Master's Thesis

Fake News Evolution - Linguistic Characteristics

Chrysovalantis Christodoulou

University of Cyprus

Department of Computer Science

December 2020

Acknowledgment

I would like to express my heartfelt gratitude to my supervisor Dr. George Pallis, Associate Professor in the Computer Science Department of University of Cyprus, for his endless willing and the opportunity that gave me to work my thesis in one of the biggest new-age problems. He gave me all the support and encouragement throughout the elaboration of my thesis project.

Moreover, I would like to thank the Ph.D. candidate Demetris Paschalidis for his excellent support and guidance, along with my colleagues who effortlessly help me through the process.

Last but not least, I would like to acknowledge and thank my family for their valuable help during my studies. Specifically, I would like to thank my parents Stelios and Maria and my sweetheart fiancee Elena for all the psychological and physical aid they graciously provided to me.

Abstract

Recently, the misinformation pandemic, known as “infodemic”, has been flooding the internet and social media. The crisis of misinformation is rapidly infecting our daily lives, and its impact on society and democracy worsens. The influence of fake news in events like the COVID-19 pandemic and the 2020 US presidential election indicate the urgent need to reduce the spread and penetration of misinformation and fake news, made from users either knowingly or inadvertently. Many recent approaches seek to mitigate misinformation through automatic identification with artificial intelligence techniques. However, due to the complexity of the problem, most of the algorithms are too specific with no generalized outcomes. In this work, we believe that the only way to stop this “infodemic”, is hidden in analyzing and understanding the evolution of it through time. Therefore, we developed a crawler to gather articles from trusted and untrusted sources, covering a period of 10 years. By collecting these articles, we construct a comprehensive dataset containing almost daily basis data with 500K documents. Having the dataset allows extracting a plethora of different linguistic features, for each article, indexing and visualized them to study their evolution. To the best of our knowledge, no previous work proposed any similar contribution.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

The widespread of online social networking and media platforms have changed dramatically the production and consumption of digital information. Any individual equipped with an Internet connection and a social media account can create and circulate content that can reach people at unprecedented speed and scale, without any prior moderation for inaccuracy or inappropriateness. Recently, the World Health Organization characterized the spread of misinformation, during the COVID-19 pandemic era, as an "infodemic"[1]. WHO director declares that Fake News are spread faster than the virus, and they actually harm people. We have dozen of examples, starting from the Donald Trump's 2016 US

[1] https://www.who.int/director-general/speeches/detail/munich-security-conference

presidential election campaign [2], the UK referendum for leaving the European Union, the 2020 United States presidential elections [3], and the ongoing pandemic crisis (COVID-19)[4].

Thinking about Fake News as a virus, a critical question arises: "How to stop a virus?". Taking COVID-19 as an example, at first you are trying to develop countermeasures to mitigate the direct effects. For example, there are measures aiming at the goodwill of the people, like washing hands and keeping social distances or even measures that forcing people to stay at home and wear masks. Corresponding measures are followed by the majority of the research community to mitigate misinformation. Multiple approaches are trying to develop algorithms to detect Fake News delivered through browser plugins to warn users Paschalides et al. 2019; Gupta et al. 2014. Some other researchers, attempting to stop the propagation of a story from the very beginning, and international organizations like the EU trying to set regulations and penalties about the spreading of misinformation [5].

In both cases, measures are crucial to reducing the effects of a virus but are not enough. Viruses are living beings, which are evolving and fighting back to survive. Fraud journalists, which are also living from writing misleading content, are also evolving to avoid our countermeasures. The actual way to stop it is to analyze the building blocks that make it up and studying how evolves over time. In this work, we are focusing on providing the essential ingredients to analyze the evolution of linguistic characteristics on both fake and real news over the years. More specifically, we construct a half million articles dataset covering a period from 2009 to 2019, with almost daily data. Then, we extract and analyze a plethora of linguistic features to study their evolution over time and identify possible explanations.

[2] `https://www.bbc.com/news/world-us-canada-37896753`

[3] `https://tinyurl.com/y3m9hpfl`

[4] `https://www.un.org/en/battling-covid-19-misinformation-hands`

[5] `https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation`

1.2 Challenges

By definition, Fake News is a challenging issue from every perspective. The confirmation of article veracity usually needs thorough research around the topic, and even expert fact-checking sites might delay 2-3 days to confirm a story. The reason is that Fake News are written by expert journalists who on purpose trying to bias the common thought to serve either political interests as it happens in the US 2016 elections and in the UK referendum for leaving the European Union or other interests like related to economics, religious, and psychology.

The issue is that those journalists are experts to avoid common mistakes such as grammatical or syntax and hide their purpose behind strong and powerful words. Thus, our natural language analysis becoming harder and harder, and we have to take into account those parameters. Therefore, we had to study the relative literature to select every possible feature that might give us a better understanding of the text and help us analyze the differences and similarities between fake and real articles over the years.

However, analyzing features from the news over the past decade requires the collection of them, which is a challenging procedure. Firstly, we have to decide how we should classify an article's credibility. There is no straightforward way to address this issue, except for manual checking, which is not possible due to the amount of data. Another crucial challenge is the scraping of articles from 2009 until 2019. We need instances from a variety of different domains over the years. Hopefully, there is WebArchive[6], which is a digital archive of the World Wide Web and contains more than 452 billion pages from 2001. Through WayBackMachine, which is a tool of, WebArchive, we can visit and scrape these pages. However, some pages do not have instances from 2009, or they do not have samples

[6] https://web.archive.org/

for each day. Furthermore, the response time of the WayBackMachine is not satisfactory. Thus we have to be very efficient in order to collect the number of articles we want from the past decade.

Having a comprehensive volume of data, leading us to new challenges such as the storage and visualization arrangement. The nature of data does not indicate the use of traditional databases. We do not need any of the ACID properties of such a database, and besides, the amount of data requires a more flexible schema. Regarding the visualization of our data, we need a tool that will handle the amount of data and will be capable of presenting a variety of data aspects.

The last challenge we will describe is the most crucial one, especially for a programmer. As we explain, our goal is to study the evolution of fraud and real journalists over the years. Despite the visual representation of features, which can provide plenty of conclusions and help us understand where to focus, we need mathematical proves. For example, even if a graph shows that fake stories are using more capital letters than real stories during the years, we need to prove that there is a statistically significant difference between the values of fake and real. To do so, we have to identify the distribution of data and also clarify the possible types of noise that the data have. Then we have to apply the correct test to ensure the validity of our results.

1.3 Contributions

In this work, our ultimate goal is to provide the essentials for analyzing the evolution of linguistic characteristics over the years to fight the “infodemic”, namely Fake News. To do so, we had to gather news pieces over a wide period from both trusted and untrusted sources. Therefore, we implemented a highly-parallel and scalable crawler, which gathered articles from the Web Archive, covering ten years. We manage to collect, for now, more than 500K articles from 2009 until 2019, building a comprehensive dataset suitable for various applications. Then, based on the author's previous work Paschalides et al. 2019, we extract more than 500 linguistic characteristics from the article's title and content. We divided these features into three major classes: Stylistic, Complexity, and Psychological features. The last step we made was the indexing of our data and the utilization of a visualization platform to execute our initial analysis. Based on the outcomes, we divide the feature into three categories: i) Features alignment over time, ii) Features variable over time, and iii) Features evolved over time. We provide a visual representation of the results, along with possible explanations about each category. To sum up, our contributions are as follows:

- The implementation of a highly-parallel configurable web crawler, which is capable of collecting a vast amount of articles from Web Archive.
- The construction of a comprehensive dataset, containing more than 500K articles, covering a period from 2009 to 2019.
- The indexing and visualization of a plethora of different linguistic characteristics, along with the analysis of their evolution over time.

1.4 Outline Contents

The organization of our contents is as follows.

Chapter 1: Introduction

The introduction defines the motivation of our research and making clear how significant is the addressing of the Fake News problem. Moreover, explains the contribution of our study and the challenges we faced during our research.

Chapter 2: Related Work

Chapter 2 focus on an analytical review of the literature, which consists of work related to fake news classification. More specifically we examine previous studies on fake news detection using natural language processing characteristics, along with the available datasets.

Chapter 3: Methodology

The methodology chapter defines our methodology and explains each step we made very precisely. We described how we collected and constructed our datasets and the role of it in our study. Moreover, we analyze the feature extraction process and how we divide our features into three main different categories.. Then we present how we indexed and visualized our features to get a further understanding of them and study their evolution.

Chapter 4: Dataset

Chapter 4 focuses on presenting our constructed dataset, containing statistical information about the data. Moreover, in this chapter, we explained, in detail, each column of our dataset, and we also provide potential applications that can utilize it.

Chapter 5: Results

Chapter 5 focuses on presenting the outcomes after the visualization of the extracted features. We explain how we divided the features into three categories depending on the correlations of fake to real values, and we provide possible explanations for these effects.

Chapter 6: Conclusion

Chapter 6 defines our conclusions thoughts and summarize them to provide a brief outcome. Finally, we propose possible meaningful future works that will extend our work and help the research community to the mitigation of the virus, called Fake News.

Chapter 2

Related work

Contents

2.1 Datasets

Understanding the evolution of fake and real news through time can play a significant role in the development of thorough machine learning models. Though there exist several datasets for the identification of misinformation, most of them focusing on an inadequate period of time and covering a specific subject like politics, gossips, and tweets. To help researchers understand the roots of misinformation and create more generalize models, we provide a dataset that includes articles from 2009 until 2019 from a variety of different domains. For a better comparison, we perform a comprehensive analysis of well-known datasets used in the identification of misinformation, and we summarize the information in Table 2.1.

LAIR (W. Y. Wang 2017) dataset contains 12.4K short statements collected from

Dataset	# Rows	Collection Period	Classification
LIAR (W. Y. Wang 2017)	12.8K	2007-2016	Pants on fire, False, Barely true, Half true, Mostly true, True
FakevsSatire (Golbeck et al. 2018)	486	JAN/2016-OCT/2017	Fake, Satire
BuzzfeedPolitical (Horne and Adali 2017)	120	JAN/2016 - OCT/2016	Fake, True
NewsTrustData (Mukherjee and Weikum 2015)	82K	2006-2014	Qualitative scores
Buzzfeed (Shu, Mahudeswaran, et al. 2018)	182	-	Fake, True
PolitiFact (Shu, Mahudeswaran, et al. 2018)	488 (ongoing)	-	Fake, True
GossipCop (Shu, Mahudeswaran, et al. 2018)	3570 (ongoing)	-	Fake, True
FakeNews Corpus	9.4M	2016 - 2018	Fake News, Satire, Extreme Bias, Conspiracy Theory, Junk Science, Hate News, Clickbait, Proceed with caution, Political, Credible
FacebookHoax (Tacchini et al. 2017)	15.5K	JUL-DEC/2016	Scientific, Conspiracy

Table 2.1: List of existing datasets

PolitiFact, covering nine years from 2007 to 2016. Each statement is manually labeled by PolitiFact's experts into six fine-grained categories. The distribution of data along the six classes is relatively balanced, except for the pants-on-fire class, which contains approximately half of the data that each of the other categories has.

FakevsSatire (Golbeck et al. 2018) dataset contains 486 covering almost two years from January 2016 to October 2017. All of the articles are related to American politics and particularly to the US 2016 elections. The researchers annotate the content into two categories, Fake and Satire. In contrast with the LAIR dataset, FakevsSatire provides the whole text of the article along with the title and some extra details.

BuzzfeedPolitical (Horne and Adali 2017) dataset constructed by Horne et. al using Buzzfeed's 2016 article on fake election news on Facebook (Silverman 2016). They filter 120 articles during 9 months before the election's period. The dataset also contains the title and the body of the article for analyzing the journalists' writing style.

NewsTrustData (Mukherjee and Weikum 2015) dataset contains more than 82K articles from 2006 until 2014. NewsTrustData focuses on the article's credibility. They create a community, and each member manually answers 15 different questions about the article, like how factual is? , Is it fair? , Is the writing style proper, Is it credible, etc. Each member of the community has a different weight in the aggregation based on expertise, previous knowledge, and other characteristics.

BuzzFeed, **PolitiFact**, and **Gossipcop** (Shu, Mahudeswaran, et al. 2018) datasets are part of the FakeNewsNet project. BuzzFeed contains 182 comprises a complete sample of news that was published on Facebook, originating from9 news outlets over the period of a week during the 2016 U.S.elections. Each Facebook post is attached with a news article that was fact-checked by 5 BuzzFeed journalist. PolitiFact contains a set of 240 articles labeled by PolitiFact journalists as fake or real were collected, along with a scraped version of the analogous news articles. GossipCop dataset is related to gossips and contains 3570 stories which 19% of them are fake. Both PolitiFact and GossipCop are ongoing projects with new entries every couple of months.

FakeNewsCorpus is the largest of the listed datasets with over 9 million articles collected from a variety of different domains. The dataset is divided into 11 classes including, Fake News, Satire, Extreme Bias, Conspiracy Theory, and Credible news. The articles are characterized by their domain based on a list from Opensources. The dataset includes articles from 2016 until 2018.

FacebookHoax (Tacchini et al. 2017) dataset contains 15.5K Facebook posts coming from 32 pages (14 conspiracy and 18 scientific). The posts were collected from the second semester of 2016. Each post comes from a conspiracy page consider fake, and each post from a scientific page true.

All of the previously mentioned datasets, except FakeNewsCorpus, contain a relatively small number of articles for analyzing the behavior of fraud and real journalists over the years. FakeNewsCorpus, which contains a suitable number of articles, covers only 3 years of data.

2.2 Linguistic Characteristics on Fake News detection

Several researchers underestimate the analysis of solely the article's content for fake news detection and move to multi-modal approaches. For example, Shu et al. (Shu, S. Wang, and H. Liu 2019) explain the inability of news content to distinguish the difference between real and fake due to the fact that some fake stories are written intentionally to mimic real ones. The assumption is reasonable. However, a variety of other works identify significant deviations between trusted and untrusted articles (Potthast et al. 2017; Horne and Adali 2017; Paschalides et al. 2019).

To analyze how linguistic features evolve over time and verify the assumption, we focus on the author's previous work (Paschalides et al. 2019), with features produced from the headline and body of an article and categorizes them into three broad groups: stylistic, complexity, and psychological.

Stylistic Features focus on NLP methods to understand the syntax and textual style of each article's body and headline. Such features include the frequency of stop-words, punctuation, quotes, negations, and words that appear in all capital letters, whereas syntactical features include the frequency of Part-of-Speech (POS) tags in the text. Horne and Adali (Horne and Adali 2017) claim that the use of proper nouns and capital words in titles significantly differentiates fake from real news. (Pérez-Rosas et al. 2017) calculate the number of periods, commas, dashes, question marks, and exclamation marks in the text, and they achieve 71% accuracy.

Complexity Features require extensive NLP analysis, aiming to capture the overall intricacy of an article or a headline, and can be computed based on several word-level metrics that include readability indexes and vocabulary richness. Examples of readability indices are the *Gunning Fog*, *SMOG Grade*, and *Flesh-Kincaid* grade level. Each measure computes a grade-level reading score based on the number of complex words (e.g., over 3 syllables). A higher index means a document takes a higher education level to read. Methods to capture the vocabulary richness of the content include the *hapax legomenon* and *dislegomenon*, which correspond to a phrase that occurs only once and twice within a context. Additionally, methods such as *Yule's K* measure (Miranda-García and Calle-Martín 2005), *Sichel's S* measure, and *Honore's R* measure capture the complexity of the text. Such features have been used by Zollo et al . (Zollo et al. 2017) and Perez-Rosas et al. (Pérez-Rosas et al. 2017) achieving the higher classification accuracy in contrast to other feature combinations.

Psychological Features are based on frequencies of words defined in expert dictionaries, associated with sentiment and different psychological processes. A known example of such a dictionary is the *Linguistic Inquiry and Word Count (LIWC)* dictionary (Tausczik and Pennebaker 2010) that is used to measure various psychological processes and is also widely used on fake news detection (Horne and Adali 2017; Potthast et al. 2017; Shu,

S. Wang, and H. Liu 2019). There are also different sentiment and opinion lexicons that capture the writer's mindset along with any subjective claims (Loughran and Mcdonald 2011; Qiu et al. 2009). Zollo et al. (Zollo et al. 2017), and Horne and Adali (Horne and Adali 2017) use these dictionaries to show that fake content is considerably more negative, in general, than trusted news.

Chapter 3

Methodology

Contents

3.1 Methodology Overview

In this section, we will describe, in detail, our methodology for collecting and analyzing fake and real articles from 2009 until 2019. It consists of three major components: *i) a data collection component*, *ii) a feature extraction component*, and *iii) a data visualization & analysis component*. An overview of our architecture is depicted in Figure 3.1. Firstly, we need to collect the articles from the past decade by crawling the WebArchive[1] and

[1] http://web.archive.org/

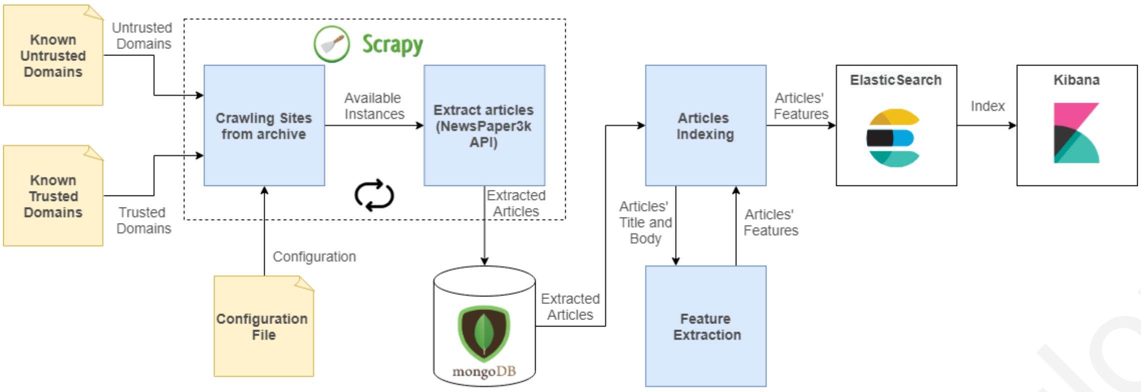


Figure 3.1: An overview of our methodology

divide the data into trusted and untrusted using two domain's credibility lists. Secondly, we have to extract a plethora of different linguistic features to gather all the available knowledge of the text. Finally, due to the amount of our dataset, we need an efficient way for indexing and data visualization. In the following sections, we describe each component thoroughly.

3.2 Data collection

In this section, we will describe the Data Collection component, which is responsible for gathering a meaningful dataset by collecting articles from various domains during the 2009-2019 period. To thoroughly explain this component, we divide it into three subcomponents: a) Domain Lists & Configuration File, b) Crawler, and c) Article's Extraction.

3.2.1 Domain Lists & Configuration File

In order to classify the veracity of an article, we adopted the BS Detector technique of having lists defining the credibility of a domain. In our case, we have two lists, namely Known Untrusted Domains, and Known Trusted Domains. The former contains domain names that usually publish Fake News and are highly scrutinized by fact-checking organizations, including Snopes, PolitiFact, and others. The latter includes high reputation domains, which rarely or never been criticized by fact-checking sites. In our crawling system, we read domains from these lists in a round-robin fashion and declare every article depending on the domain's credibility.

An important observation showing the validity of the untrusted domain list is that the majority of untrusted domains are no longer exist due to many reasons, such as bans, low visibility, and loss of purpose. For example, 70news.wordpress.com, the day after the US2016 elections, announced that Trump won the public vote by 700,000 votes. It became Google's first result for the public vote. The news was fake, and a few months later, the domain was banned. Hopefully, due to WebArchive[2] , we can crawl pages that no longer exist.

The configuration file allows users to modify the execution of the crawler slightly. More specifically, the user can edit the configuration file to change the crawling period. The user can set a start and an end date in the following format: %Y%m%d, where %Y is the full year (e.g., 2012), %m is the month (e.g., 11), and %d is the day (e.g., 07). Month and day are optional parameters. Furthermore, the user can provide a list targeting specific dates in a period. For example, the user can set the start date = "2009", end date = "2019" and give a list with specific dates within this period. The crawler will return articles only for the particular dates the user provides. Moreover, the user can provide different domain lists and also specify the least amount of article's body words for storing the article.

3.2.2 Crawler

The crawler is the heart of the Data Collection component. This subcomponent is responsible for the crawling of the specific domains in WebArchive. As we explain in Section

[2] https://web.archive.org/

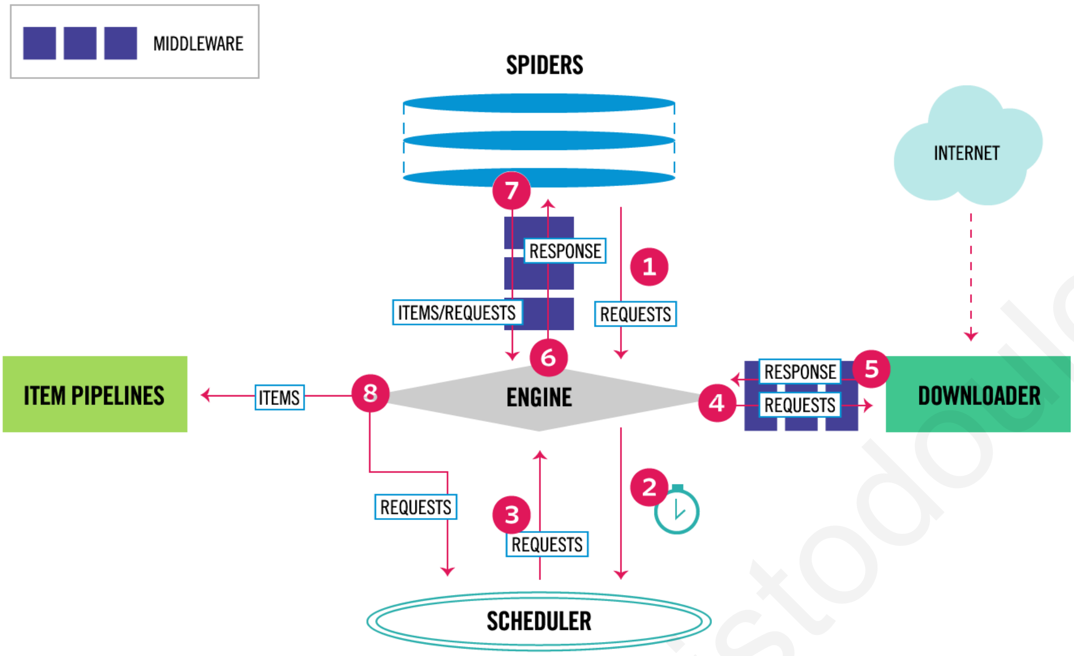


Figure 3.2: Crawler's Architecture and Data Flow

1.3, crawling from WebArchive.com is a challenging and time-consuming process. To be as effective, we can, we chose to utilize the Scrapy[3] library of Python.

Crawler's Architecture

Scrapy considers as one of the fastest web-crawling tools, especially for complex crawling and scrapping application. The architecture allows it to perform the fastest possible crawling and scraping while maintaining CPU and memory utilization as low as possible. For the needs of the thesis, we will dig inside the architecture of Scrapy to fully understand how it works.

Figure 3 illustrates the major components of Scrapy and presents the typical dataflow. Firstly, we will describe each component, and then we will see the dataflow.

Spiders are the main workhorses of Scrapy. More specifically, Spiders are custom classes,

[3] https://scrapy.org/

written by Scrapy users, containing the logic of the crawler. They are responsible for sending and handling requests.

Engine is responsible for controlling the data flow between all components of Scrapy and triggering events when a specific action occurs.

Scheduler is responsible for orchestrating the parallel nature of the system. Scrapy follows the asynchronous programming model for efficiency, which means that Spiders does not wait for the request's responses to move to the next one. The Scheduler keeps queues for request handling based on the engine commands.

Downloader solely performs web pages fetching.

Item Pipeline is responsible for handling the extracted items. Scrapy users can perform several tasks such as cleaning and storing the items.

The workflow presented in Figure 3 can be explained as follows:

1. Spiders (one or many) send requests to the engine.
2. The engine sends the requests to the scheduler's queues.
3. Depending on the system's availability, the engine asks the scheduler to dispatched a request.
4. The engine sends the requests to the downloader for fetching.
5. The downloader feeds the engine with a response.
6. The engine returns the response to the spider for further actions.
7. Depending on application logic, a spider can request the engine to pass the item to the item pipeline.
8. Item pipeline receives the items and performs any additional processing.

The whole process is iterative until no other requests are left.

3.2.3 Article's Extraction

Now that we have a full picture of how domain lists, configuration file, and Scrapy works, we will explain our crawler logic step-by-step.

1. At first, we are loading the domain lists, and we are setting crawler depending on the configuration file.

2. Then Scrapy spiders start sending requests to WebArchive API. The requesting URL is: $'http://web.archive.org/cdx/search/cdx?url = site\&from = start\_date\&to = end\_date\&filter = statuscode : 200'$.

3. WebArchive response includes only valid URL(home page url) snapshots from the domain during start_date and end_date.

4. Taking these URLs, we are making new crawling requests. Then we are extracting all the links included on the URLs' home page.

5. The extracted links are parsed as request elements to a new spider.

6. Each link is scraped and parsed as input to the NewPaper3K[4] to extract the article JSON object presented in Figure 3.3.

7. Finally, if the number of words of the article's body is greater than the words variable in the configuration file, we will store the article object into a MongoDB[5].

[4] https://newspaper.readthedocs.io/en/latest/

[5] https://www.mongodb.com/

_id: 5fbfa5a99e49313b0c91b05d

url: https://web.archive.org/web/20160505231528/http://order-or-er.com/2016/05/02/another-labour-politician-wants-to-relocate-jews-to-america/

Author: None

Content: This video shows Labour Director of Strategy and Communications Seumas Milne praising Hamas' "spirit of resistance" and saying "they will not be broken" , to huge cheers...

Domain: order-order.com

Fake: True

Publish_date: 2016-05-06T00:00:00.000+00:00

Timestamp: 20160506

Title: Labour Councillor: Israel Behind ISIS, "Zionist Jews are a Disgrace to Humanity"

Year: 2016

(a) Fake Article

Figure 3.3: A random example of an article's object

3.3 Feature Extraction: Linguistic Features for Fake News Detection

We compute different linguistic features that can be found in the headline and the body of articles, to extract discriminating characteristics for the detection of fake news. We group these features into *stylistic*, *complexity* and *psychological*, as presented in Section 2.1.

For the calculation and extraction of stylistic features, including *word frequencies, punctuation, quote, all-capital words* etc., we use Python Natural Language Toolkit (NLTK) [6]. For the complexity features, we compute the *Gunning Fog, SMOG Grade, and Flesh-Kincaid*

[6] `https://www.nltk.org/api/nltk.tag.html`

Examples of the Features Extracted		
Stylistic	**Complexity**	**Psychological**
Number of ”I” pronouns	Gunning_fog	Number of analytical words
Number of all capital letters	SMOG Grade	Number of negations
Number of stop words	Flesh-Kincaid	Number of slang words
Number of Verbs	Yules_k	Number of power words
Number of quotes (”)	Coleman Liau	Number of casual words
Number of adverbs	Dale Chall	Number of emotion words
Number of ”We” pronouns	Brunets W	Number of risk words
Number of full stops (.)	Honores R	Number of certainty words
Number of words	Number of happax legomena	Number of power words
Number of lines	Number of happax dislegomena	Number of affiliation words

Table 3.1: A sample of the extracted features divided into the three categories

grade level readability indexes, hapax legomenon and dis legomenon, and complexity measures such as *Yule's K, Sichel's S, Honore's R* and *Type-Token Ratio (TTR).*

Regarding the psychological features, we use the LIWC dictionary to extract features related to the author's cognitive process, emotions, and attentional focus. To capture the opinion of the writer, we used dictionaries including the negative and positive opinion lexicon B. Liu, Hu, and Cheng 2005, and the moral foundation dictionary Graham, Haidt, and Nosek 2009. The sentiment score is computed via the AFINN sentiment lexicon Nielsen 2011, a list of English terms manually rated for valence.

3.4 Data Indexing & Visualization

Due to the amount of data and the nature of the problem, we need to examine them further. To do this, we utilize ElasticStack [7], which contains as a core the ElasticSearch engine and also allows us to visualize the indexed data through Kibana. Elasticsearch is a highly scalable open-source full-text search and analytics engine. We can store, search, and analyze vast volumes of data quickly and in near real-time. Kibana lets us visualize the Elasticsearch indexed data. It allows us to select how to give shape to the data with the classic charts including, histograms, line graphs, pie charts, sunbursts, bar charts. Moreover, we can also perform advanced time series analysis and find visual relationships in the data or any anomalies that may occur.

In this work, we created two indexes in order to have a better representation of the data. The first index, namely articles_in_time, contains the collected articles along with their information. More specifically, the index includes the title and the body of the article, the article's URL, the authors' lists, the domain, the label, the publish date, weather is available, the timestamp, and the article's title and full content. The purpose of this index is to give us the ability to search through the data and find possible anomalies, for example, articles that refer to the privacy policy of the page. Moreover, through this index, we can examine the domains' distribution and confirm the validity of the articles' URL. The second index, namely features_in_time, contains the 535 numerical features we produce, along with the article's timestamp and label. The second index facilities the examination of the linguistic characteristics. Using the rich interface of Kibana, we can visualize feature evolution through time and compare the two-classes (Fake VS True). Moreover, because of the timestamp, we can "zoom-in" into a specific period and examine the data. An extensive representation of the visualization results is presented in Chapter 5.

[7] https://www.elastic.co/

Chapter 4

Dataset

Contents

4.1 Dataset Overview

In this section, we provide an extensive description of the proposed dataset. Our dataset consists of half a million articles from 500 unique domains. From the whole set, 77% of them originating from untrusted sources, and the rest 23% from trusted. Dataset designed to capture daily data achieving, for now, 52% completeness. A detailed representation of data statistics is presented in Table 4.1. Moreover, Figure 4.1 depicts the top 20 domains, for each class, that we collect data. The majority of untrusted domains are Blogspot and WordPress websites, which usually indicates individual effort. We highly recommend visiting these pages to realize the validity of them.

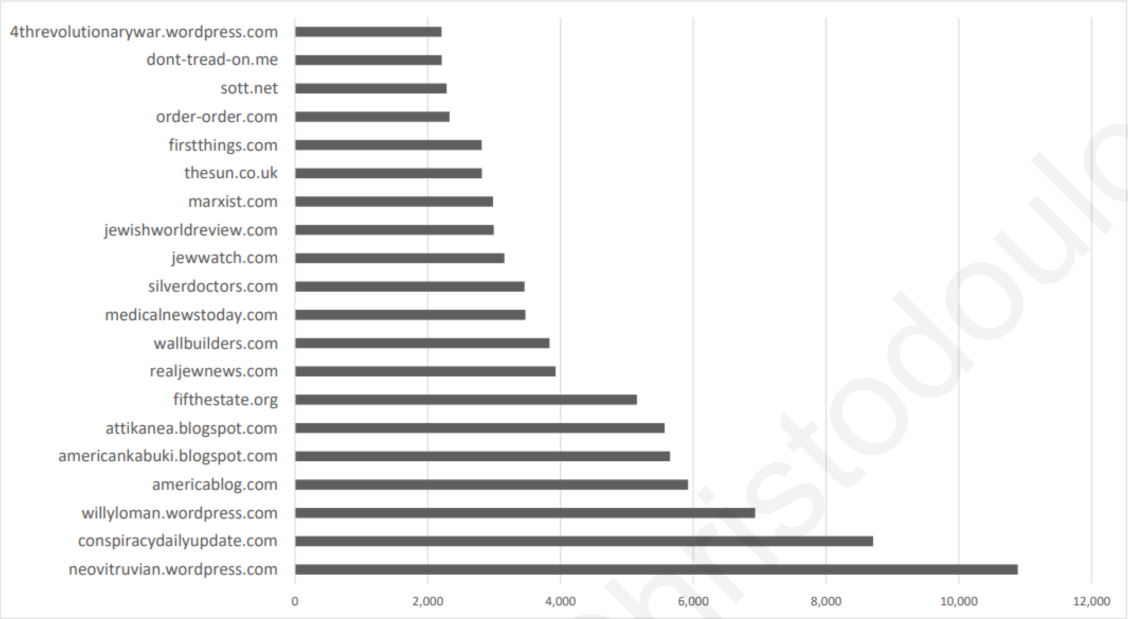
Year	Total	Fake	Real	Missing Days
2009	19,828	10,764	9,064	240
2010	31,455	19,401	12,054	246
2011	83,870	59,977	23,893	129
2012	65,841	50,736	15,105	79
2013	20,584	19,264	1,320	234
2014	40,951	28,012	12,939	213
2015	36,701	33,742	2,959	231
2016	64,251	51,180	13,071	58
2017	55,589	47,761	7,828	110
2018	35,766	32,348	3,418	172
2019	59,628	46,189	13,439	236
Total	**514,464**	**399,374**	**115,090**	**1,948**

Table 4.1: Dataset Overview

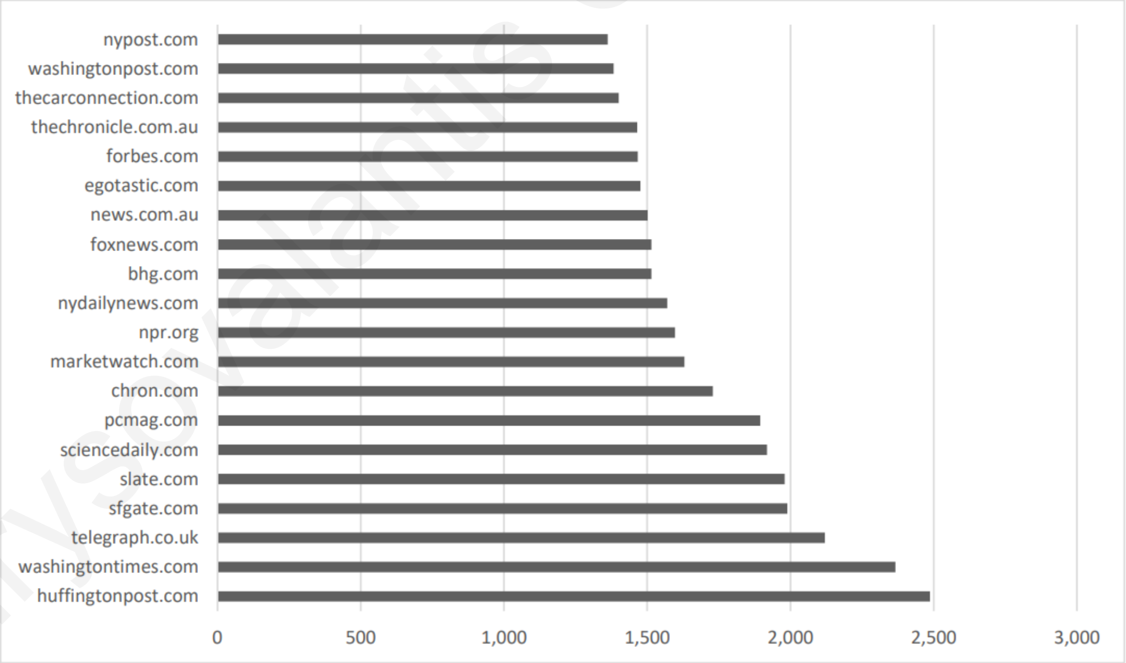
4.1.1 Dataset Columns

In this section, we will introduce the data format for each of the captured fields, along with a short description.

- _id (String): A unique 24-bit identifier for each document
- URL (String): The URL of the captured article. It's always a snapshot from Web Archive.
- Authors (List): A list contains the names of the authors. Usually, the list is empty, especially for untrusted sources.



(a) Top Untrusted Domains



(b) Top Trusted Domains

Figure 4.1: Top-20 Trusted VS Top-20 Untrusted domains

- Domain (String): The domain that the article is originated.
- Publish Date (Date): The article's publishing date. Usually, it's not available, especially for untrusted sources.
- Timestamp (Date): The date that Web Archive set as the article's publishing date. Please note that timestamp and publish date may differ because Web Archive takes as publish date the first day that the article was available on the site. The timestamp is always available, thus we recommended it as the field for time series analysis.
- Year (String): The published year of the article.
- Fake (Boolean): Whether the label is True, the article coming from an untrusted source, whether the label is False the article originating from a trusted source.
- Title (String): The article's title.
- Content (String): The content of the captured article. Please note that content may need some preprocessing in order to remove some unexpected newline (
n) characters.

4.2 Potential Applications

In this section, we will describe some applications that we believe would benefit from our dataset. Due to the amount of data and period, our dataset can be useful for analyzing fake news evolution, detect fake news, and expose specific events that trigger the way journalists write, either from trusted or untrusted sources.

4.2.1 Fake News Evolution

As we explain, our goal is to analyze the evolution of linguistic characteristics from trusted and untrusted sources. The dataset was created and implemented following this specific consideration. Our dataset contains more than 500k articles from 2009 until 2019, almost on a daily basis. Therefore, it's suitable for analyzing the evolution of these articles over the years and find potential similarities and differences between trusted and untrusted sources. First, using the article's title and body, you can extract a variety of different linguistic characteristics and analyze how they evolve. For example, instead of focusing on a large feature space, as we proposed, you can focus on the features((Chen, Conroy, and V. L. Rubin 2015; Agrawal 2017)) that show how related is the title and the content to identify how clickbait evolves. Second, using the article's URL you can visit and analyze the website that posted the news to identify how websites evolve over the years. For example, by utilizing our dataset you can examine how trusted sources evolved over the years and correspond these changes with untrusted sources.

4.2.2 Fake News Detection

One of the challenges of fake news detection using deep learning approaches is the lack of comprehensive datasets, that covers many years. Long Short-Term Memory (LSTM) is a tree-structured recurrent neural network used to analyze variable-length sequential data. Many approaches ((Ling et al. 2015; Sholar, Chopra, and Jain 2017; Long et al. 2017)) utilized recurrent neural networks for the identification of misinformation. However, this kind of neural network needs an extensive amount of data to perform well. Moreover, having an extended period of data increases their performance even further. Considering these arguments, our dataset can be considered an august fit for these approaches.

4.2.3 Event Detection

As we explained, we captured almost daily data from 2009 until 2019. A great work is to identify specific events over the years by analyzing significant change points in the features' behavior. For example, by investigating how the mean and standard deviation of one or multiple features change over time, you can identify possible real-world events that may affect the way journalists write. On the other side, you can reverse this process to analyze how specific real-world events affect the title, content, or the design of a piece of news.

Chapter 5

Results

Contents

5.1 Feature Analysis

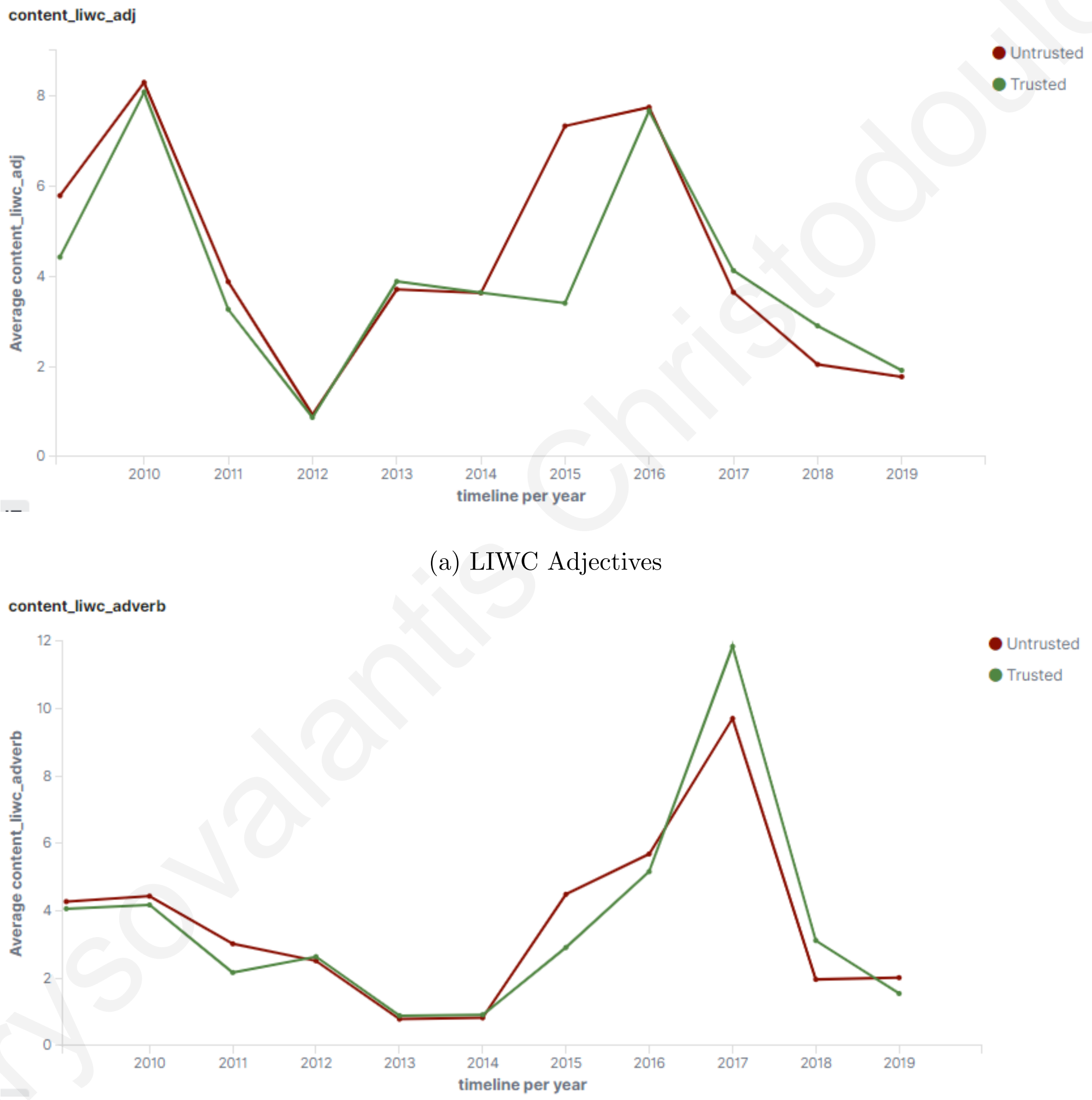
In this section, we will present the outcomes after the visualization of our comprehensive dataset. Each graph corresponds to a feature. In the figure, both lines representing the average value of the feature across the years. The red line corresponds to the average value of all the articles coming from untrusted sources, and the green one represents the average value of all the articles originating from trusted sources. Due to the nature of the results, we choose to divide them into three different groups: i) Features alignment in time, ii) Features variable in time, and iii) Features Fake to Real.

5.1.1 Feature alignment in time

During the visualization of the features, we observed many of them follow similar fluctuations, between trusted and untrusted sources, over the years. This outcome leads us to some conclusions.

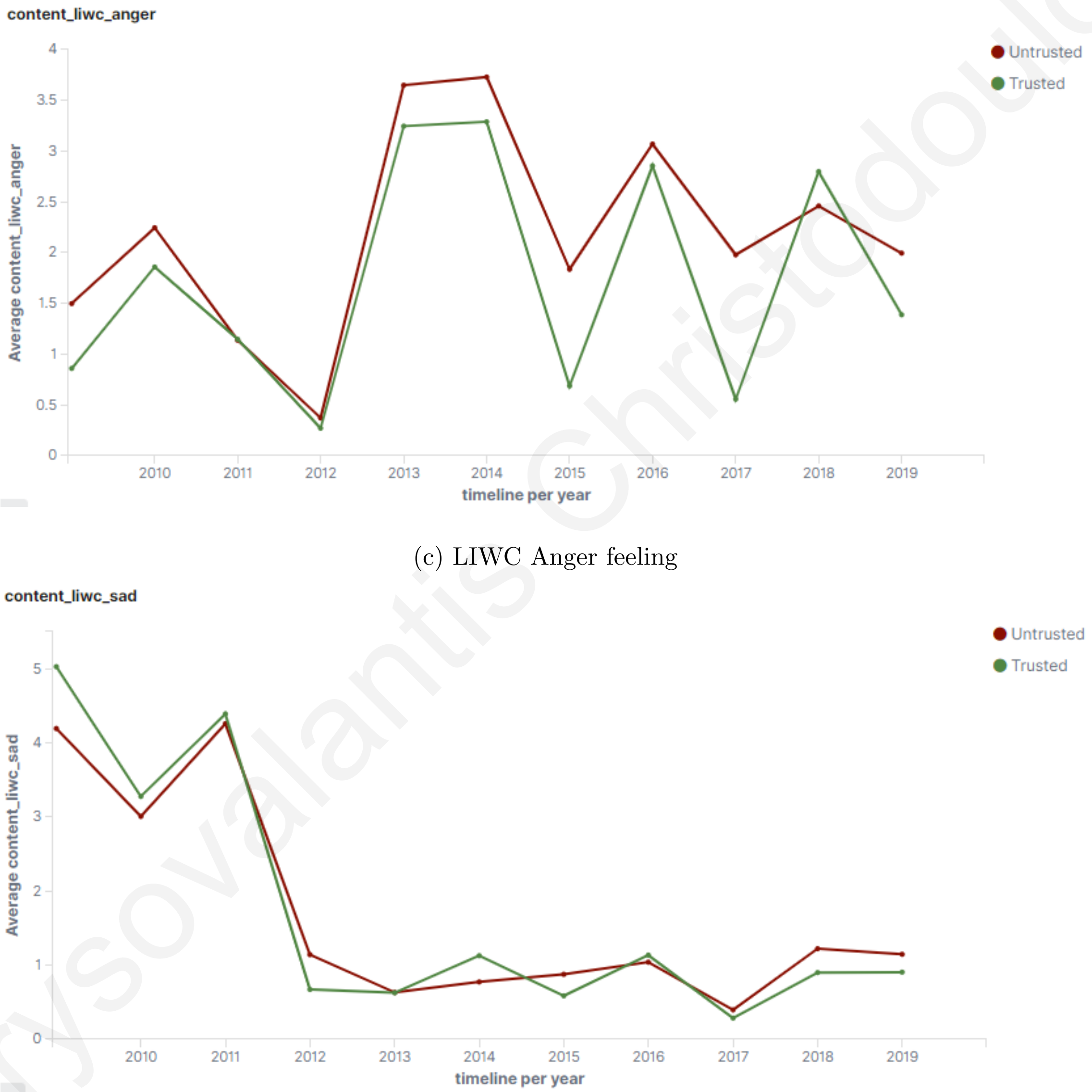
Firstly, those features are not capable for distinguishing between real and fake articles. To the best of our knowledge, none of the features belong to this category has ever been mentioned in the literature as helpful in the identification of misinformation. However, we intentionally chose characteristics coming from LIWC, which is a highly used dictionary in the literature (Potthast et al. 2017; V. Rubin et al. n.d.; Paschalides et al. 2019). Usually, it provides poor results and is utilized to provide a comparison baseline for more advanced approaches. As Figure 5.1 depicts, LIWC characteristics flow in the same way for both trusted and untrusted sources. Except for some small variations, the great majority of metrics show that the difference can be considered insignificant.

The second conclusion is related to the factors that may affect these similarities. As the figures show, despite the similarity between trusted and untrusted sites, there are significant fluctuations in the data, which leads us to the conclusion that there are factors that affect the behavior of both sites. Unfortunately, the examination of these factors is not related to the purpose of this research, however, further research can be considered as challenging and very interesting work.



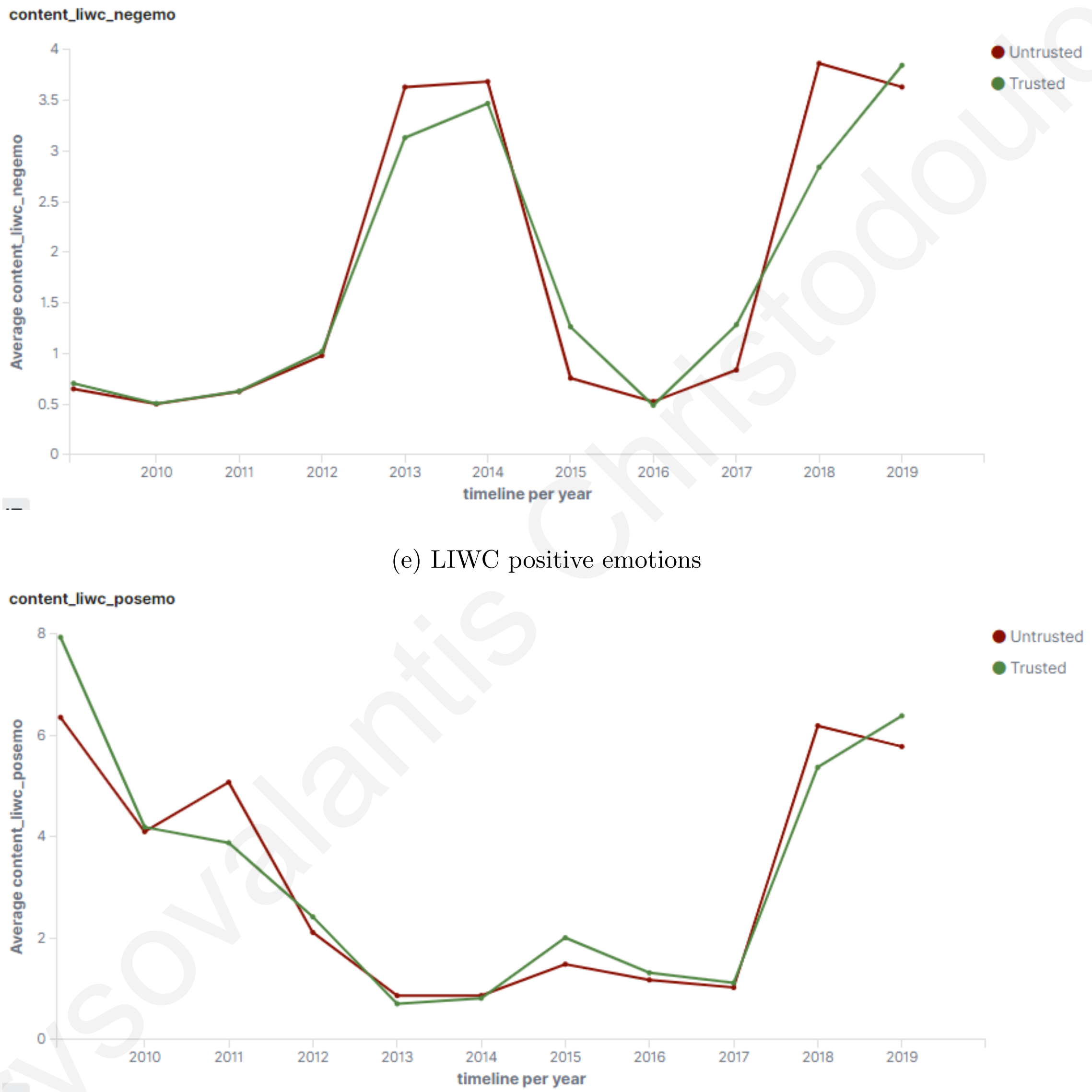
(a) LIWC Adjectives

(b) LIWC Adverbs



(c) LIWC Anger feeling

(d) LIWC Sad feeling



(e) LIWC positive emotions

(f) LIWC negative emotions

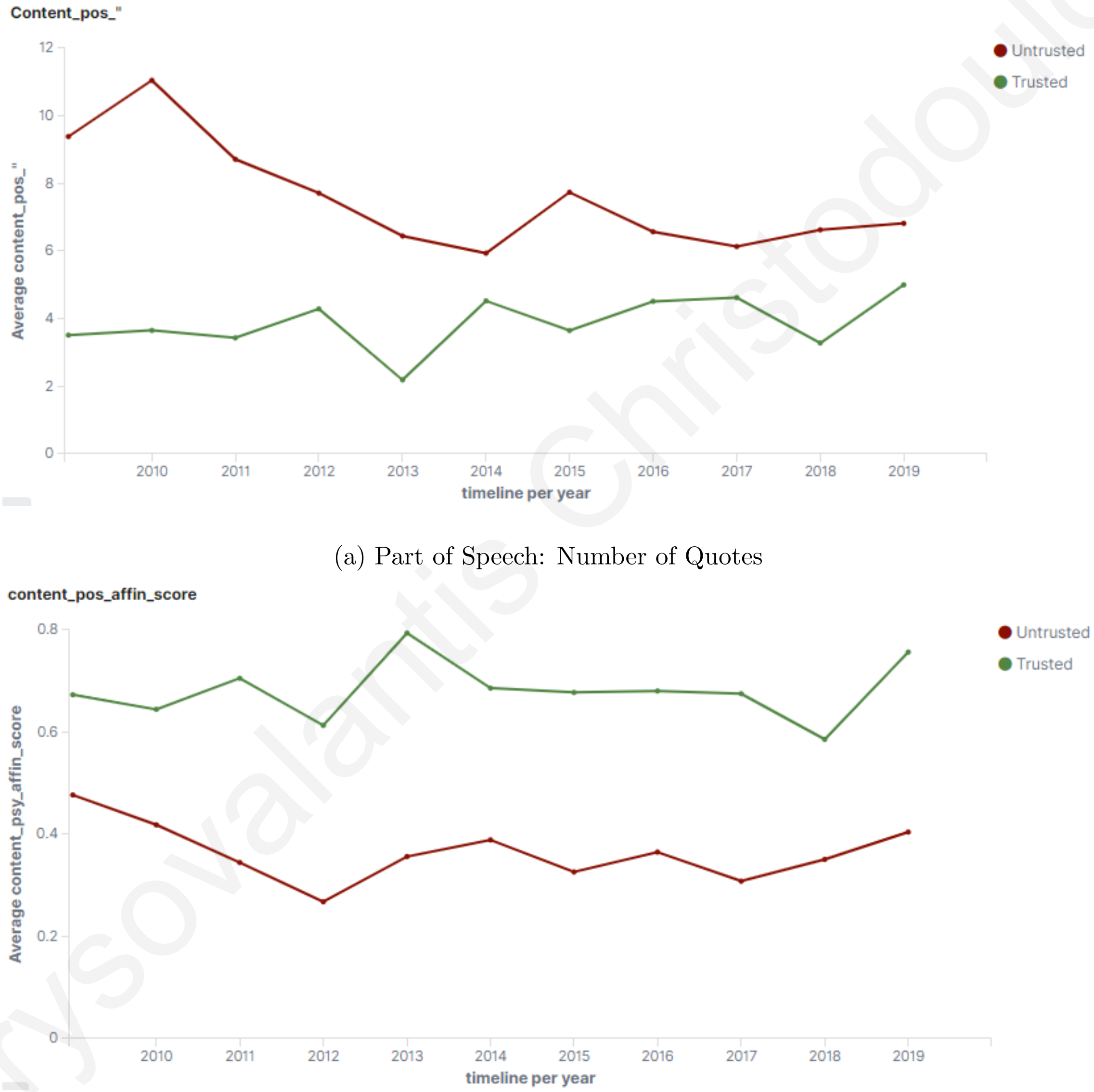
Figure 5.1: Examples of Features alignment in time

5.1.2 Feature variable in time

In the second category, we include features with a continued distinction between trusted and untrusted sources. As Figure 5.2 depicts, during the period of 10 years, the differences between the averages of the two classes steadily have the same sign. In a nutshell, there is no point in the graph that the two-line may intercept.

Generally, as the literature indicates, Fake News are mostly related to negative stories, except for the purposes which are related to gossips. (Horne and Adali 2017; Zollo et al. 2017). Graph 5.2d, 5.2e, 5.2f clearly show that stories coming from untrusted sources contain more negative emotions. In contrast with the LIWC dictionary, the rest of our emotion detection dictionaries agree with this fact. People are attracted much more intensely to a negative story than a positive one [1]. This argument is valid for both trusted and untrusted sources. Therefore, news agencies promote as main stories bad stories to get more attention. However, trusted sources usually cover a wider area of subjects, including stories that do not capture the users' attraction. As we explain in the previous section, most of the untrusted domains we captured are small Blogspot and WordPress sites, which are continuously delivered misinformation to attract users. As a consequence, there is no motivation or resources to publish a positive story that will attract much fewer users.

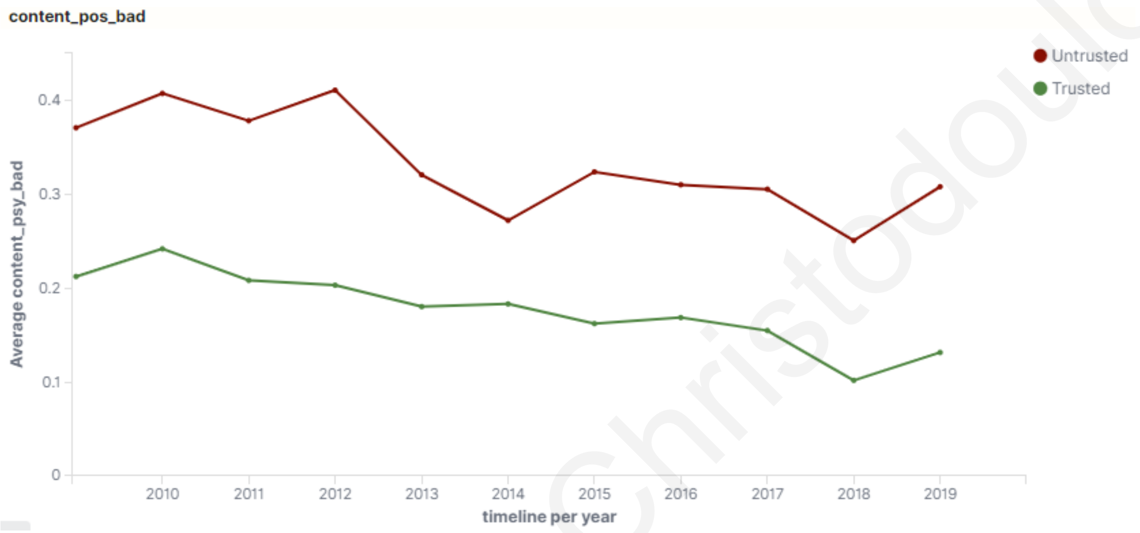
[1] https://www.bbc.com/future/article/20140728-why-is-all-the-news-bad



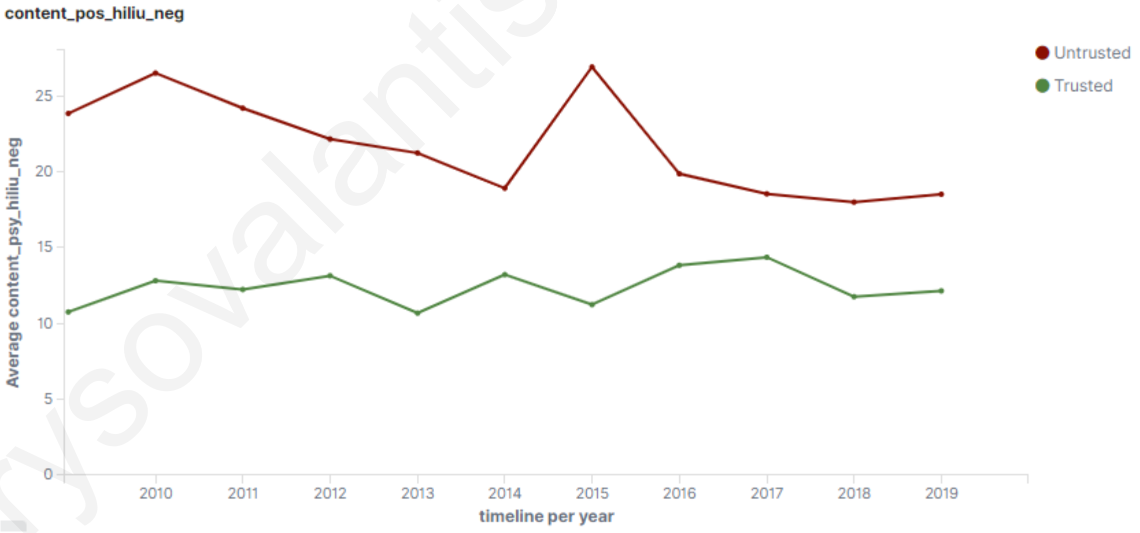
(a) Part of Speech: Number of Quotes

s

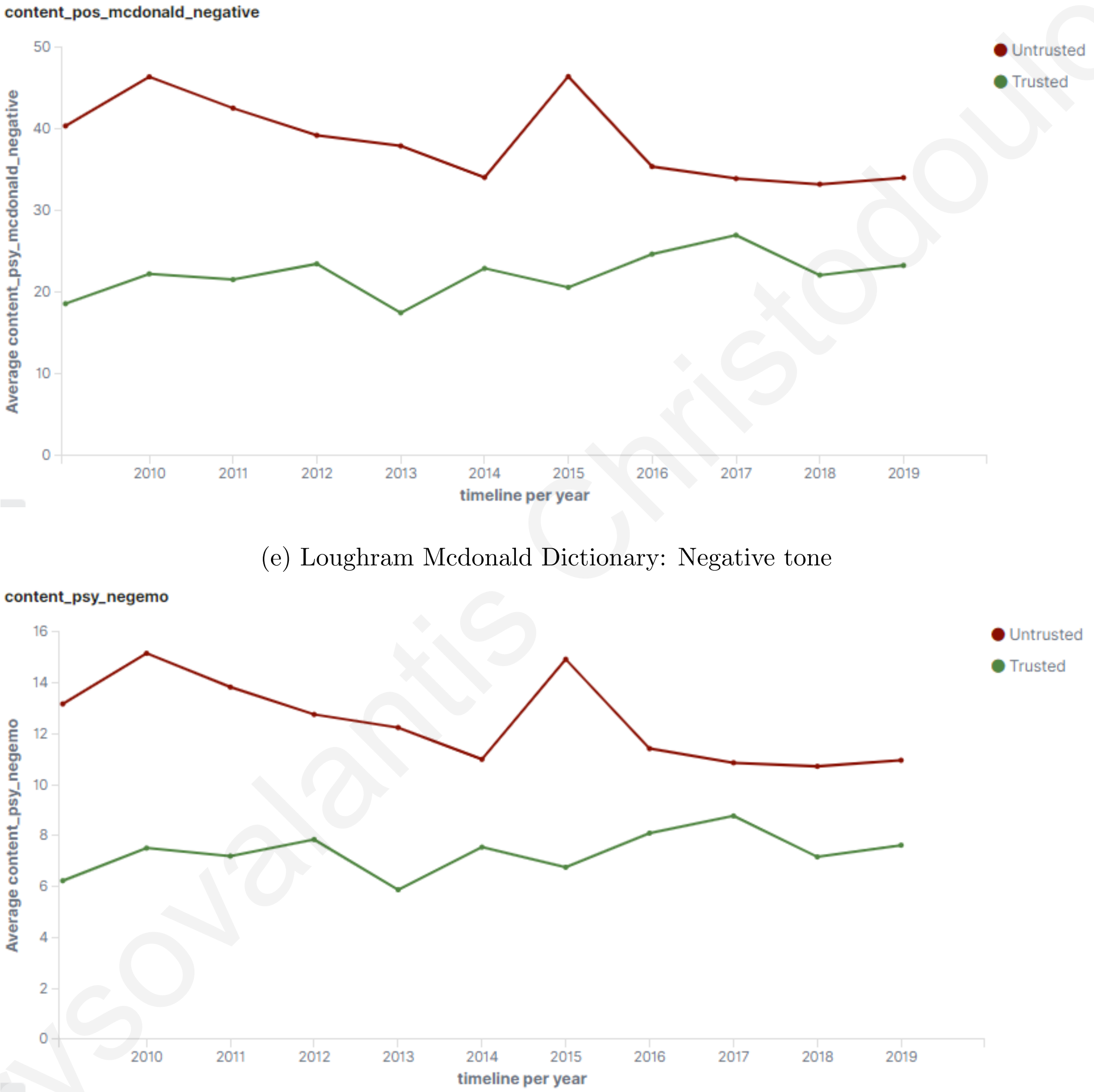
(b) AFFIN Sentiment Score



(c) Psychology: Bad emotions



(d) Hu Liu dictionary: Negative words



(e) Loughram Mcdonald Dictionary: Negative tone

(f) Psychology: Negative emotions

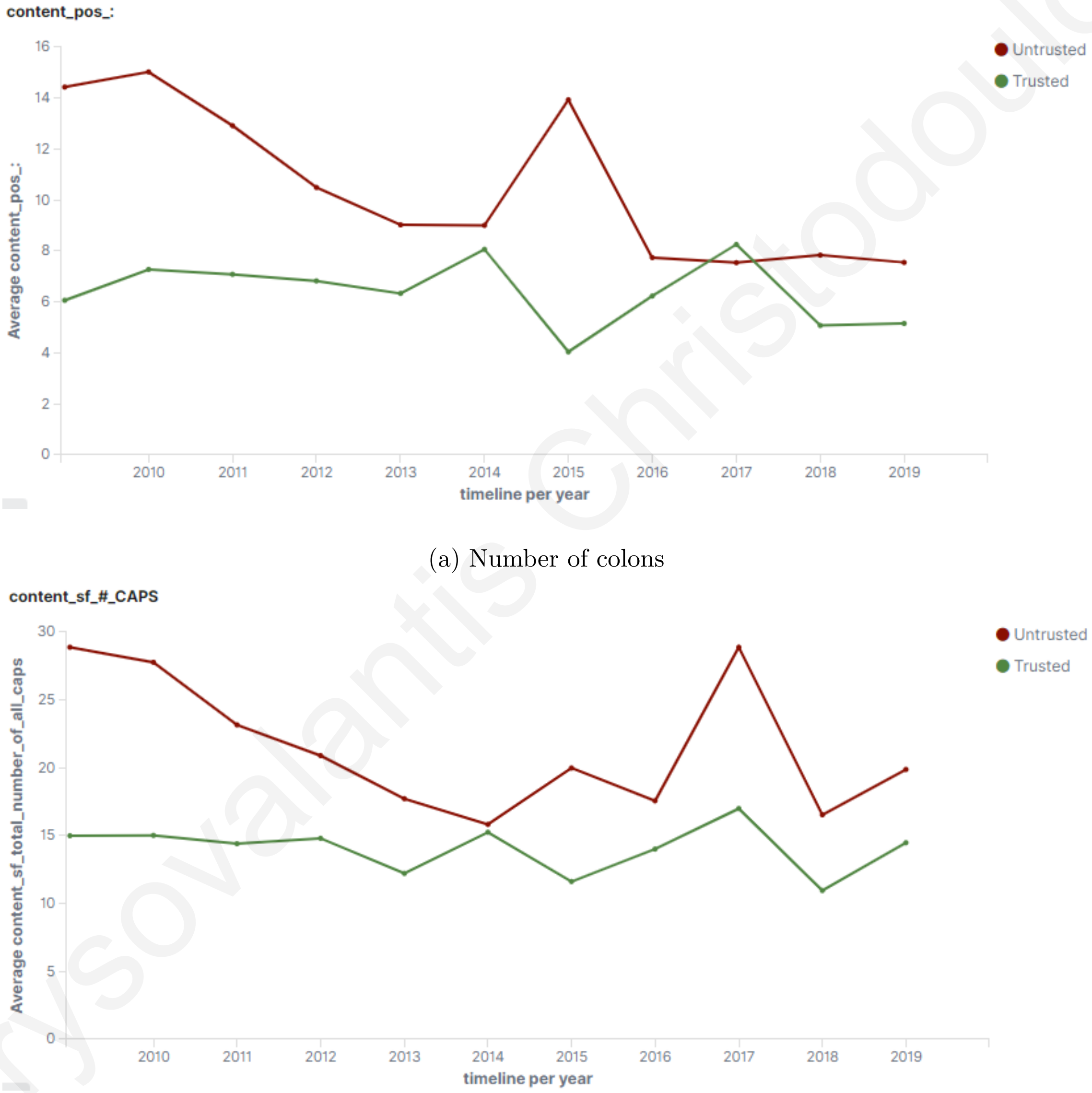
Figure 5.2: Examples of Features distinguish in time

5.1.3 Feature Fake to Real

One of the principal assumptions that motivate as to start this project was the fact that fake stories became more realistic as time passes. Fraud journalists evolved and trying to mimic the way credible ones write. The third category of features corresponds to this particular assumption.

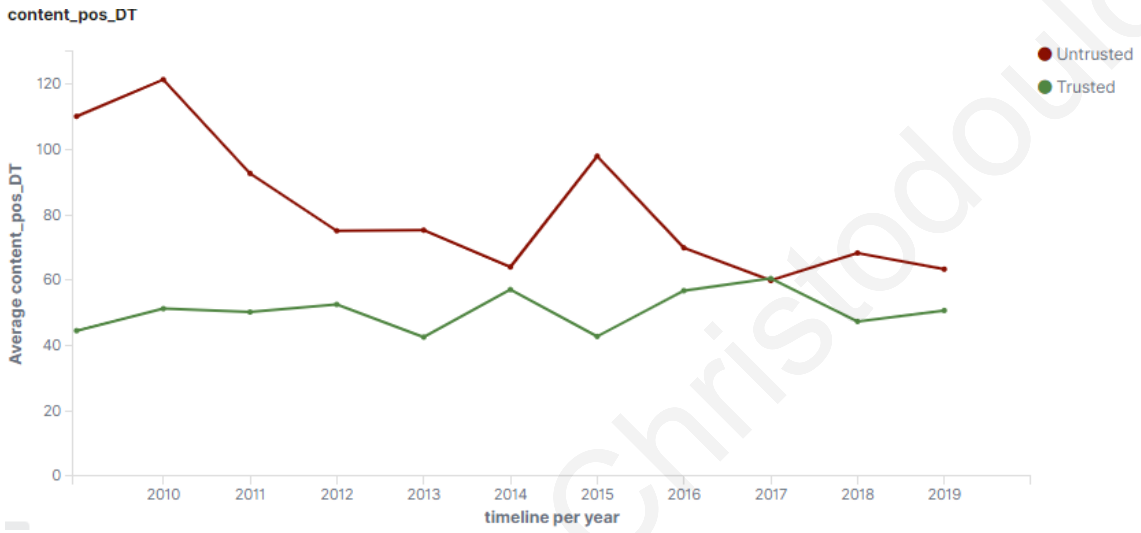
Both number of colons and capital words are presented in the bibliography as important features for Fake News detection ((Horne and Adali 2017; Yang et al. 2018)). As Figures 5.3a, 5.3b depict that indeed, the differences between the two classes look distinguishable, especially for the 2009 to 2012 period. However, as time passes, the distinction becoming less significant. The same occurs in every feature belong to this category.

Another notable observation coming from Figure 5.3 is the rhombus effect, which is presented between 2015 and 2016. During this period, there is an inversely proportional relationship between stories from trusted sources and stories from untrusted. The effect looks very promising and intriguing for research and may be related to the U.S 2016 elections. During the presidential campaign for the U.S 2016 elections , there was vast dissemination of misinformation. A possible explanation is that, due to the extensive polarization, ”amateur” writers disseminate Fake News to increase the political interests of their party. On the other hand, credible news agencies may tried to be as objective as possible to avoid criticism. This is just a possible explanation for the existence of this rhombus effect, but further investigation is needed to provide concrete conclusions.

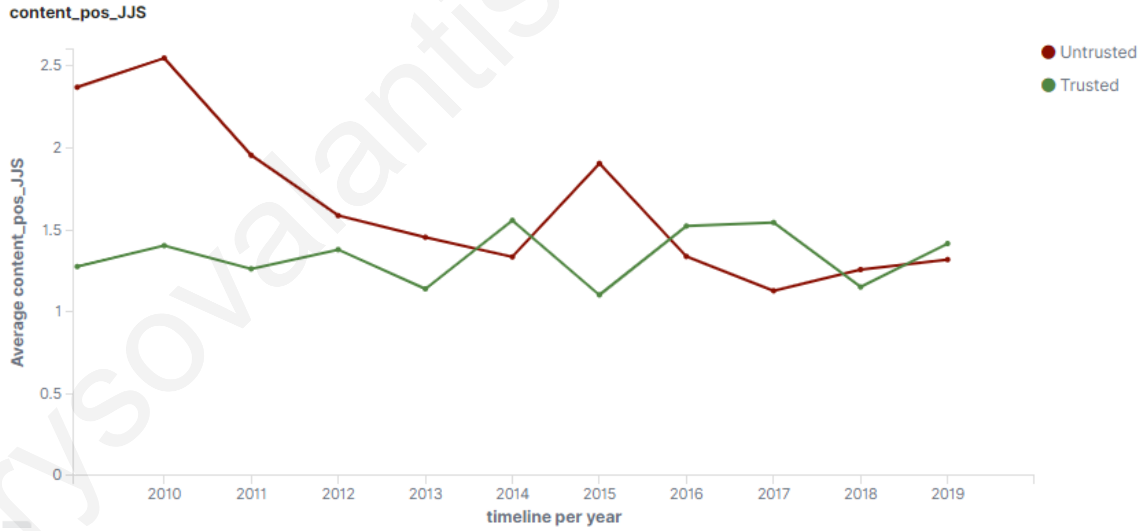


(a) Number of colons

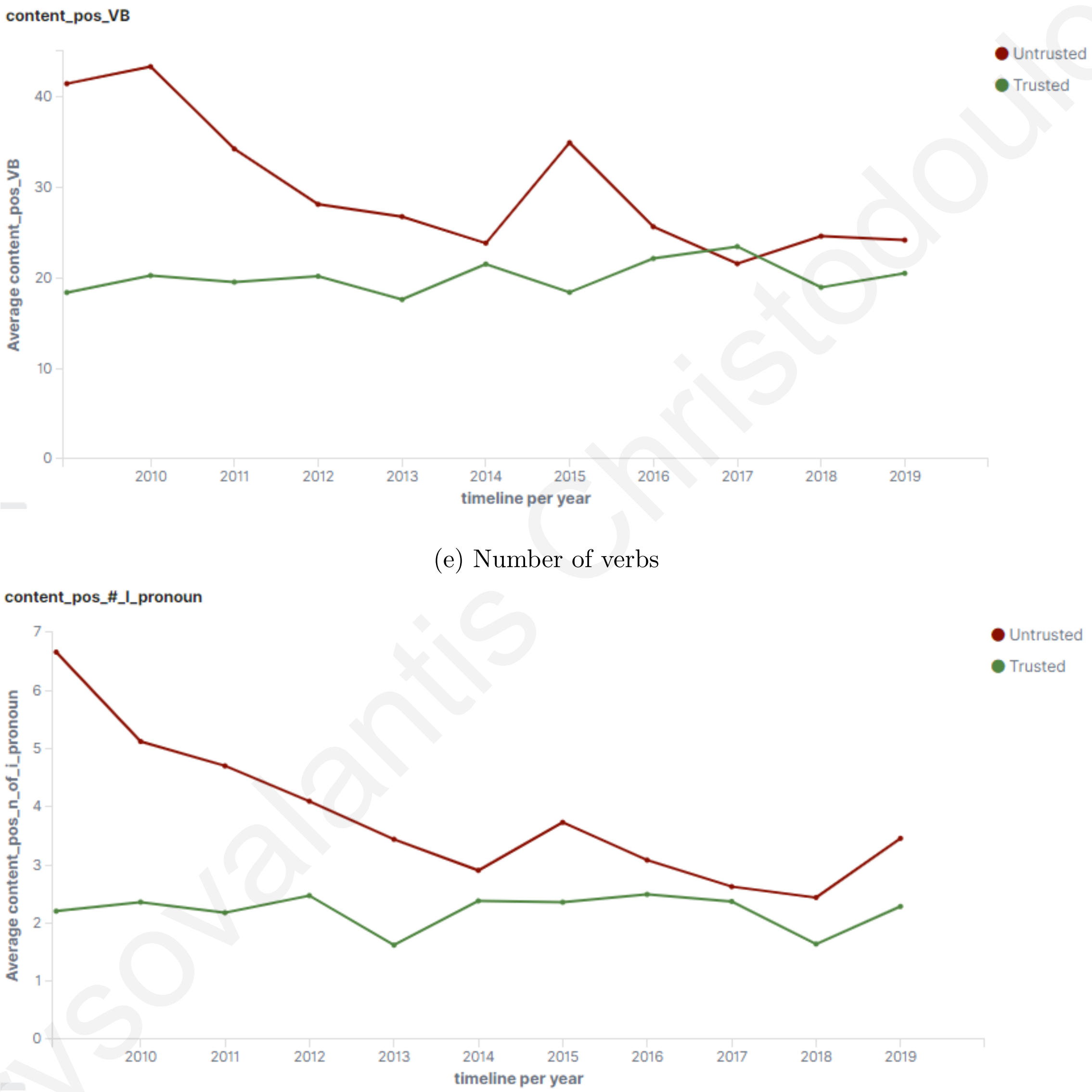
(b) Number of capital words



(c) Determinants



(d) Superlative Adjectives like "biggest"



(e) Number of verbs

(f) Number of I pronouns

Figure 5.3: Examples of Features becoming similar over time

Chapter 6

Conclusion

Contents

6.1 Conclusion

Taking everything into consideration, our goal was to provide the essentials for analyzing the evolution of linguistic characteristics over the years. We strongly believe that we achieve this objective, and we offer some valuable information that will help researchers fight the virus of Fake News. Through the proposed dataset, which contains 500K articles, for now, coming from trusted and untrusted sources and covering a decade of data, researchers can examine not only the evolution of linguistic characteristics but also utilized it for many other applications. Moreover, in this work, we manage to visualize and investigate the evolution of more than 500 linguistic features over time, and we divided them into three categories to provide a solid understanding. The first category refers to features that align for both trusted and untrusted sources. The second includes features that have a clear and continuous distinction during the whole period. Finally, the third category, which is the most interesting one, contains features from untrusted sources that evolved over time to approach values coming from credible sources.

6.2 Future Work

We took the assumption of declaring the article's veracity depending on the domain's credibility. Obviously, some pages may regularly publish fake news but also post articles describing a real story. Moreover, there is a possibility of a page, usually publishing misinformation to change ideas and transform into a credible site. Therefore, we have some false positive samples in our data, which may affect the outcomes. To address this drawback, we plan to implement a tool that will help us search into known fact-checking sites to verify if some of our stories manually label as fake. There different approaches to accomplish this goal. For example, we can extract the claim of the article, or we can utilize text similarity algorithms like SimHash to identify news that is manually labeled as fake. Then we will collect those verified stories and analyze the features to examine the differences and produce new outcomes. Moreover, in this work, we focused on the visual representation of the results to get a better understanding of them. However, we need to examine the data thoroughly and perform all the essential statistical analysis tests to ensure the validity of our results. As a first step, we can sample our data through time and produce the same graphs multiple times to study the deviation. Lastly, the application of change-point detection techniques can sufficiently add some light to the events that cause changes to the articles' linguistic characteristics. To mitigate misinformation, it's crucial to understand what kind of events affect these characteristics, and find possible patterns that would let us deeply realized how fake and real news evolve over time.

Bibliography

Paschalides, Demetris et al. (2019). “Check-It: A Plugin for Detecting and Reducing the Spread of Fake News and Misinformation on the Web”. In: arXiv: 1905.04260. URL: http://arxiv.org/abs/1905.04260.

Gupta, Aditi et al. (2014). “Tweetcred: Real-time credibility assessment of content on twitter”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8851. Springer Verlag, pp. 228–243. ISBN: 9783319137339. DOI: 10.1007/978-3-319-13734-6_16. arXiv: 1405.5490. URL: https://link.springer.com/chapter/10.1007/978-3-319-13734-6%7B%5C_%7D16.

Wang, William Yang (2017). “”Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection”. In: arXiv: 1705.00648. URL: http://arxiv.org/abs/1705.00648.

Golbeck, Jennifer et al. (2018). “Fake News vs Satire”. In: pp. 17–21. DOI: 10.1145/3201064.3201100.

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

Mukherjee, Subhabrata and Gerhard Weikum (2015). “Leveraging joint interactions for credibility analysis in news communities”. In: *International Conference on Information* *and Knowledge Management, Proceedings* 19-23-Oct-, pp. 353–362. DOI: 10.1145/2806416.2806537.

Shu, Kai, Deepak Mahudeswaran, et al. (2018). “FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media”. In: arXiv: 1809.01286. URL: http://arxiv.org/abs/1809.01286.

Tacchini, Eugenio et al. (2017). “Some like it hoax: Automated fake news detection in social networks”. In: *arXiv* December. arXiv: 1704.07506.

Shu, Kai, Suhang Wang, and Huan Liu (2019). “Beyond news contents: The role of social context for fake news detection”. In: *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining* December, pp. 312–320. DOI: 10.1145/3289600.3290994. arXiv: 1712.07709.

Potthast, Martin et al. (2017). “A Stylometric Inquiry into Hyperpartisan and Fake News”. In: arXiv: 1702.05638. URL: http://arxiv.org/abs/1702.05638.

Pérez-Rosas, Verónica et al. (2017). “Automatic Detection of Fake News”. In: arXiv: 1708.07104. URL: http://arxiv.org/abs/1708.07104.

Miranda-García, A. and J. Calle-Martín (2005). “Yule’s characteristic K revisited”. In: *Language Resources and Evaluation* 39.4, pp. 287–294. ISSN: 1574020X. DOI: 10.1007/s10579-005-8622-8.

Zollo, Fabiana et al. (2017). “Debunking in a world of tribes”. In: *PLoS ONE* 12.7, pp. 1–27. ISSN: 19326203. DOI: 10.1371/journal.pone.0181821. arXiv: 1510.04267.

Tausczik, Yla R. and James W. Pennebaker (2010). “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods”. In: *Journal of Language and Social Psychology* 29.1, pp. 24–54. ISSN: 0261-927X. DOI: 10.1177/0261927X09351676. URL: http://journals.sagepub.com/doi/10.1177/0261927X09351676.

Loughran, Tim and Bill Mcdonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *Journal of Finance* 66.1, pp. 35–65. ISSN: 00221082. DOI: 10.1111/j.1540-6261.2010.01625.x.

Qiu, Guang et al. (2009). "Expanding Domain Sentiment Lexicon through Double Propagation Zhejiang Key Laboratory of Service Robot Department of Computer Science College of Computer Science University of Illinois at Chicago". In: *Ijcai*, pp. 1199–1204.

Liu, Bing, Minqing Hu, and Junsheng Cheng (2005). "Opinion observer". In: *Proceedings of the 14th international conference on World Wide Web - WWW '05*. New York, New York, USA: Association for Computing Machinery (ACM), p. 342. DOI: 10.1145/1060745.1060797. URL: http://portal.acm.org/citation.cfm?doid=1060745.1060797.

Graham, Jesse, Jonathan Haidt, and Brian A Nosek (2009). "PERSONALITY PROCESSES AND INDIVIDUAL DIFFERENCES Liberals and Conservatives Rely on Different Sets of Moral Foundations". In: DOI: 10.1037/a0015141. URL: www.moralfoundations.org..

Nielsen, Finn Årup (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *CEUR Workshop Proceedings* 718, pp. 93–98. ISSN: 16130073. arXiv: 1103.2903.

Chen, Yimin, Niall J Conroy, and Victoria L Rubin (2015). "Misleading Online Content". In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection - WMDD '15*. New York, New York, USA: ACM Press, pp. 15–19. ISBN: 9781450339872. DOI: 10.1145/2823465.2823467. URL: http://dl.acm.org/citation.cfm?doid=2823465.2823467.

Agrawal, Amol (2017). "Clickbait detection using deep learning". In: *Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016* October, pp. 268–272. DOI: 10.1109/NGCT.2016.7877426.

Ling, Wang et al. (2015). “Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation”. In: September, pp. 1520–1530. arXiv: 1508.02096. URL: http://arxiv.org/abs/1508.02096.

Sholar, John Merriman, Shahil Chopra, and Saachi Jain (2017). “Towards Automatic Identification of Fake News : Headline-Article Stance Detection with LSTM Attention Models”. In: 1, pp. 1–15.

Long, Yunfei et al. (2017). “Fake News Detection Through Multi-Perspective Speaker Profiles”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing* Volume 2:8, pp. 252–256. URL: http://www.aclweb.org/anthology/I17-2043.

Rubin, Victoria et al. (n.d.). “Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News”. In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection.* URL: http://www.academia.edu/24790089/Fake%7B%5C_%7DNews%7B%5C_%7Dor%7B%5C_%7DTruth%7B%5C_%7DUsing%7B%5C_%7DSatirical%7B%5C_%7DCues%7B%5C_%7Dto%7B%5C_%7DDetect%7B%5C_%7DPotentially%7B%5C_%7DMisleading%7B%5C_%7DNews.

Yang, Yang et al. (2018). “TI-CNN: Convolutional neural networks for fake news detection”. In: *arXiv.* arXiv: 1806.00749.

Appendices

Appendix A

A-1 Dictionary Features

Feature	Definition	Examples
Loughran Mcdonald Dictionary		
LM_NEGATIVE	Loughran Mcdonald's words show negative tone	abducates, burden, careless
LM_POSITIVE	Loughran Mcdonald's words show positive tone	advancement, dream, innovator
LM_UNCERTAINTY	Loughran Mcdonald's words show uncertainty	approximate, doubted, speculate
LM_LITIGIOUS	Loughran Mcdonald's words show litigious tone	absolved, crime, executory
LM_CONSTRAINING	Loughran Mcdonald's words show constraining tone	confines, forbids, unavailability
LM_SUPERFLUOUS	Loughran Mcdonald unnecessary words	assimilate, theses, whilst
LM_INTERESTING	Loughran Mcdonald interesting words	extraordinary, rabbi, toxic

LM_MODAL WORDS STRONG	Loughran Mcdonald's words show strong modal	always, must, never
LM_INTERESTING	Loughran Mcdonald interesting words	extraordinary, rabbi, toxic
Laver Garry Dictionary		
LG_CULTURE-HIGH	Laver Garry's words show high culture	artistic, music, theatre
LG_CULTURE-POPULAR	Laver Garry's words show popular culture	media
LG_CULTURE-SPORT	Laver Garry's words show sport culture	angler, civil war, people
LG_ECONOMY	Laver Garry's words related with economy	accounting, earn, loan
LG_ENVIRONMENT	Laver Garry's words related with environment	green, planet, recycle
LG_GROUPS_ETHNIC	Laver Garry's words related with ethnic groups	Asian, race, ethnic
LG_GROUPS_WOMEN	Laver Garry's words related with women	girls, woman, women
LG_INSTITUTIONS_CONSERVATIVE	Laver Garry's words related with conservative institutions	authority, inspect, rule
LG_INSTITUTIONS_NEUTRAL	Laver Garry's words related with neutral institutions	chair, scheme, voting

LG_LAW_and_ORDER	Laver Garry's words related with law and order	police, punish, victim
LG_RUDAL	Laver Garry's words related with countryside	farm, forest, village
LG_VALUES_ CONSERVATIVE	Laver Garry's words with conservative values	glories, past, proud
LG_VALUES_LIBERAL	Laver Garry's words with liberal values	cruel, rights, sex
RID Primary Needs		
RID_ORALITY	RID's words show orality	belly, cook, eat
RID_ANALITY	RID's words show anality	anal, dirt, fart
RID_SEX	RID's words related with sex	lover, kiss, naked
RID Primary Sensation		
RID_TOUCH	RID's words related with touching	contact, sting, touch
RID_TASTE	RID's words related with tasting	flavor, savor, spicy
RID_ODOR	RID's words related with smelling	aroma, nose, sniff
RID_GEN_SENSATION	RID's words related with general sensation	awareness, charm, fair
RID_SOUND	RID's words related with sounds	bell, ear, music

RID_VISION	RID's words related with vision	bright, gray, spy
RID_COLD	RID's words related with cold	Alaska, ice, polar
RID_HARD	RID's words related with feels hard in touching	crispy, metal, rock
RID_SOFT	RID's words related with feels soft in touching	feather, lace, velvet
RID Primary DEFENSIVE_SYMBOL		
RID_PASSIVITY	RID's words related with passivity	bed, dead, safe
RID_VOYAGE	RID's words related with trips	journey, nomad, travel
RID_RANDOM MOVEMENT	RID's words related with random movements	jerk, spin, wave
RID_DIFFUSION	RID's words related with diffusion	fog, mist, shadow
RID_CHAOS	RID's words related with chaos	char, discord, random
RID_CHAOS	RID's words related with chaos	char, discord, random
RID Primary Regressive Cognition		
RID_UNKNOW	RID's words shows unknown feelings	secret, strange, unknown

RID_TIMELESSNES	RID's words related with infinity time	eternal, forever, immortal
RID_COUNSCIOUS	RID's words shows consciousness alteration	dream, sleep, wake
RID_BRINK-PASSAGE	RID's words shows brink passage	road, wall, door
RID_NARCISSISM	RID's words shows narcissism	eye, heart, hand
RID_CONCRETENESS	RID's words shows something specific	here, tip, wide
RID Primary Icarian Imagery		
RID_ASCEND	RID's words shows that something ascend	climb, fly, wing
RID_HEIGHT	RID's words related with height	bird, hill, sky
RID_DESCENT	RID's words shows that something descent	dig, drop, fall
RID_DEPTH	RID's words related with depth	cave, hole, tunnel
RID_FIRE	RID's words related with fire	solar, coal, warm
RID_WATER	RID's words related with water	ocean, sea, pool
RID Secondary feeling		

RID_ABSTRACT_ TOUGHT	RID's words related with abstraction	know, may, thought
RID_SOCIAL_ BEHAVIOR	RID's words related with social behavior	ask, tell, call
RID_INSTRU_ BEHAVIOR	RID's words related with instrumental behavior	make, find, work
RID_RESTRAINT	RID's words related with restraint behavior	must, stop, bind
RID_ORDER	RID's words related with order(form)	measure, array, system
RID_TEMPORAL_ REPERE	RID's words related with temporal references	when, now, then
RID_MORAL_ IMPERATIVE	RID's words related with moral imperatives	should, right, virtue
RID Emotions		
RID_POSITIVE_ AFFECT	RID's words related with positive emotions	cheerful, enjoy, fun
RID_ANXIETY	RID's words related with anxiety emotions	avoid, horror, shy
RID_SADNESS	RID's words related with sad emotions	hopeless, pain, tragic
RID_AFFECTION	RID's words related with affection	bride, like, mercy
RID_EXPRESSIVE_ BEH	RID's words related with expressive behavior	dance, sing, art

RID_GLORY	RID's words related with glory	elite, kingdom, royal
RID_GLORY	RID's words related with glory	elite, kingdom, royal
Other Dictionaries		
Uncertainty_words	Words that shows uncertainty	assume, could, maybe
Bad words		bastards, tits, porn
Common words	Words that commonly used	the, of, come
Emotional tone words		angry, happy, tolerant
Emotional words		bored, helpless, hurt
Negative words		abandon, abuse, concern
Positive words		boost, easy, enjoys
Hu_Liu Negative words	Hu Liu negative words	abrade, bankrupt, cataclysm
Hu_Liu Positive words	Hu Liu positive words	accurate, brighten, fascination
Litigious words		appeal, dockets, indict
Strong modal words		best, never, will
Weak modal words		could, depend, may
Slang words	Slang words are very informal language	hello, 2mr, 4give
AFINN Dictionary		

AFINN score	The AFINN lexicon is a list of English terms manually rated for valence with an integer between -5 (negative) and +5 (positive) by Finn Årup Nielsen	abuses: -3, amazing: 4, avoid: -1

Table 1: Dictionary Features

A-2 Complexity Features

Feature	**Definition**
Readability Index	
Flesch reading ease	$206.835 - 1.015(\frac{total\#of words}{total\#of sentences}) \quad (1)$
Flesch–Kincaid	$0.39\left(\frac{total\#of words}{total\#of sentences}\right) + 11.8\left(\frac{total\#of syllables}{total\#of words}\right) - 15.59 \quad (2)$
SMOG	$1.0430\sqrt{\#of polysyllables * \frac{30}{\#of sentences}} - 15.59 \quad (3)$

Automated readability index	$0.39\left(\frac{total\#ofwords}{total\#ofsentences}\right)+11.8\left(\frac{total\#ofsyllables}{total\#ofwords}\right)-15.59$ (4)
Dale-Chall	$0.1579\left(\frac{difficultwords}{total\#ofwords}*100\right)+0.0496\left(\frac{total\#ofwords}{total\#ofsentences}\right)$ (5) *Dale-Challe declare a list with difficult words

Coleman–Liau	$0.0588L-0.296S-15.8$ (6)
	L = Total # of Letters / Total # of Words * 100
	S = Total # of Sentences / Total # of Words * 100

Gunning fog	$0.4\left[\left(\frac{Total\#ofwords}{Total\#ofsentences}\right)+100\left(\frac{Total\#ofcomplexwords}{Total\#ofwords}\right)\right]$ (7)
Vocabulary Richness	
Yule K	Miranda-Garcia et al. Miranda-García and Calle-Martín 2005
TTR	$(Total\#ofuniquewords/Total\#ofwords)*100$
Brunets Index	$N^{V^{-a}}$, where N is the text length, V is the number of unique words, and –a is a scaling constant that is usually set at –0.172
Sichel	$Total\#ofhappaxdislegomena/Total\#ofwords$

Table 2: Complexity Features

A-3 Stylistic Features

Feature	**Meaning**
Part Of Speech Tags	
CC	Coordinating conjunction
CD	Cardinal digit
DT	Determiner
EX	Existential there (like: “there is” ... think of it like “there exists”)
FW	Foreign word
IN	preposition/subordinating conjunction
JJ	adjective ‘big’
JJR	adjective, comparative ‘bigger’
JJS	adjective, superlative ‘biggest’
LS	list marker 1)
MD	modal could, will
NN	noun, singular ‘desk’
NNS	noun plural ‘desks’
NNP	proper noun, singular ‘Harrison’
NNPS	proper noun, plural ‘Americans’
PDT	predeterminer ‘all the kids
POS	possessive ending parent’s
PRP	personal pronoun I, he, she
PRP$	possessive pronoun my, his, hers
MD	modal could, will
RB	adverb very, silently

RBR	adverb, comparative better
RBS	adverb, superlative best
RP	particle give up
TO,	to go 'to' the store.
UH	interjection, errrrrrrrm
VB	verb, base form take
VBD	verb, past tense took
VBG	verb, gerund/present participle taking
VBN	verb, past participle taken
VBP	verb, sing. present, non-3d take
VBZ	verb, 3rd person sing. present takes
WDT	wh-determiner which
WP	wh-pronoun who, what
WP$	possessive wh-pronoun whose
WRB	wh-abverb where, when

Table 3: Part Of Speech Features

Feature	**Meaning**
Structural	
total_number_of_sentences	
total_number_of_words	
total_number_of_characters	
total_number_of_begin_upper	Words with first capital letter

total_number_of_begin_lower	Words with first lowercase letter
total_number_of_all_caps	Word with all capital letters
total_number_of_stopwords	
total_number_of_lines	
number_of_I_pronouns	
number_of_we_pronouns	
number_of_you_pronouns	
number_of_he_she_pronouns	
number_of_exclamation_marks	
number_of_quotes	
number_of_happax_legomena	Word types that occur only once in text
number_of_happax_dislegomena	Word types that occur only twice in text
has_quoted_content	
ratio_alphabetic	
ratio_uppercase	
ratio_digit	
avg_number_of_characters_per_word	
avg_number_of_words_per_sentence	
avg_number_of_characters_per_sentence	
avg_number_of_begin_upper_per_sentence	
avg_number_of_all_caps_per_sentence	

avg_number_of_begin_lower_per_sentence	
avg_number_of_stopwords_per_sentence	

Table 4: Structural Features