detection: An Annotation, Feature Construction, and Classification Approach to the Web of Parody''. In: *Data Analytics in the Digital Humanities*. Springer, 2017, 67–89. DOI: `10.1007/978-3-319-54499-1_3`.

[38] Nick Wingfield, Mike Isaac, and Katie Benner. ``Google and Facebook Use Ad Policies to Take Aim at Fake News Sites''. In: *The New York Times* (2016), B1. URL: `https://www.nytimes.com/2016/11/15/technology/google-will-ban-websites-that-host-fake-news-from-using-its-ad-service.html`.

[39] Fan Yang, Arjun Mukherjee, and Eduard Constantin Dragut. ``Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features''. In: *EMNLP*. 2017. DOI: `arXiv:1709.01189v1`.

[40] Le Zhang, Jingbo Zhu, and Tianshun Yao. ``An Evaluation of Statistical Spam Filtering Techniques''. In: 3.4 (Dec. 2004), pp. 243–269. ISSN: 1530-0226. DOI: `10.1145/1039621.1039625`. URL: `http://doi.acm.org/10.1145/1039621.1039625`.

[41] Chunting Zhou et al. ``A C-LSTM Neural Network for Text Classification''. In: *CoRR* abs/1511.08630 (2015). arXiv:`1511.08630`. URL: `http://arxiv.org/abs/1511.08630`.