A Systematic Approach to Analyzing Syndication of Viral News Stories

17-631 Information Security, Privacy, & Policy

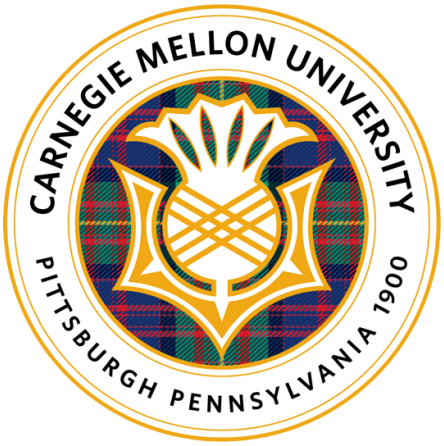
Gaurav Deshpande
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
gdeshpan@andrew.cmu.edu

Sam E. Teplov
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
steplov@andrew.cmu.edu

Alon Peer
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
apeer@andrew.cmu.edu

Dr. Nicolas Christin
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
nicolasc@andrew.cmu.edu

December 7, 2019



Abstract

In this work, we present a novel, semi-autonomous methodology for analyzing syndicated news articles. We first detail our manual process of news syndication analysis from which we observe a number of unsettling patterns. From there, we go on to categorize and define two different forms of syndication: top-down and bottom-up. We then transform our manual process into a semi-automated one which combines web scraping and machine learning techniques to create and then analyze a data set of syndicated news stories. From the analysis, we are able to make some inferences about the nature of syndicated news articles, as well as trace back a viral news story to its original publisher and article. By presenting this work, we hope to minimize the effects that blind news syndication have on the spread of misinformation.

Keywords Fake news · News propagation · News syndication

Contents

1 Introduction

Misinformation, disinformation, and fake news have become a common occurrence in our interconnected society. The idea of intentionally spreading fake news to sway public opinion, also known as disinformation, was brought to light by the U.S. Senate Select Intelligence Committee's report on Russian meddling in the 2016 U.S. presidential election [1]. Misinformation, or the unintentional spread of inaccurate news, has also become ubiquitous thanks to social media, which allows unreliable news to be proliferated quickly and without any verification [2]. The use of a combination of misinformation and disinformation by nation states, non-state actors, and other politically motivated groups, has become so prevalent that it has affected over seventy countries [3]. This problem has become so extreme that the integrity of the U.S. election system, and even the fundamental democratic ideals of the U.S., have been called into question by some [4].

But what exactly is fake news? According to the Cambridge Dictionary, fake news is "false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke" [5]. In addition to this definition, fake news may also include clickbait stories, propaganda, satire, biased news, or even sometimes just bad journalism [6].

These days, anyone can publish content on the internet. Whether it is on a website, a blog, or a social media profile, it can potentially reach large audiences. The fact that people nowadays rely on the internet more than ever to consume news does not make this problem any easier to solve. Fake news is very profitable for the publisher as viral stories generate large amounts of web traffic, thus increasing advertisement revenue [6]. For that reason, fake news is widely common around the internet. Another motive to publish fake news is to promote a political agenda. Fake political news stories are meant to persuade consumers to accept biased false beliefs. This is the main reason detecting fake news is such a challenge. These false stories are created to sound true and interesting.

Due to the global nature of this problem, over fifty countries have begun to implement measures to combat misinformation and disinformation [7]. The measures implemented by these countries span everything from laws and legislation, to creating task forces and launching official investigations to try to weed out disinformation [7]. Besides nation states trying to tackle this behemoth of a problem, many researchers and academics have also been trying to implement various methods to try to detect and classify misinformation, disinformation, and fake news. This area of research has become commonly known as *fake news detection* [8].

1.1 Problem Statement

One can try and detect fake news by checking the source of the story, read beyond the title (which is intended to attract users and thus contains more interesting and controversial topics), try to trace the article back to the original source, or even make sure the article is not satirical in nature [6]. Even though it is possible, doing so is not an easy task, and that is why there have already been many different approaches proposed to detecting and classifying fake news. Many proposed solutions have mostly relied on various machine learning techniques, such as using deep neural networks, or recurrent learning models [9]. These models for the most part work by training the model on a variety of features ranging anywhere from text-based features, to network features that show how the article has been spread through space and time [9]. The problem with these models is that they rely on a lot of information to be fed into them, which take a long time to be extracted. Also, many of these approaches have only been tested on a few media outlets and social media platforms, so they could very well not work on a global scale.

One of the largest problems that has not be widely discussed or researched is that many media outlets simply take news articles from other news sources and republish them as their own. This process is known as news syndication. This is an extremely dangerous practice that can lead to the problem of fake news being further exacerbated. There are certain types of news syndication which are generally safe and do not lead to misinformation being spread. As we discuss later, other types of news syndication are dangerous and could easily be exploited by nation states seeking to carry out a disinformation campaign against another country.

1.2 Objectives

Our first objective of this work is to build a system that given a viral article, can find related articles that were either syndicated off of the original article, or articles that were syndicated by the original article. This allows us to achieve our second objective of building a propagation path which shows where the original article was published, and by whom. The third objective of this work is to then analyze the propagation path of the article and see what features change over time, how they change over time, and identify any patterns.

The last two objectives of this work were not accomplished in this research, but we hope they will be completed in future work. For one, we hope to eventually fully automate this process by incorporating various models for clickbait title identification, so that the clickbait model could feed our system articles to be analyzed. Our second future objective is to eventually incorporate existing fake news detection models so that once our system identifies the original article from the propagation path, it can then be classified as fake news or not. In the end, we hope to be able to provide the media consumer with a message stating where the article originated from, and if its veracity can be verified or not.

1.3 Contributions

We make a number of contributions throughout this work. First, we present out initial manual methodology which outlines how we first identified viral articles, how we traced them back to their origin, and the findings from our analysis. From this analysis, we present and define two different types of news syndication: top-down and bottom-up. We then present a refined, semi-autonomous methodology which is derived from our initial methodology, but with a few revisions. This methodology, given a viral news story and a couple related articles, will automatically scrape the web and identify articles within a time window which were either syndicated from the original set of articles, or articles that the original set of articles syndicated. We use a combination of similarity score and key words that eventually feed into a dendrogram which clusters similar articles together. We then take this clustered data set and build a timeline which allows the original article to be identified. We also are able to perform a wide range of analysis on various article features and categorize how certain features change over time. This analysis then allows us to identify patterns about syndicated articles and eventually draw conclusions.

1.4 Definitions

The realm of fake news detection has become inundated with a variety of various terms that are used interchangeably depending on the article. It is important to define these various terms in order to avoid ambiguity and misunderstandings. The following definitions are taken from the Ethical Journalism Network [10]:

- **Fake News**: *"Fake news is information deliberately fabricated and published with the intention to deceive and mislead others into believing falsehoods or doubting verifiable facts"*
- **Disinformation**: *Information that is false and deliberately created to harm a person, social group, organization or country*.
- **Misinformation**: *Information that is false, but not created with the intention of causing harm.*
- **Malinformation**: *Information that is based on reality, used to inflict harm on a person, organization or country*.

The main difference between these definitions is the intent and what the information is being used for. These terms may be used interchangeably in this work. This is because our work does not intend to differentiate between the intent of the news. Our work focuses on building a methodology that is able to semi-autonomously trace back an article to its origin and analyze how article features change over time. Future work involves trying to classify news as fake or not, but that involves using already developed models for classification. Not even these models attempt to differentiate between different intentions. Therefore, although misinformation, disinformation, and fake news do technically have different meanings, they can be used interchangeably in this line of work since most research on fake news detection just attempts to classify news as accurate or inaccurate, and does not try to determine the intent behind the article.

1.5 Limitations

As stated before, this work does not attempt to make a determination of whether the content of an article is true or false. Instead, this research is primarily focused on presenting a semi-autonomous methodology which can be used to identify the propagation path of a viral article, as well as conduct analysis on article features. In the future, we hope to fully automate this methodology, as well as incorporate existing fake news classification models in order to actually make a determination about content accuracy. However, for now, the scope of this work is strictly limited to semi-autonomously identifying the propagation path of an article, as well as conducting time series analysis on articles features.

2 Background

The field of misinformation and fake news has become widely researched by academics ever since the Russian misinformation campaign against the 2016 United States presidential election came to light [1]. The most straightforward

approach to detecting fake news is applying various machine learning techniques, and this is what many researchers have focused on thus far.

2.1 Fake News Detection

Shu et al. go through a comprehensive overview of all of the various ways of attacking the problem of fake news detection and classification from a data mining perspective [9]. From the perspective of detecting fake news via machine learning techniques, Shu et al. outline that there are generally four different types of features that can be used to detect fake news: news content features, linguistic-based features, visual-based features, and social context features. These features are then applied to one of two different types of models: news content models, or social context models . Just as they sound, news content models focus on fact checking claims in the news article or looking at stylistic features that may indicate that the article is fake. Social context models, on the other hand, focus on how users interact with the article, the virality of the article, and the propagation path of the news story. These two different methods are rarely used independently, but rather compliment each other in many hybrid approaches, such as ensemble models [9].

2.2 Open Issues in Fake News Detection

Shu et al. close by giving an overview of many of the various open areas of research in this field, as well as proposing some future research ideas. They outline four future research directions: data-oriented, feature-oriented, model-oriented, and application-oriented. Within the area of data-oriented research, Shu et al. bring forward a couple of issues that need to be addressed. For one, there is no large, encompassing fake news data set which can be used as a benchmark to facilitate future research [9]. In fact, this was such a large problem that in 2018, Shu et al. published another work detailing their methodology for building a large scale, open source, data set consisting of fake news, which is now available on Github [11]. Another idea that Shu et al. propose for future research within the data-oriented direction is early fake news detection [9]. Many of the current state-of-the-art fake news detection methods rely on data that needs to be accumulated over time and information that is not readily available when news is just starting to go viral [12]. Shu et al. suggest creating a system that provides fake news alerts during the dissemination process so that misinformation can be flagged before it is viewed by a large audience. Another suggestion that is made is to look at social media posts made within some time delay of the original posts as a means of verifying the veracity of the post [9].

3 Related Work

There has been work done on early fake news detection by analyzing propagation paths and time series, as well as analyzing the dynamics of news syndication. Most of this work, however, has been limited in scope in terms of the domains that it covers, as well as the type of media being analyzed. For example, past research has focused on either a few social media platforms, a specific media outlet, or a single country [12],[13], [14]. The research that has spanned across multiple different domains of news (social media, various news outlets, etc.) has been limited in scope when considering the type of media that is being analyzed [15]. There has been little work, if any, done specifically on the syndication of viral news and trying to automatically analyze and trace back a viral news story to its original publisher.

3.1 Domain-Specific Approaches

Liu et al. propose a method of early fake news detection by modeling the propagation paths as a multivariate time series. They build a classifier that uses recurrent and convolutional networks to detect fake news. They were able to achieve an accuracy of 85% on Twitter and an accuracy of 92% on Sina Weibo when it came to correctly classifying fake news. Liu et al. claim that they were able to detect the misinformation five minutes after it began to spread [12]. The downside to this approach is that it relies on user interactions with the article to generate the data needed to feed into their classifier. This means that some users will be exposed to the fake news before it can be flagged as misinformation just due to the nature of the dataset that they are using. Also, this approach has only been tested on social media platforms and micro blogs, meaning that this methodology cannot be applied on a global scale that spans hundreds of different news outlets and social media sites.

Jin et al., in their work, tackle the problem of misinformation on the popular website Micro blog. They propose a three-layer hierarchical model that establishes a credibility score for each post and then propagates this score throughout the network. They then formulate this propagation process as a graph optimization problem and find a globally optimal solution. Jin et al. were able to boost accuracy by 6% over the baseline model [13]. Again, just like Liu et al.'s approach, Jin et al. focus solely on one website: Micro blog. Their model is reliant on data to be taken specifically from Micro blog. Although their technique could in theory be applied to other platforms, their approach is mostly domain-specific and does not analyze misinformation on a global scale.

Wang et al. explore and analyze the patterns of news propagation and syndication in Chinese news media. They draw comparisons between the way news diffuses through time and space and how an epidemic spreads. Something interesting that they found was that 80% of news outlets in China were responsible for re-printing news articles directly from the source [14]. Some of our findings also substantiate this claim. Like much of the other research on this topic, Wang et al. focus only on Chinese media outlets and they also did not attempt to trace an article back to its original source without knowing what that source was in the first place. Although Wang et al.'s research brings to light some interesting patterns in the ways that news articles propagate and are syndicated, the work is confined mostly to Chinese news media, which is known to be heavily regulated by the Chinese Communist Party [16].

3.2 Media-Specific Approaches

Zannettou et al. take a different approach to analyzing propagation; instead of looking at social media posts or news stories, they focus on politically motivated memes. Unlike any of the previous research discussed, Zannettou et al. pool data from a number of various sources: Twitter, Reddit, 4chan, and Gab. Using memes from these sites, they are able to trace the propagation path of each meme and even draw conclusions about the influence that each meme outlet has. Zannettou et al. use Hawkes process to model how memes from various sources interact with each other and quantify the influence that each meme has on other memes [15]. Although this research encompasses many different media outlets and social media platforms, it fails to analyze any other form of media besides memes. Some of the processes outlined in this work could be applied to news articles, but much of it is very specific to memes and images and cannot be generalized to news articles due to the difference in data between an image and a set of text. Nevertheless, Zannettou et al. showed that memes are often syndicated by larger media outlets which causes fringe web communities to have much broader reach than they ever could have before [15].

4 Phase I - Manual Approach

Our research and methodology can be split up into a manual process (phase I), and a semi-autonomous process (phase II). In the manual approach, we created a data set of popular news articles and then attempted to trace them back to the original publisher. Once we identified the propagation path, we attempted to analyze how various article features changed over time. Although inefficient, the takeaways from phase I of our research helped inform our much more efficient and effective methodology in phase II.

4.1 Dataset Creation

We started by collecting a diverse range of articles from a number of different news sources. As seen in Appendix A, we selected five articles from each news source and record each articles' associated comment count. These articles were picked between October 5 2019 and October 8 2019. We looked for articles that were published in the past two weeks. We picked media outlets that tended to be biased either towards the political left or right as defined by the chart in Appendix B. The reason we did not pick any mainstream, unbiased news outlets (New York Times, Washington Post, Economist, etc.) is because these news sources typically do their due diligence in ensuring that the news that they are publishing is true, and the public typically trusts these sources [17]. It is also worth noting that we did not consider any articles that were about, or mentioned, President Donald Trump, as these could possibly skew results. Aside from that, we did not consider the actual content of the article when picking articles for the initial dataset.

4.2 Case Study: Chinese Gold

One of the first articles that stuck out to us from the initial dataset was the article from the Daily Mail titled *Thirteen and a half tonnes of gold worth up to £520million is found in a corrupt Chinese official's home and £30BILLION in suspected bribe money in his bank account*.

4.2.1 Propagation Analysis

The reason this article jumped out at us is because the number of comments for this article was much higher than any article in the dataset. Once we identified this specific story, we decided to try to trace it back to its original source. The process of trying to trace an article back to its origin is an extremely tedious and time consuming task. This is because there is no systematic approach already out there that can help us gather the information that we need. The data used to build the propagation path for the Chinese Gold Story can be found in Appendix C.

As it turns out, the original content for the news story was actually a twitter post. What is fascinating is that none of the articles actually state that Twitter was the first platform where the content for the story appeared; many of the news

Figure 1: First Article About Chinese Gold That We Found

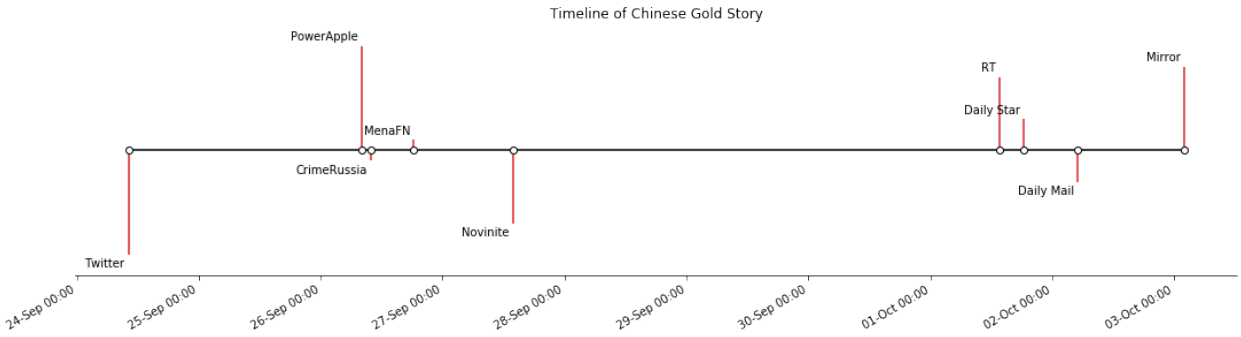


Figure 2: Timeline of Chinese Gold Story

outlets did link to the Twitter post, but not once did any of the news sources explicitly say "this content was originally posted on Twitter by a user." It is also important to note the time delays between when the story was first posted on Twitter, and when it was picked up by mainstream media. The initial twitter post was made on September 24. It took about two days for international news outlets (PowerApple, CrimeRussina, MenaFN, Novinite) to pick up this news story. It took about seven days for mainstream news outlets (RT, Daily Star, Daily Mail, Mirror) to pick up this news articles. It is important to note these time delays as they can be used to categorize bottom-up syndication, which we will define later in this work.

4.2.2 Article Feature Analysis

Besides just looking at how the article travels through space and time as it is syndicated, we were interested in looking at how the articles' properties change over time.

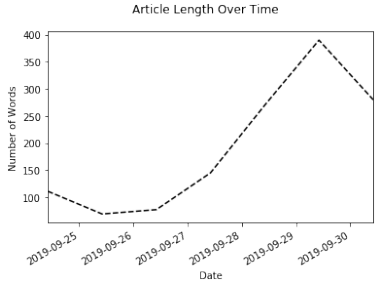


Figure 3: Article Length Over Time

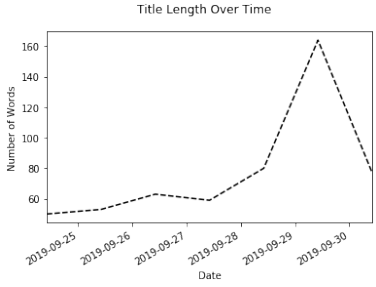


Figure 4: Title Length Over Time

From Figures 3 and 4, we can see that the general trend for both the article length and title length is that as the article is syndicated, both features become longer. By actually analyzing the content of the article, it seems as each article is syndicated by another news outlet, they make more inferences about the article. For example, the original Twitter post just shows a video, and short description about the video. By the time Daily Mail posts their version of the article, they are making inferences about the official's total net worth, his assets, as well as general claims about anti corruption in China. None of these topics are ever discussed by the original Twitter post. The Daily Mail just assumes these as fact and wrap everything up into one story.

4.3 Case Study: University of Virginia

Once we found the story about the Chinese gold story, we wanted to see if other similar articles would exhibit the same properties and characteristics. We started by looking for articles whose origin could be traced back to either a social media post, or a local news media outlet. One of the first articles that we identified was the story titled *University of Virginia cancels 21-gun salute from Veterans Day ceremony* from Fox News.

Figure 5: First Article About UVA That We Found

4.3.1 Propagation Analysis

When we traced the propagation path back to the original published article, it turned out that it was actually an opinion piece from a local newspaper which is located around the University of Virginia. Once again, we have to use a manual process to trace back the article to its origin since we started with the Fox News article, which had already been syndicated multiple times.

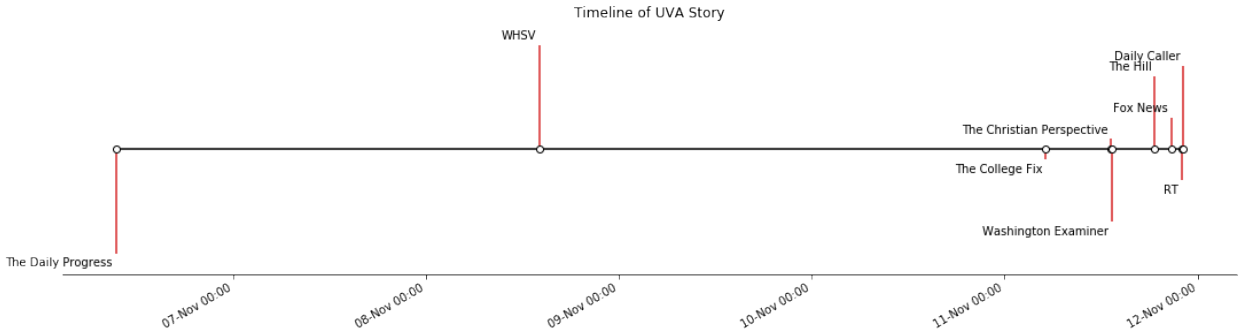


Figure 6: Timeline of UVA Story

Analyzing the propagation path, it took about five days for the story to move from the local newspaper, to national news. It is difficult to compare this timeline to that of the Chinese Gold story since that story was international in scope and this story is local to the United States. Nevertheless, it is interesting to note the time delay between when the article was first posted by the local newspaper, and when a national media outlet, such as Fox News, syndicated the article.

4.3.2 Article Feature Analysis

We wanted to see if the general trends from the Chinese gold story of article length increasing over time and title length increasing over time would hold true with this article. If they did, this could lead to an interesting correlation that could be drawn between syndicated articles that originate from local news outlets, or social media platforms.

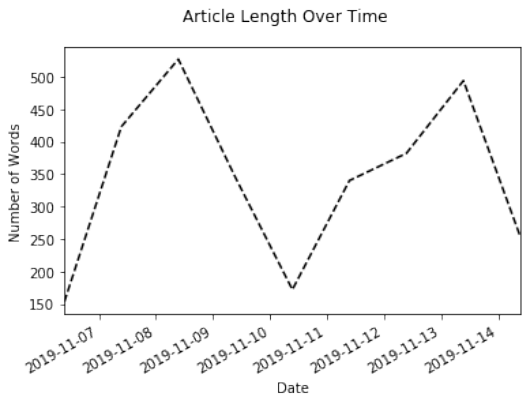


Figure 7: Article Length Over Time

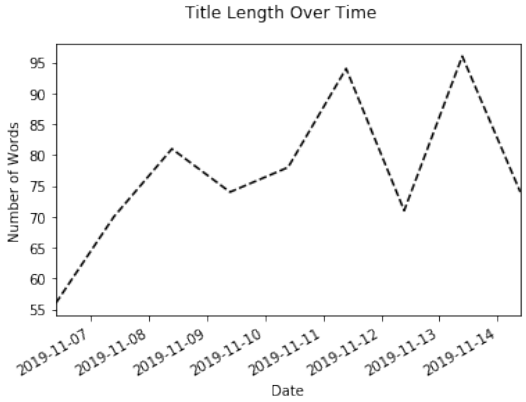


Figure 8: Title Length Over Time

Unlike the Chinese gold story, there was no clear correlation between time since original publication of the article, and the article length. However, the length of every syndicated article was longer than the original. This same pattern can be seen with the title length over time; every syndicated article title length was longer than the original article's title length.

4.4 Findings

Since our manual approach is extremely tedious and time consuming, the data that can be used to draw conclusions is quite limited. However, the findings from the manual approach help us better inform our phase II methodology. We are also able to make some hypothesises about patterns in syndicated news articles which are then either substantiated or disproved by the findings from phase II of our research. Also, using our observations from phase I, we are able to characterize and define two different types of news syndication.

4.4.1 Patterns and Trends

One of the first patterns that we observed was that syndicated articles are generally not syndicated immediately; there is typically a delay of about five to seven days from when the article is first posted and when a mainstream media outlet syndicates it. We also learned that the article length and title length of syndicated articles are always longer than that of the original article. Upon manual inspection of syndicated articles, we found that they rarely, if ever, referenced that their article was a syndication of another article. They did sometimes cite where they got their data from, but ironically, this was usually just another syndicated article. This is why tracing back a syndicated article back to its origin is such a tedious and time consuming task.

4.4.2 Different Types of Syndication

Something else that we noticed during the course of phase I of our research is that there seemed to be two different kinds of syndication going on, and one seemed to be much more dangerous than the other. One type of syndication, as seen with the two case studies presented above, we titled *bottom-up syndication*. In this case, news stories from either local news outlets, or social media platforms, are syndicated by more mainstream platforms. As time goes on, the media outlets that syndicate the article increase in popularity. The other type of syndication that we categorized was *top-down syndication*. In this case, the original publisher is a credible publisher such as The Washington Post, New York Times, Economist, Wall Street Journal, etc. [17]. An example of *top-down syndication* can be seen with the Washington Post Article titled *Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans* [18]. This article was first posted on November 11 by the Wall Street Journal. Almost immediately, the article was syndicated by news outlets such as St. Louis Post-Dispatch, Seeking Alpha, and CNN within the same day [19], [20], [21]. The difference between these syndicated articles and the syndicated articles from the previous two case studies is told fold: (1) all of the syndicated articles gave clear attribution to the Wall Street Journal, and (2) there was a much smaller time delay between the when the original article was posted and when the syndicated articles became popular.

We propose the fact that *bottom-up syndication* poses a larger threat when it comes to the spread of misinformation than *top-down syndication* does. This is due to the fact that articles that are syndicated from a mainstream, verified, news outlet are usually well researched, accurate, and are generally well received by the public [17]. Because of this, the media outlets that syndicate from these trusted sources have no issue with citing the original media source. On the other hand, articles that originate from lower level media outlets, such as local news, satirical news sites, or social media, may not put in the same amount of resources and effort in verifying the accuracy of their claims; in fact, their claims may be completely fabricated intentionally. Therefore, the media outlets that syndicate these articles from lower levels may just be doing so because the title can serve as clickbait, which would in turn bring more readers to their website and ultimately generate more revenue from ads. We are not claiming that this is the case with every article that is syndicated following a bottom-up approach, but rather that there is a bigger risk for misinformation to be spread when news is syndicated in this manner.

4.5 Limitations of a Manual Approach

The manual approach that we applied in phase I of our research has quite a few limitations. First off, we did not have a systematic process for finding viral articles. This is something that we fix in phase II of our research by applying concrete criteria for categorizing an article as viral. Next, there was a major limitation in the way that we found syndicated articles. By looking for similar articles by hand, there was no way that we could find every single related article. The ones that we could find would take a while to track down and overall, the whole process of tracing the propagation path was extremely tedious and time consuming. This drawback also caused us to not be able to develop a large dataset for analysis. This issue was fixed in phase II of our research by applying web scraping techniques, various APIs, and machine learning techniques to automatically find and group together related articles. Overall, phase I of our research was inefficient, but was still extremely valuable because it exposed us to two different types of syndication, brought to light some interesting patterns in news syndication, and allowed us to better inform our phase II methodology.

5 Phase II - Semi-autonomous Approach

In the semiautomated approach, we wanted to see if our findings from phase I would still hold true against a larger data set. We automated the process of tracing the propagation path and conducting analysis on the resulting articles. The resulting methodology from phase II is the main contribution of this work.The overall algorithm can be seen in Algorithm 1 and Algorithm 2

The approach can be broadly categorized into two stages.

1. *Dataset Generation*: As a part of this data generation we identify the viral news articles, scrape for articles that are similar to it and gather metadata on these set articles. The raw data is passed through machine learning models to group them as relevant articles and discard non relevant articles.

2. *Analysis*: Numerous data visualization methods were applied based on the findings of phase I. These included time series analysis and histograms on certain features.

Algorithm 1: Dataset generation algorithm

Result: Generates raw dataset of news articles
for *url in list* **do**
    title, desc = get_title_desc(url);
    title, desc = remove_html_tag(title,desc);
    add_to_meta(title,desc);
end
for *news in meta* **do**
    related_urls = query_news_api(news);
    add_unique_to_meta(related_urls)
end
for *news in meta* **do**
    extract_article(news);
    extract_reactions_upvotes(news);
    add_to_final_dataset();
end

5.1 Dataset Generation

The one of the key parts of phase II is to generate a large enough dataset for analysis. As a part of this stage we identify news articles, that can potentially be called viral news stories. Based on these news articles we build a larger data-set comprising of all other relevant articles, metadata about these articles and the response these stories received on platforms like Reddit and Facebook.

```
In [12]: url_data = pd.read_csv("./dataset/url.csv")
         url_data.columns

Out[12]: Index(['url', 'date', 'group'], dtype='object')
```

Figure 9: Columns in the url dataset

5.1.1 Identification of Viral News Articles

As in phase I, we did not use articles published by unbiased media sources and steered clear of stories related to Donald Trump. An article with more than *ten thousand reactions*[1] on facebook was quantified as a viral news article. The articles selected in the end were a heterogeneous mixture of local and international events. These new articles were used to manually build *url.csv* dataset. The features present in the dataset are listed in Figure 9

This list of URLs is used to generate an intermediate dataset, *meta.csv*, which consists of features like title, description and content of the articles. For extracting contents of the web pages, we use *embed.ly* [22]. It provides */1/extract/* API which can be used to retrieve article text, title, and other meta-data. This raw data is processed to remove HTML tags and HTML symbols.

```
In [13]: url_data = pd.read_csv("./dataset/meta.csv")
         url_data.columns

Out[13]: Index(['Unnamed: 0', 'description', 'image', 'publishedAt', 'source.name',
                'source.url', 'title', 'url'],
               dtype='object')
```

Figure 10: Columns in the meta dataset

5.1.2 Scraping Similar News Articles

To correctly identify the origin of a particular article, or to see the propagation path for an article, it is crucial to extract all the relevant articles of the web. In phase II, we used various news aggregator APIs like google news [23], NEWSAPI [24] to identify more articles on the relevant news topic.To generate better results, the *stop words* are removed. All the results generated by the query are added to the *meta.csv* dataset. The columns for *meta.csv* can be seen in Figure 10.

Stop Words Words that do not provide any discriminative power are referred to as stop words[25]. We used the classic method to remove the stop words. In this method, any word present in a precompiled list is removed from the text. We used Standford NLP's list of stop words [26].

5.1.3 Building The Dataset

Once all of the relevant urls are gathered, we query reddit APIs [27] and sharedcount APIs [28] to gather relevant metadata about the articles. The final dataset consists of the column as in the Figure 11.

[1] reaction is the sum of likes, comments and shares

```
: url_data = pd.read_csv("./dataset/data_v1.csv")
  url_data.columns

: Index(['Unnamed: 0', 'Unnamed: 0.1', 'language', 'title', 'content',
         'description', 'provider_url', 'keywords', 'provider_display',
         'provider_name', 'total_count', 'comment_count', 'share_count',
         'reaction_count', 'reddit_upvotes', 'reddit_comments', 'date',
         'title_len', 'content_len', 'description_len', 'text', 'group'],
        dtype='object')
```

Figure 11: Columns in the raw dataset

5.2 Data Clustering

Algorithm 2: Clustering algorithm

Result: Clusterized dataset
for *article in data* **do**
    article_text = article.title+article.description+article.content;
    vectorized = tfidf_vectorizer(article_text);
    cosine_similarity(vectorized);
end
hierarchical_cluster(vectorized_articles);
cluster_trim(level=3);

The dataset generated in the previous stage was raw. The articles present in the dataset were merely scraped from the APIs, without enquiring the relevance of articles from the response of these APIs to the original articles that were manually identified. Clustering models help remove the outliers from the dataset and also group the articles on the basis of there similarity scores. We experimented with numerous clustering approaches like K-Nearest-Neighbors, KMeans and Hierarchical clustering.

5.2.1 Similarity Score

Similarity scores helps us answer the question, as to how close two pieces of text are. Similarity score, thus, is a measure of lexical and semantic similarities [29]. The similarity score in the current algorithm is calculated using a TF-IDF vectorizer with cosine similarity score.

The Figure 12, the yellow diagonal is congruent with the fact that, the figure represents a symmetric matrix. However, the square blots with varying shades of yellow and green, suggests high probability plagiarism. The yellow blots, in bottom right part of the image implies the articles were simply copied from the other source. In some cases these are legitimate news websites, while in some cases they are search engines like yahoo-news and msn-news.

TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency. TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user [30].

Cosine Similarity Score

The cosine of two non zero vectors A and B is defined as an Euclidean Dot Product of the two vectors. The similarity between two vectors is given by

$$similarity = cos\theta = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}\,||\;||\vec{b}\,||} \tag{1}$$

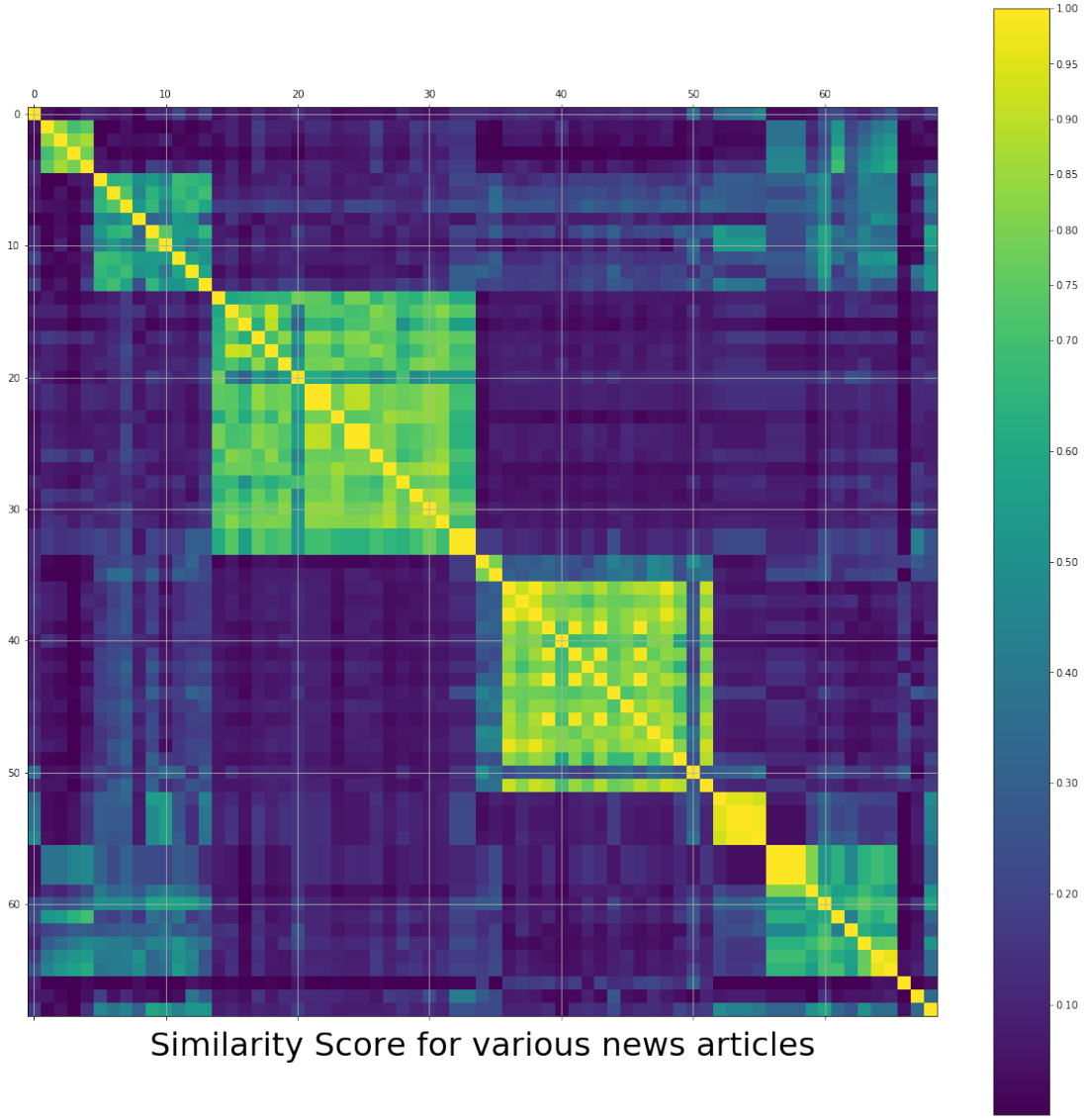


Figure 12: Similarity Score Matrix
Similarity Score of 1 means articles are exactly same. While Score of 0 means they do not match at all.

5.2.2 Semi Supervised Hierarchical Clustering

Hierarchical clustering algorithms are unsupervised methods to generate tree-like clustering solutions. They group the data points into a hierarchical structure using bottom-up (agglomerative) or top-down (divisive) approaches [31]. In agglomerative approach, the articles present as leaf nodes are most similar. Hierarchical Clusters generate the dendrogram, which represents various levels of clusters based on similarity of lower clusters. To group the news articles, we identified the best cut was at level 3. This provided us with a accuracy in grouping of **96%**.

In Figure 13 we can see that the news articles with similar titles are grouped closely. The vertical red line determines the level on which the tree is cut. Thus is clearly separates unrelated articles.

Thus, by the end of Data Clustering stage, we had clean, grouped data, which was later used for analysis.

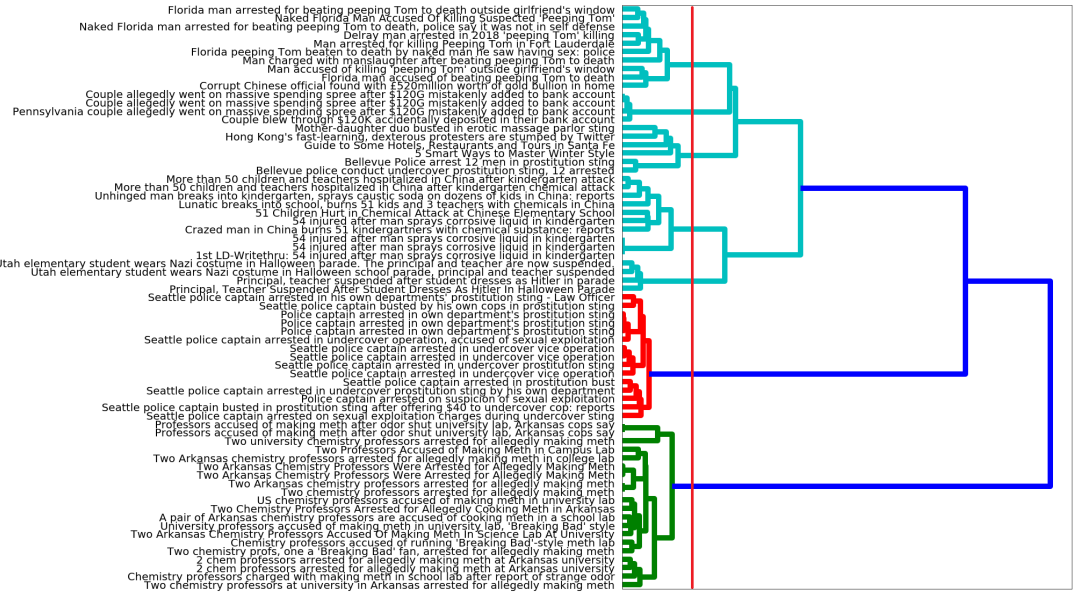


Figure 13: Dendrogram for news articles
The vertical red line cuts the dendrogram into smaller clusters

5.3 Data Analysis

From using the semi-autonomous methodology presented in phase II of our research, we were able to conduct propagation and feature analysis on a much larger dataset. Some of our findings in phase I held true even with a larger dataset, whereas some other findings and hypothesises from phase I were shown to be false.

5.3.1 Propagation Analysis

Like in phase I, we were able to identify the original article after plotting the propagation path. What was different was when we used a semi-autonomous approach in phase II, we were much more certain that we actually pinpointed the original article, whereas in phase I we were just assuming that we found the original article. For this section, we selected one story to display results for; the results for the other news stories can be found in appendix E. A complete dataset used in phase II can be found in the GitHub Repository (https://github.com/kratos-an0nym0us/spectacle) .

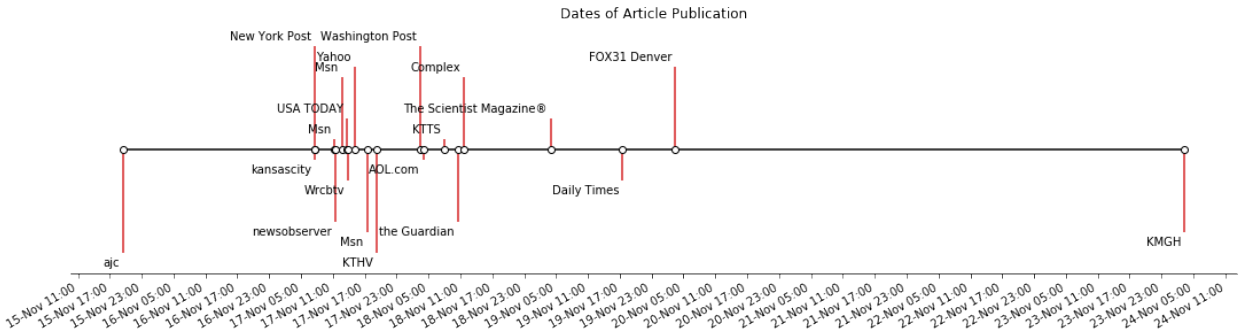


Figure 14: Timeline of Kentucky Story

As seen in the example of the Kentucky story, the story took about two days to spread from a local news source, to a more mainstream news source. Although this does not follow the pattern identified from phase I (five to seven days), it still does show that there is some sort of delay between when a story is published by a local media outlet or on a social media platform, and when it is picked up by a mainstream news source. In appendix E, it is clear that articles took anywhere from one to seven days to be syndicated by a mainstream news outlet. Although this is a large range, it can

still be used as a categorical measure to distinguish between *top-down syndication* and *bottom-up syndication*, since articles that follow top-down syndication are typically syndicated within the same day.

5.3.2 Article Feature Analysis

Looking at figures 15 and 16, there is no clear correlation between article length and time, or title length and time. Even the hypothesis from phase I that every syndicated article would have a longer title length and article length than the original article did not hold true.

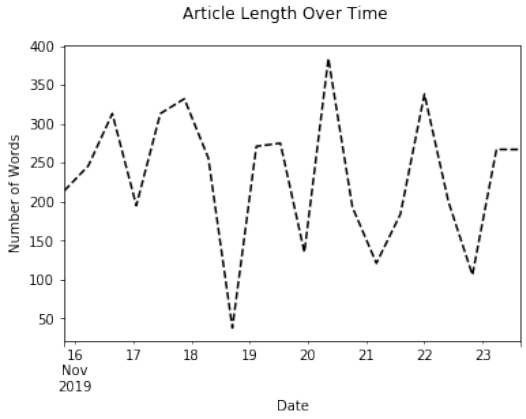


Figure 15: Kentucky Article Length Over Time

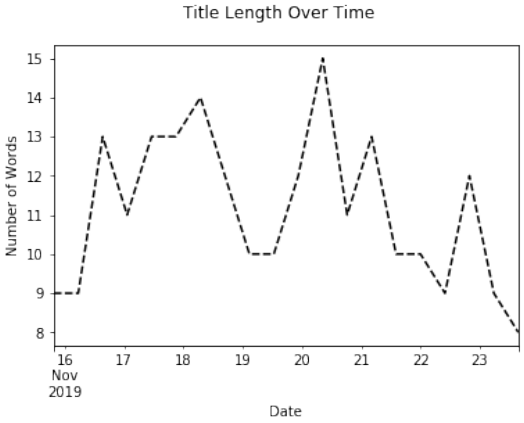


Figure 16: Kentucky Title Length Over Time

Although the correlations from phase I did not hold true when looking at a larger dataset, there are many other article features that can be analyzed in the future. Not seeing any correlations is a finding within itself and is important for showing that categorizing and finding syndicated article just based off of article features is a difficult task.

5.4 Findings

There were a number of findings from phase I that we actually showed to be false in phase II of our methodology. Nevertheless, showing something to be false is just as important as showing something to be true. There are also a number of findings that we confirmed in phase II of our research, as well as some new findings we made with a semi-autonomous methodology. All in all, our findings for phase I and II can be summarized as follows:

1. In the case of *bottom-up syndication*, original articles were generally published one to seven days before renowned media outlets syndicated the articles.
2. In the case of *top-down syndication*, media outlets syndicated the article within half a day of the publication of the original article. Most of these articles did mention a reference to the original article.
3. The hypothesis that the length of the article increases over time did not hold up to be true.
4. The hypothesis that the length of the title increases over time did not hold up to be true.
5. The hypothesis that syndicated articles always had longer titles and longer article content than the original did not hold up to be true.
6. A significant number of articles hold a similarity score of greater that 0.8 within their cluster, suggesting plagiarised content.

5.5 Limitations

Although our methodology presented in phase II does not have as many limitations as phase I, there are still some notable limitations that need to be discussed. For one, the methodology from phase II is not completely autonomous; it still requires a user to input a viral article, as well as some related articles. Also, the dendrogram does not get a high enough accuracy for conducting analysis, so the dataset had to be manually inspected and edited in order to enable data analysis. In addition, the model presented in this work does not make a determination about the accuracy of an article. Instead, this methodology provides a means to track an article back to its original publisher, as well as enables trivial analysis on article features.

6 Future Work

Future work can be grouped into two separate categories: improvement in methodology, and incorporating relevant fake news detection and clickbait detection models.

6.1 Methodology Improvement

We intend to automate the project. In its autonomous state, the system will be able to scan the headlines and/or top stories for various news outlets, track its propagation path and generate a report. We also plan to broaden the scope of search for articles, as currently we work in a narrow domain of news APIs. We plan to publish a dataset comprised of features mentioned in Figure 11.

6.2 Incorporating Existing Models

Our methodology is platform agnostic. We also plan to extend our methodology by implementing existing work in fake news detection, clickbait detection, and sentiment analysis. Therefore, in the future, our system would be able to take a single article as input, trace it back to its origin, and then classify the original article as fake news, clickbait, or neither. This integration would also allow any article along the propagation path to be categorized as fake news, clickbait, or neither. In the future, we hope to have a system which is able to tell the news consumer where the article originated, if that source can be trusted, and what the propagation path looks like.

7 Discussion and Conclusion

News syndication is a common phenomenon across the internet and it does not seem like it is going to get better anytime soon. After all, the goal of news websites is to publish as much news as possible, and to gain as much web traffic as possible to their websites. That is why it comes as no surprise that sites do not do their due diligence when it comes to verification of facts. The way to combat this worrisome matter is to automate the process of analyzing syndicated news and identifying the propagation path of viral articles. That is, tracing back an article to its origin and analyzing how article features change over time.

When a news consumer reads an article on a news site, he has no way of knowing the origins of the story or what steps that news site took to verify the authenticity of the article contents before publishing it. It is only fair that the consumer would have some information about the origin of the story, and that is exactly what our work achieves. That way, if the article originated from a credible news source, the news consumer could read the article confidently knowing that it is true. On the other hand, if there is no information about the article coming from a credible news source (i.e. the first time the story was published on the web was on a social media platform or a unknown news source), the consumer can take everything they read with a grain of salt, and not take every single statement that is said in the article as fact.

References

[1] Constanze Stelzenmüller. The impact of russian interference on germany's 2017 elections. *Testimony before the US Senate Select Committee on Intelligence June*, 28, 2017.

[2] Pardis Pourghomi, Fadi Safieddine, Wassim Masri, and Milan Dordevic. How to stop spread of misinformation on social media: Facebook plans vs. right-click authenticate approach. In *2017 International Conference on Engineering & MIS (ICEMIS)*, pages 1–8. IEEE, 2017.

[3] Davey Alba and Adam Satariano. At least 70 countries have had disinformation campaigns, study finds, Sep 2019.

[4] Robert S Mueller. Report on the investigation into russian interference in the 2016 presidential election. *US Dept. of Justice. Washington, DC*, 2019.

[5] Fake news | definition in the cambridge english dictionary, Dec 2019.

[6] Explained: What is fake news? social media and filter bubbles, Dec 2019.

[7] Flamini Funke. A guide to anti-misinformation actions around the world, 2019.

[8] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 2019.

[9] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[10] "fake news": Disinformation, misinformation and mal-information.

[11] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[12] Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining*, pages 230–239. IEEE, 2014.

[14] Youzhong Wang, Daniel Zeng, Xiaolong Zheng, and Feiyue Wang. Propagation of online news: dynamic patterns. In *2009 IEEE International Conference on Intelligence and Security Informatics*, pages 257–259. IEEE, 2009.

[15] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202. ACM, 2018.

[16] Qiuqing Tai. China's media censorship: A dynamic and diversified regime. *Journal of East Asian Studies*, 14(2):185–210, 2014.

[17] Michael Kearney. Trusting news project report 2017. *Reynolds Journalism Institute*, 2017.

[18] Rob Copeland. Wsj news exclusive | google's 'project nightingale' gathers personal health data on millions of americans, Nov 2019.

[19] Reuters. Google, clayton-based ascension ink deal to gather health data of millions of americans, Nov 2019.

[20] Brandy Betz. Google collecting health data without patient knowledge - wsj, Nov 2019.

[21] Google's health initiative collects personal data - cnn video, Nov 2019.

[22] Explore extract: https://www.washingtonpost.com/nation/2019/11/04/nazi-costume-utah-elementary-school-creekside/ | embedly. `https://embed.ly/docs/explore/extract?url=https%3A%2F%2Fwww.washingtonpost.com%2Fnation%2F2019%2F11%2F04%2Fnazi-costume-utah-elementary-school-creekside%2F`. (Accessed on 12/07/2019).

[23] Google news api - an api to search articles from google news. `https://gnews.io/`. (Accessed on 12/07/2019).

[24] News api - a json api for live news and blog articles. `https://newsapi.org/`. (Accessed on 12/07/2019).

[25] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

[26] stanfordnlp. CoreNLP, Dec 2019. [Online; accessed 7. Dec. 2019].

[27] reddit.com: api documentation. `https://www.reddit.com/dev/api/`. (Accessed on 12/07/2019).

[28] Sharedcount. https://www.sharedcount.com/. (Accessed on 12/07/2019).

[29] Text similarities : Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894. (Accessed on 12/07/2019).

[30] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ, 2003.

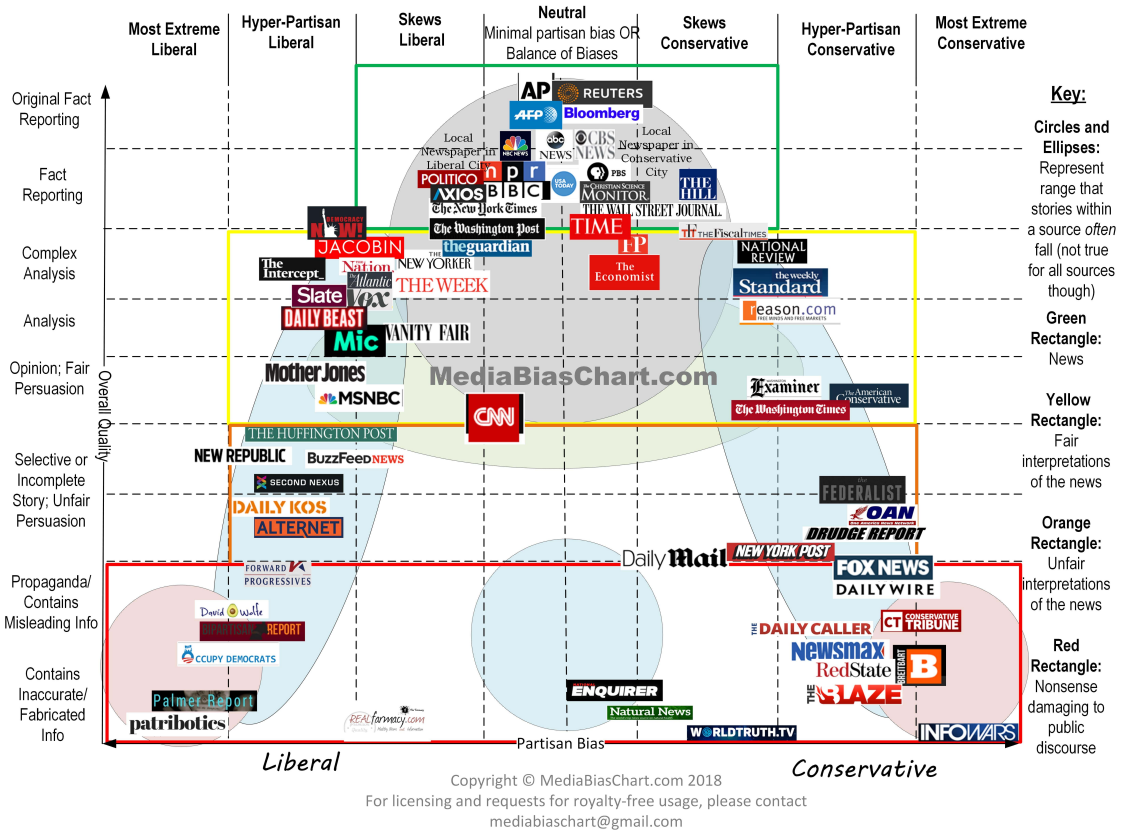
[31] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

A Initial Dataset Used for Phase I Analysis

Media Outlet	Article	Comments
Patribotics	Dear Mr. Putin, Let's Play Chess	148
Patribotics	Wikileaks is Connected to Russia – Despite Their Claims	42
Patribotics	Planespotting: Michael Cohen's Amazing Journey	49
Patribotics	Putin's Hacker, Wikileaks Host Pyotr Chayanov, Hacked America's Vote System And the DNC	11
Patribotics	Wikileaks Hands "Keys" to Putin's Russian Hacker – Readers, Leakers Tracked	19
InfoWars	De-Dollarization: Europe Joins The Party	6
InfoWars	Unemployment Falls to Lowest Level Since 1969	13
InfoWars	China Unveils 'doomsday Bomb' While U.s. Military Concentrates on "diversity"	83
InfoWars	'no Precedent in Human Experience': Study Finds Nuclear War Between India and Pakistan Could Leave 125 Million Dead	6
InfoWars	China Reveals New Photos of Strange Substance From Dark Side of Moon	26
Daily Caller	Iranian Foreign Minister Uses Instagram To Resign	22
Daily Caller	Israel Holding Early Elections As Bribery Allegations Engulf Netanyahu	12
Daily Caller	Turkey's President Arrests More Than 100 People For Connections To Failed 2016 Coup	4
Daily Caller	Saudi Crown Prince Fires Entertainment Chief Because Of Tightly Clad Female Circus Performers	4
Daily Caller	Hong Kong Police Unload Live Rounds On Protesters, Shoot 18-Year-Old: Report	23
Daily KOS	How to Support the Hong Kong Protesters	2
Daily KOS	Who knew? Ukraine-gate is actually a Rick Perry crime spree!	76
Daily KOS	NYT: Second Ukraine-related whistleblower may soon come forward	96
Daily KOS	Have we learned nothing about wars in the Middle East?	119
Daily KOS	Open thread for night owls: 'Itching for a War' with Iran	93
Daily Mail	Russia is helping China build a new missile attack warning system in 'response' to US plans to deploy missiles in Asia	44
Daily Mail	British sausage makers claim nation's bangers are under threat from pork shortage in China that has seen prices rise by 45 per cent	184
Daily Mail	Thousands of pro-democracy activists rally in Hong Kong ahead of four days of protests to overshadow anniversary celebrations in Beijing	0
Daily Mail	Thirteen and a half tonnes of gold worth up to £520million is found in a corrupt Chinese official's home and £30BILLION in suspected bribe money in his bank account	451
Daily Mail	'Most-wanted' Chinese fugitive, 63, hides in a cliff-side cave for 17 YEARS after escaping from prison	6
The Washington Times	Man pulls gun in road rage incident over Elizabeth Warren sticker, police say	2
The Washington Times	Man pulls gun in road rage incident over Elizabeth Warren sticker, police say	5
The Washington Times	Man pulls gun in road rage incident over Elizabeth Warren sticker, police say	124
The Washington Times	FBI runs Russian-language Facebook ads asking for help neutralizing 'hostile foreign intelligence'	1
The Washington Times	FBI runs Russian-language Facebook ads asking for help neutralizing 'hostile foreign intelligence'	0
Mother Jones	It's No Coincidence That the Top Presidential Candidates Are All So Old	6
Mother Jones	Columnist at the Center of Ukraine Scandal Joins Fox News	5
Mother Jones	Microsoft Says Iranian Hackers Are Targeting a 2020 Presidential Campaign	9

Media Outlet	Article	Comments
Mother Jones	Researchers Assembled over 100 Voting Machines. Hackers Broke Into Every Single One.	7
Mother Jones	The Biden Campaign Is Demanding That TV Execs Stop Booking Guiliani	69
Reason	Supreme Court Will Finally Hear Arguments Over Federal LGBT Discrimination Protections	100
Reason	China Banned South Park After the Show Made Fun of Chinese Censorship	105
Reason	The NBA Cares More About Making Money in Mainland China Than Supporting Freedom in Hong Kong	94
Reason	The U.K. Must Ban Pointy Knives, Says Church of England	76
Reason	The New York Times Says 'Free Speech Is Killing Us.' But Violent Crime Is Lower Than Ever.	120

B Media Bias Diagram



Media Bias Diagram

C Chinese Gold Story Dataset

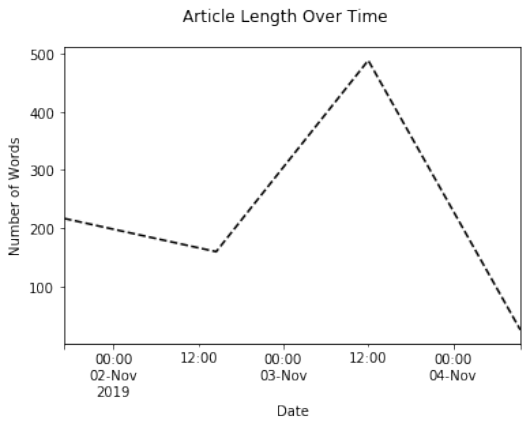
Media Outlet	Article Title	Date of Publication
Twitter	N/A (Twitter post by user @h1300062810)	. 09-24-2019 10:09
PowerApple	Secretary of Haikou copied 13.5 tons of cash, booked 268 billion in gold (translated)	09-26-2019 08:08
CrimeRussia	Chinese official hides 13 tons of gold in basement	09-26-2019 09:50
MenaFN	13.5 Tons Of Gold Found In Chinese Ex Mayors Basement	09-26-2019 18:20
Novinite	13 Tonnes of Gold Found in the Basement of Former Chinese Mayor	09-27-2019 13:49
RT	13.5 TONS of gold found piled in Chinese ex-governor's home	10-01-2019 13:37
Daily Star	13.5 tons of gold and $37billion cash found during police raid on mayor in China	10-01-2019 18:21
Daily Mail	Thirteen and a half tonnes of gold worth up to £520million is found in a corrupt Chinese official's home and £30BILLION in suspected bribe money in his bank account	10-02-2019 04:54
Mirror	Corrupt Chinese official found with £520million worth of gold bullion in home	10-03-2019 02:01

D University of Virginia Story Dataset

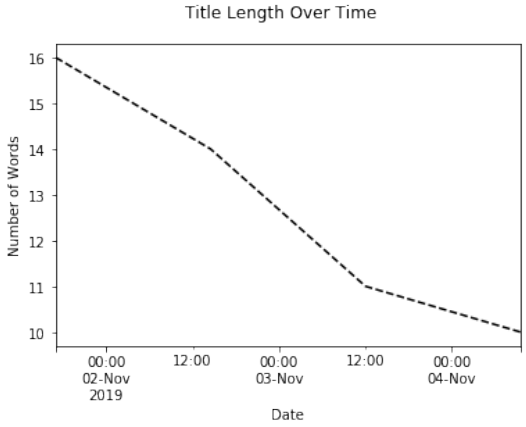
Media Outlet	Article Title	Date of Publication
The Daily Progress	Opinion/Letter: UVa should rethink Veterans Day decision	11-06-2019 09:26
WHSV	VA reinstates 21-gun salute on Veterans Day after wide-scale backlash	11-08-2019 14:11
The College Fix	Citing 'gun violence,' UVA cancels 21-gun salute portion of Veterans Day ceremony	11-11-2019 05:05
The Christian Perspective	University of Virginia Cancels 21-Gun Salute to Appease Snowflake Students	11-11-2019 13:15
Washington Examiner	University of Virginia ending 21-gun salute over potential 'panic' by students	11-11-2019 13:20
The Hill	University of Virginia cancels 21-gun salute to veterans over concern it might 'cause a panic'	11-11-2019 18:41
Fox News	University of Virginia cancels 21-gun salute from Veterans Day ceremony	11-11-2019 20:42
RT	University of Virginia excoriated for ditching Veterans Day 21-gun salute 'because gun violence'	11-11-2019 21:59
Daily Caller	Students Force University Of Virginia To End 21-Gun Salute On Veterans Day	11-11-2019 22:14

E Figures From Phase II Analysis

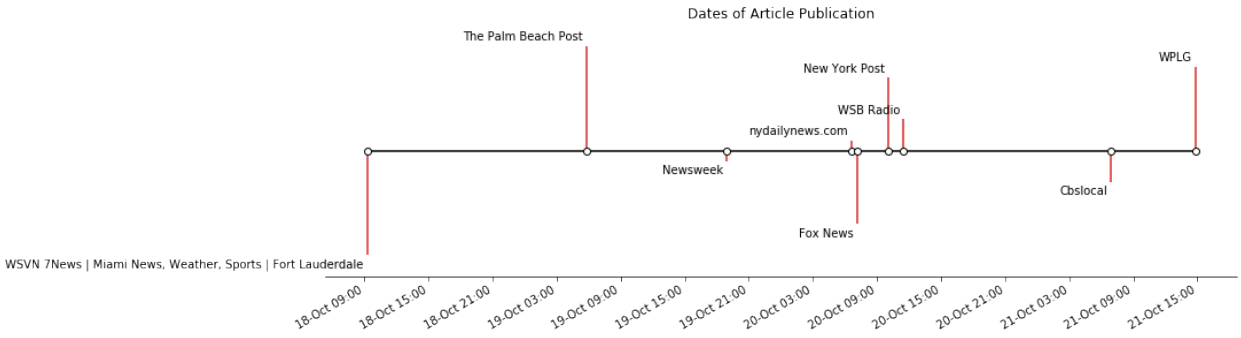
Timeline of Utah Halloween Story



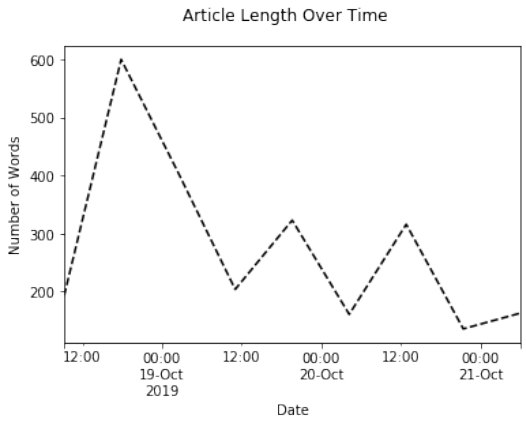
Utah Halloween Article Length Over Time



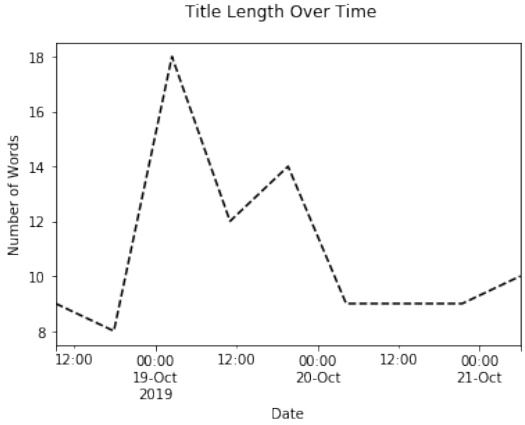
Utah Halloween Title Length Over Time



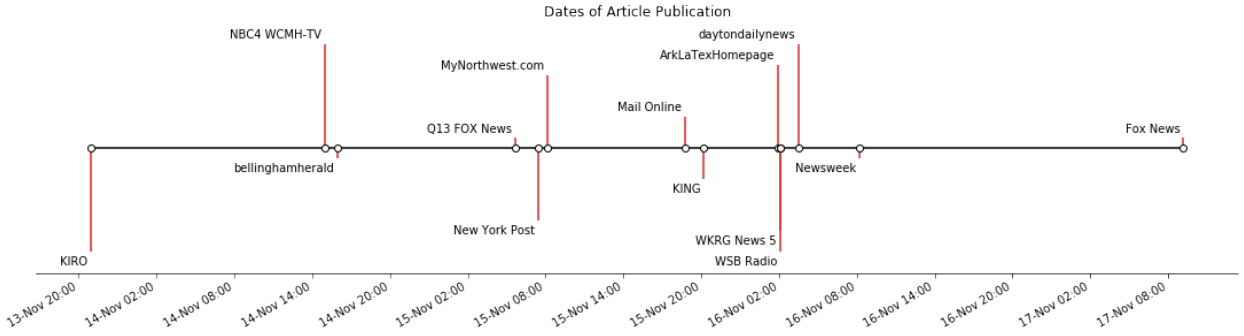
Timeline of Florida Man Story



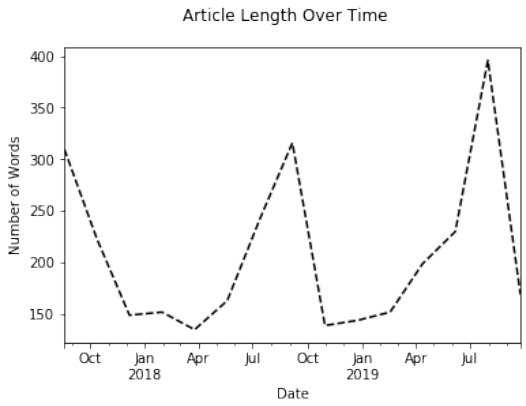
Florida Man Article Length Over Time



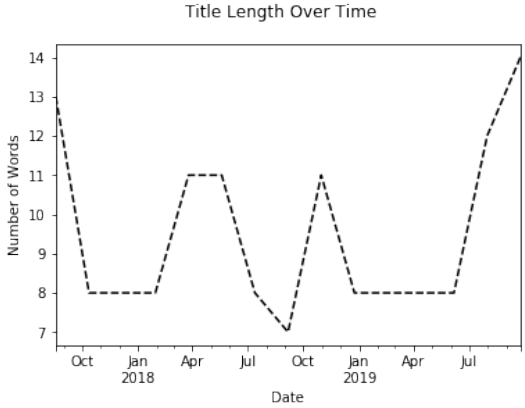
Florida Man Title Length Over Time



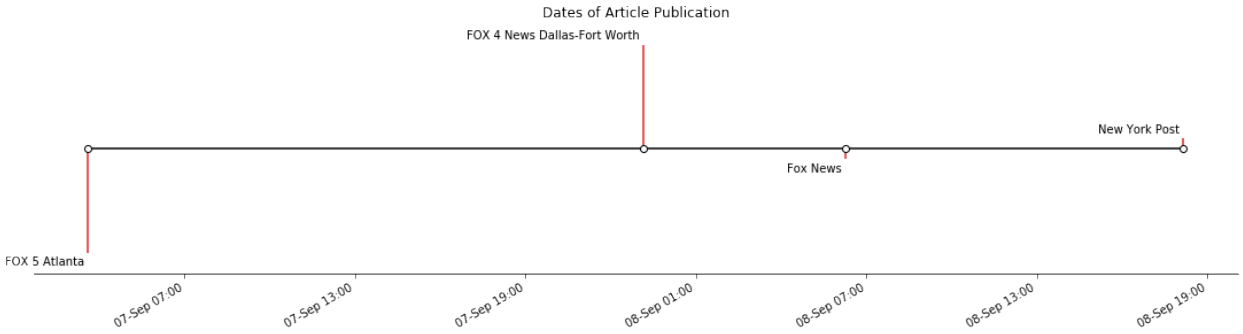
Timeline of Seattle Police Story



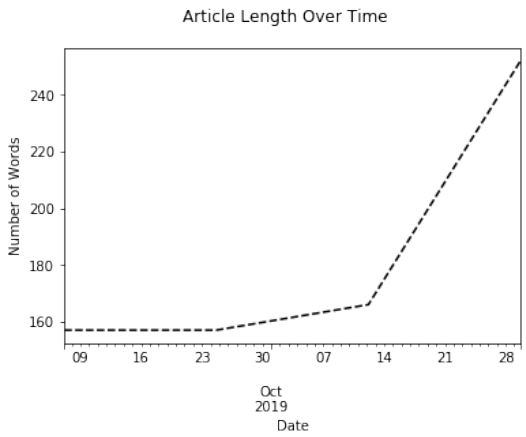
Seattle Police Article Length Over Time



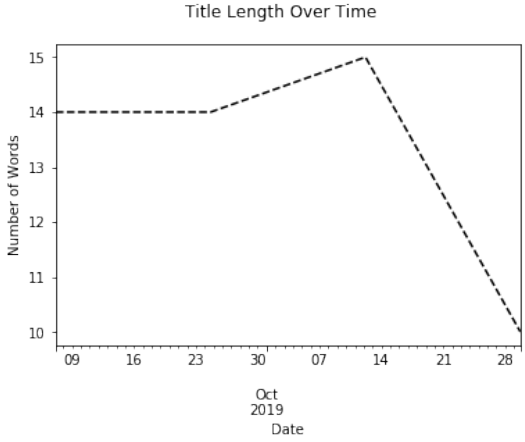
Seattle Police Title Length Over Time



Timeline of Spending Spree Police Story

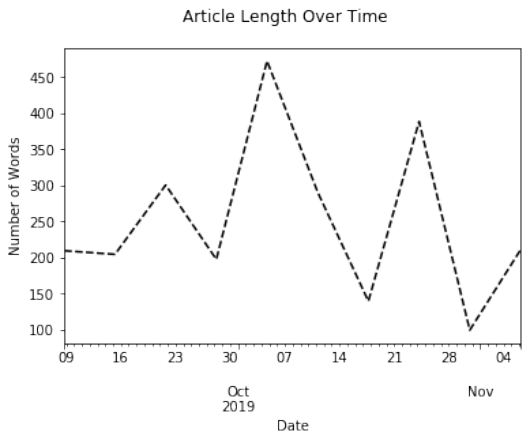


Spending Spree Article Length Over Time

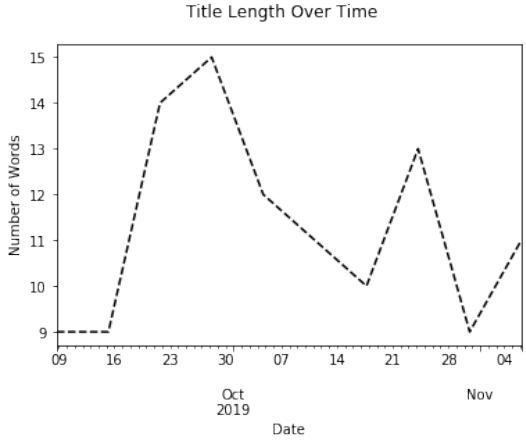


Spending Spree Title Length Over Time

Timeline of Chinese Burn Police Story



Chinese Burn Article Length Over Time



Chinese Burn Title Length Over Time

F Contributions of Each Group Member

Gaurav Deshpande: Wrote code for web scraping, article classification and clustering. Collected dataset used for phase II.

Alon Peer: Collected data for phase I, conducted manual propagation tracing, conducted analysis for phase I.

Sam Teplov: Wrote code for data analysis, conducted data analysis, analyzed patterns between different propagation paths.