Using Machine Learning to Identify Fake Political News

By

Peter D. Kelley

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In

Data Science

The University of Wisconsin – Green Bay

DS 785: Capstone Project Paper

Green Bay, Wisconsin

December 2021

ABSTRACT

Using Machine Learning to Identify Fake Political News

Peter D. Kelley

Fake news is a growing issue and is made worse because online fake news propagates “significantly farther, faster, deeper, and more broadly than the truth in all categories of information (Vosoughi, Roy, & Aral, 2018).” This project was a case study that successfully showed how machine learning (ML) algorithms can help identify fake vs. reliable or real news. The tested theory is that fake news differs from real news in several statistically significant ways. If it is true that fake news has patterns that differentiate it from real news, then ML algorithms can detect that difference and help people have more confidence in what they are reading.

The training and testing portion of the project incorporates serval stages to create the final machine learning algorithms that score the truthfulness of news articles. The first stage is data intake, scraping data off URLs from the secondary training and testing datasets. The next stage is data cleaning and pre-processing. The cleaned pre-processed data has two text analysis paths. One supports a linguistic base analysis, and the other supports a term based analysis. The cleaned and pre-processed data from both approaches are used to train and test machine learning algorithms on the article title and then again on the article body.

The user interface portion allows a user to supply news article URLs, cleans and pre-processes the title and body text of the news article, runs the result through the training ML algorithms, and finally provides the user with a dashboard readout.

The code and data used in the project can be found at <https://github.com/petedk/UWCapstone>.

Keywords: data science, fake news, text analysis, URL scraping, machine learning

TABLE OF CONTENTS

[ABSTRACT iii](#_Toc89608522)

[TABLE OF CONTENTS iv](#_Toc89608523)

[LIST OF TABLES vi](#_Toc89608524)

[LIST OF FIGURES vi](#_Toc89608525)

[CHAPTER 1 1](#_Toc89608526)

[INTRODUCTION 1](#_Toc89608527)

[1.1 Project Background 1](#_Toc89608528)

[1.2 Objectives 2](#_Toc89608529)

[CHAPTER 2 2](#_Toc89608530)

[LITERATURE REVIEW 2](#_Toc89608531)

[2.1 Automated Fake News Detection: 3](#_Toc89608532)

[2.1.1 Knowledge-Based Fake News Detection 3](#_Toc89608533)

[2.1.2 Style-Based Fake News Detection 4](#_Toc89608534)

[2.1.3 Propagation-Based Fake News Detection 5](#_Toc89608535)

[2.1.4 Source-Based Fake News Detection 5](#_Toc89608536)

[CHAPTER 3 6](#_Toc89608537)

[METHODOLOGY 6](#_Toc89608538)

[3.1 Data Source 8](#_Toc89608539)

[3.1.1 Media Bias and Accuracy Data Source 8](#_Toc89608540)

[3.1.2 Kaggle Competition Data Source I 8](#_Toc89608541)

[3.1.3 Kaggle Competition Data Source II 9](#_Toc89608542)

[3.1.4 FakeNewsNet Data Source 9](#_Toc89608543)

[3.1.5 Newsbot-Master Data Source 9](#_Toc89608544)

[3.2 Webpage Scraping 9](#_Toc89608545)

[3.3 Text Pre-Processing 10](#_Toc89608546)

[3.3.1 Text Cleaning 11](#_Toc89608547)

[3.3.2 Text Stemming and Lemmatization 11](#_Toc89608548)

[3.3.3 Term Dictionary 12](#_Toc89608549)

[3.3.4 Term Frequency Inverse Document Frequency (TFIDF) 12](#_Toc89608550)

[3.3.4.1 Term Frequency (TF) Dictionary. 13](#_Toc89608551)

[3.3.4.2 Inverse Document Frequency (IDF) Dictionary. 13](#_Toc89608552)

[3.3.4.3 Term Frequency Inverse Document Frequency (TFIDF) Dictionary. 14](#_Toc89608553)

[3.4 Linguistic Inquiry and Word Count Matrix (LIWC) 15](#_Toc89608554)

[3.5 Machine Learning Algorithms 15](#_Toc89608555)

[3.6 User Supplied URL Intake 16](#_Toc89608556)

[3.7 Dashboard Construction 16](#_Toc89608557)

[CHAPTER 4 16](#_Toc89608558)

[RESULTS 16](#_Toc89608559)

[4.1 Technical Results 16](#_Toc89608560)

[4.1.1 Feature Analysis 16](#_Toc89608561)

[4.1.2 Machine Learning Analysis 21](#_Toc89608562)

[4.1.2.1 Article Body Term Analysis. 21](#_Toc89608563)

[4.1.2.2 Article Title Terms. 21](#_Toc89608564)

[4.1.2.3 Article Body LIWC. 22](#_Toc89608565)

[4.1.2.4 Article Title LIWC. 23](#_Toc89608566)

[4.2 Project Output 23](#_Toc89608567)

[4.3 Algorithms 24](#_Toc89608568)

[4.4 Dashboard Graphs 25](#_Toc89608569)

[4.5 Dashboard Linguistic List 26](#_Toc89608570)

[4.5 Dashboard Examples 26](#_Toc89608571)

[CHAPTER 5 39](#_Toc89608572)

[CONCLUSION 39](#_Toc89608573)

[5.1 Conclusion 39](#_Toc89608574)

[5.2 Future work 39](#_Toc89608575)

[APPENDIX A 45](#_Toc89608576)

LIST OF TABLES



LIST OF FIGURES



CHAPTER 1

INTRODUCTION

1.1 Project Background

The internet and social media have facilitated the ability of people to access and share information; unfortunately, not all of the information is reliable. The following definition is used for this paper: fake political news are “articles that are intentionally and verifiably false, and could mislead readers” (Allcott & Gentzkow, 2017). By design, this definition excluded other types of false news that aren’t intentionally false, like reporting errors. It also excluded content that wasn’t created to mislead, like satire.

The risk of sharing false or fake information increases during times of uncertainty when public information is in high demand and facts are uncertain (Jang & Kim, 2018). People are not always good at spotting fake news. Many factors reduce a person’s ability to distinguish real news from fake news. Exposure to a fake news item increases the likelihood a reader will believe the same false information the next time they come across it (Martel, Pennycook, & Rand, 2020). The intensity of emotions felt while reading news stories reduces the ability to detect fake news (Martel, Pennycook, & Rand, 2020). Additionally, a person’s priory beliefs about the topic in the news story decrease the likelihood they will recognize fallacies when those fallacies support their prior beliefs (Allcott & Gentzkow, 2017).

Fake news is a significant threat to democracy and corrosive influence on a population’s trust in the government (Zhou & Zafarani, 2020). Fake news diminishes our ability to have constructive public dialog when even what is true is being debated. A healthy functioning democracy depends on citizens’ ability to access and identify factual information and make decisions on that information.

This project showed how machine learning (ML) algorithms can help identify fake vs. reliable or real news. The theory was that fake news differs from real news in several statistically significant ways. The theory further predicted that fake news would be less nuanced, more emotional, and have less word variation within the article. If it is true that fake news has patterns that differentiate it from real news, then ML algorithms can detect that difference and help people have more confidence in what they are reading.

1.2 Objectives

This project developed a tool to help people differentiate real vs. fake news. The following steps were used to accomplish this:

* Built a user interface to supply a URL that contains political news
* Built a data pipeline to scrap the visible text from the webpage and pre-processed it into clean, structured data for ingestion by ML algorithms
* Trained multiple ML algorithms on the structured data
* Combined the output of multiple ML algorithms to score the reliability of the article across several relevant dimensions
* Built a user-friendly scorecard showing the reliability aggregate score and constituent scores

CHAPTER 2

LITERATURE REVIEW

Over the last few years, there has been significant academic interest in the topic of fake news based on the number of scholarly articles written on the subject, and general interest in the topic as evidenced by the term being the 2016 Macquarie Dictionary’s word of the year (Sample, et al., 2020). The problem of identifying and protecting society from online fake news has proven to be complicated. It is partially due to the design and delivery of intentionally deceptive news using a combination of “psychology, sociology, political science, economics, linguistics, marketing, and fine arts (Sample, et al., 2020)” to maximize its effectiveness. The scope of this review was limited to literature focused on using ML to identify fake news.

2.1 Automated Fake News Detection:

Zhou & Zafarani (2020) list four strategies for automated detection of fake news:

* Knowledge-based fake news detection
* Style-based fake news detection
* Propagation-based fake news detection
* Source-based fake news detection

2.1.1 Knowledge-Based Fake News Detection

Knowledge-based fake news detection is the process of identifying a verifiable claim or fact and then checking that fact (Zeng, Abumansour, & Zubiaga, 2021). Automated fact-checking is a tool that can aid human fact-checkers by automating some of the easier and more straightforward fact-checking to free up humans to do more nuanced and labor-intensive work (Zeng, Abumansour, & Zubiaga, 2021). Fact-checking pipelines come in many forms, but they share the high-level steps of claim detection and validation (Zeng, Abumansour, & Zubiaga, 2021). Claims are often represented as tuples of SPOs (Subject, Predicate, Object) (Zhou, Jain, Phoha, & Zafarani, 2020). For example, the claim that Neil Armstrong was an astronaut would become the tuple (“Neil Armstrong”, “was an”, “astronaut”) and could be compared to a knowledge base of trusted information. Knowledge-based strategies face several challenges, including how to deal with a limited ability to span knowledge domains (Wright & Augenstein, 2020), redundant information, dependent facts (i.e., water is liquid but only in limited temperature and pressure ranges), knowledge base conflicts, low credibility sources, and incomplete data (Zhou & Zafarani, 2020).

2.1.2 Style-Based Fake News Detection

Style-based fake news detection does not directly assess whether the facts are true but instead looks at how the article was written. The assumption is that fake news seeks to mislead while real news creators seek to inform readers. This difference in intentions reveals itself in a measurable way (Zhou & Zafarani, 2020). The elements of the written document, such as the lexicon, syntax, discourse, and semantic characteristics, are transformed into ML features (Zhou & Zafarani, 2020). Examples of lexicon features include the words, frequency, and relative frequency of terms used in an article (Zhou, Jain, Phoha, & Zafarani, 2020). Examples of syntax features include the part-of-speech tags and new studies investigating the frequency of productions or rewrite rules based on Probability Context-Free Grammar Parsing Trees (Zhou, Jain, Phoha, & Zafarani, 2020). Examples of discourse features include discourse cue, semantic similarity, organization, and syntactic of the article as a whole (Wei & LiZhen, 2020). Examples of semantic features include the diversity, informality, sentiment, and quality of the words used (Zhou, Jain, Phoha, & Zafarani, 2020). One of the challenges with using style strategies to identify fake news is that fake news styles may vary with respect to time, media, domain, and language (Zhou & Zafarani, 2020). Another challenge is that as algorithms get better at detecting fake news based on style, bad actors will be incentivized to change their style to avoid their work being identified as fake (Zhou & Zafarani, 2020).

2.1.3 Propagation-Based Fake News Detection

Propagation-based fake news detection is a graphical network approach. The propagation network can be modeled as either News Cascades or Self-Defined Propagation Graphs (Zhou & Zafarani, 2020). The News Cascade approach graphs the news article from the initiator to each person (node) in the propagation chain (Zhou & Zafarani, 2020). This graph can measure the number of people who receive the article, the total and mean number of people forwarding the article, the total and mean time interval between the initial post and last share (Zhou & Zafarani, 2020). The Self-Defined Propagation approach is very flexible and can include one or more node types and edge types (Zhou & Zafarani, 2020). Node types include authors, publishers, news items, and users. Edge types include information-sharing direction, supporting vs. opposing a news item, and social relations between network nodes (Zhou & Zafarani, 2020). Online fake news propagates “significantly farther, faster, deeper, and more broadly than the truth in all categories of information (Vosoughi, Roy, & Aral, 2018).” Studies show the expanded reach of fake news over real news was driven by human peer-to-peer sharing and not automated robot influences (Vosoughi, Roy, & Aral, 2018). A significant challenge with propagation detection methods is that they are retrospective and have not shown an ability to predict fake news before the news has propagated across social networks (Zhou, Jain, Phoha, & Zafarani, 2020).

2.1.4 Source-Based Fake News Detection

Source-based fake news detection uses the previous work of the creators of the news article or the history of the accounts sharing the information and is, therefore, an indirect indicator of an article’s credibility or veracity (Zhou & Zafarani, 2020). Even though it is wrong to assume that all future news from low-credibility sources will be fake news, it is still a useful tool for identifying potentially unreliable news (Zhou & Zafarani, 2020). Credibility can be thought of as a way for individuals to determine the likelihood a news item is authentic before making that determination themselves. (Viviani & Pasi, 2017). One advantage of source-based content credibility scoring is that it can be done in real-time if the source has a score. Optimal fact-checking happens at the point of content consumption, not on a different site or at a later time (Kim, Moravec, & Dennis, 2019). Source and account credibility scores can be affixed to the article when the article is posted, and every reader can make more informed decisions. It has been shown that source rating influences what users believe and read online (Kim, Moravec, & Dennis, 2019). A source-based fake news detection challenge is maintaining an up-to-date repository of all news content creators, users, and associated credibility.

CHAPTER 3

METHODOLOGY

This chapter outlines the process of transforming a user-supplied online news article into a user-friendly dashboard with a reliable score and supporting evidence for that score. The development and final procedure are illustrated with process maps and explained in detail in the following sections.

**Fig 3.1**

Process Development Map

**Diagram

Description automatically generated**

**Fig 3.2**

Final Process Map with User Input and Dashboard

**Diagram

Description automatically generated**

3.1 Data Source

All of the data sources used in this project are secondary data sources, as described below.

3.1.1 Media Bias and Accuracy Data Source

The Media Bias and Accuracy dataset (Milligan, 2021) consists of 223 website domains that contain online news articles. The dataset provides three measurements for each site. The bias score indicates how politically biased the content on the site is. The reliability score suggests the reliability of the information in the articles on the site. The final measurement is how much traffic the site received during February 2018.

The Media Bias and Accuracy dataset is a constituent of the final reliability calculation and is presented in the dashboard as a supporting measure.

3.1.2 Kaggle Competition Data Source I

The dataset for the “Fake and Real News Dataset Classifying the News” competition (Bisaillon, 2020) contains news article records broken up into two files. One file for real news and the other for fake news. The records contain both the title and body text. The dataset does not contain news article URLs. Ideally, all of the training and testing data would have included URLs of the articles to match the user-supplied URL process. Labeled datasets for fake news detection with URLs was limited, and inclusion of these non-URL datasets was done to increase the size of the training and testing dataset.

3.1.3 Kaggle Competition Data Source II

The dataset for the “Fake News Detection” competition (Kaggle, 2017) contains news article records. The records include the URL, title, and body text.

3.1.4 FakeNewsNet Data Source

The FakeNewsNet dataset (Wales, 2019) (Shu, Mahudeswaran, Wang, Lee, & Lie, 2018) (Shu, Wang, & Liu, 2017) (Shu, Sliva, Wang, tang, & Liu, 2018) contains news article records broken up into two files: one file for real news and the other for fake news. The sample news articles are taken from a PolitiFact database. These records contain the news article URL. The datasets were two of the datasets used in an article about fake news detection (Wales, 2019)

3.1.5 Newsbot-Master Data Source

The Newsbot-Master dataset (Wales, 2019) contains news article records. The sample news articles are taken from a PolitiFact database. These records include the news article URL. The dataset is identified in the documentation for the fake news detection website <https://www.unslanted.net/newsbot/>.

3.2 Webpage Scraping

URL scraping was done to extract the visible text of the article for datasets that included the URL. The extracted text was saved as a text file for both the title and body of the article. As noted earlier, not all training datasets included URLs. The title and body of articles that did not have URLs are also saved as text files.

A few Python libraries were considered when building the web scraping logic. The first library was BeautifulSoup, one of the most common web scraping packages for Python (Zafra, 2019). The library is easy to use and has proved useful in previous projects. The downside to this library is that logic branches had to be built to accommodate the different HTML page layouts. The different approaches to laying out HTML pages change how the body and title of the article are identified and scrapped. This limitation was acceptable for the training and testing web pages, but this approach did not account for how user-supplied pages might be coded. The Python library Newspaper3k was also investigated, which was built specifically for website article scraping and curation (Helstowski, 2021). This library’s advanced algorithms significantly improved the accuracy of the scrapped document text but did not eliminate errors. The final approach was to extract the article title and body with the Newspaper3K library when possible and otherwise use an alternative logic path with BeautifulSoup. The first step proved to be fast and very good at getting the desired information from the vast majority of available web pages. The alternative path successfully extracted article content from half of the web pages that failed the first path. The most common reason a web page could not be scrapped was that it was no longer a live page, or it employed logic to refuse HTML requests from automated jobs like this one.

3.3 Text Pre-Processing

Free-form text found in writing is messy, unstructured data and doesn’t lend itself to algorithms designed to work in structured data (Brownlee, 2019). This section will highlight the steps and processes to convert the unstructured data of the news article into structured data.

Text pre-processing consisted of many steps and was required to transform data from diverse sources into consistent input for subsequent process steps. These steps included making sure that the body of the article had at least 50 words and that the website was an active site and not a dead or blocked address.

3.3.1 Text Cleaning

Text cleaning was done for all records and consisted of the following steps. Foreign characters and numbers were removed from the text. Then all characters were forced to lower case. This is all of the text cleaning required to support the Linguistic Inquiry and Word Count Matrix (LIWC) branch of the data pipeline. The term breach of the data pipeline required the additional steps of punctuation removal and stop words. Stop words are words that are so common that they provide little helpful information to a machine learning algorithm and are often dropped from natural language processes.

3.3.2 Text Stemming and Lemmatization

Text stemming and lemmatization are both approaches to text normalizing. Normalizing the text helps to remove word variation in the pre-processing steps of natural language analysis (Beri, 2020).

1)Text stemming reduces word variation by distilling words down to their base (Beri, 2020). This word variation can be due to the tense of the word or word derivatives. The stem of the words eating, eat, and eaten is eat. Stemming does not always use actual words as the stemmed base.

2) Text lemmatization is a more complicated but informative process that considers the part of speech a word is being used in and maps that to an actual word (Beri, 2020). An example of lemmatization would be to map geese to goose. The natural language tool kit stemming library nltk.stem.porter turns “geese” into “gees” and “goose” into “goos”.

This project used text stemming to reduce complexity in this initial proof of concept.

3.3.3 Term Dictionary

The term dictionary is a Bag of Words (BoW) is an easy-to-understand and implement method of document modeling (Brownlee, 2019). A term dictionary uses a term presence indicator instead of the count of words to build a dictionary or matrix. Each article has its own term dictionary that contains all of the terms found in all of the articles, also called a corpus, and indicates if a word is present or not in this document. Term presence vs. term count was used to reduce the impact of document size on the results.

An example using two documents.

doc\_1: My car hit another car on the farm.

doc\_2: My cows eat hay on my farm.

The corpus for doc\_1 and doc\_2 would be (car, another, farm, eat, cows, hay, hit) after the stop words of (my, on, the) are removed.

**Table 3.1**

Term Presence Example



3.3.4 Term Frequency Inverse Document Frequency (TFIDF)

Term Frequency Inverse Document Frequency (TFIDF) is one of the most widely used methods to convert unstructured text into structured data (Tripathi, 2018). The TFIDF is a method that scores the importance of a word in a given document relative to all the other documents in a collection. If a term is used more often in one or a subset of the documents than in the rest of the documents, this variation may be informative for an ML algorithm. Words used with the same frequency across most documents in a collection are likely to be uninformative for an ML algorithm. The TFIDF is calculated in three distinct steps:

3.3.4.1 Term Frequency (TF) Dictionary.

A TF dictionary has all of the terms in the corpus as dictionary keys and assigns the normalized term frequency for each term in a given document to the corresponding value. The count of, or frequency, a given term is used in a document divided by all the valid terms is used. Valid terms are those that survive any pre-processing steps like removal of stop words, stemming, or lemmatization, to name a few.

TF calculation:

**Table 3.2**

Term Frequency Example



Note. The total number of words in a document after stop words have been removed.

3.3.4.2 Inverse Document Frequency (IDF) Dictionary.

An IDF dictionary has all of the terms in the corpus as dictionary keys and assigns the inverse document frequency as the corresponding value. This calculation gives the highest weight to terms present in fewer documents, and the lowest weight to terms that show up in many documents.

IDF calculation:

**Table 3.3**

Inverse Document Frequency Example



Note. The word farm is an example of when a term is in many documents its weight in TF\*IDF diminishes. In the case of the term farm, it becomes zero because that term is in all documents.

3.3.4.3 Term Frequency Inverse Document Frequency (TFIDF) Dictionary.

A TFIDF dictionary has all of the terms in the corpus as dictionary keys and assigns the product of the TF and IDF as the corresponding value.

TF\*IDF calculation: TF x IDF

**Table 3.4**

Term Frequency Inverse Document Frequency Example



Note. The term car has the most weight; it is the only word used twice in one document while not present in the other. The term farm is present in all documents and therefore has a weight, or factor, of zero.

3.4 Linguistic Inquiry and Word Count Matrix (LIWC)

As a complement to the TF\*IDF approach to understanding how true and fake articles differ in how they are written, two linguistic inquiry and word count approaches were tried. The first path was a manually coded process that counted linguistic features, used libraries to determine parts-of-speech (Rachiele, 2018), reading level (Dattatreya, 2019) , and emotional tone of the text (Sen, 2020). This manual process worked but was complex and took a long time to run. The second approach was to use the LIWC2015 tool. This tool has over 90 output variables that cover categories such as text descriptors, linguistic dimensions, psychological constructs, and punctuation, to name a few (Pennebaker, Boyd, Jordan, & Blackburn, 2015). A complete list of the output variables are in the appendix. The tool analyzes each document independently of all the others and gives the document a score for each output variable. Most of the output variables range between 0 and 100, but a few, like total word count, and word count per sentence, have no upper limit.

The LIWC2015 tool was selected as part of the final process because it was much faster and produced more features than the manually coded approach.

3.5 Machine Learning Algorithms

Four machine learning algorithms were assessed in the train and testing phase of the project. These four algorithms were all from the sklearn family of libraries. The four algorithms also each represent a different approach to the classification problem. The first algorithm was a Logistic Regression, a generalized linear model. The second algorithm was a Random Forest Classifier, an ensemble learning method of decision trees. The third algorithm was a Support Vector Classifier (SVC), a margin-based classifier. The final algorithm was Neural Network Multi-layer Perceptron Classifier (MLP), a feedforward artificial neural network. Because one of the project’s primary goals is to explain why an article may or may not be fake news and due to how complex neural networks are to interpret, the MLP classifier was there as a point of comparison and not a viable final algorithm for the project.

3.6 User Supplied URL Intake

The URL intake process was a cell in a JupyterLab notebook.

3.7 Dashboard Construction

The dashboard was a collection of seven graphs and four lists of linguistic indicators copied from JupyterLab Notebook cells. The graphs include one article-level graph with the final results and six supporting charts showing the constituent dimensions used to calculate the article’s truthfulness. The four lists of linguistic indicators are a supporting and a counter-supporting list for both the article title and body.

CHAPTER 4

RESULTS

4.1 Technical Results

4.1.1 Feature Analysis

For the text content of the article, four lists were created. The lists were created by identifying the top 1000 words most likely to be found in real and fake news for both the article body and title. Below are two Wordclouds of the top 50 unique words in each list. The desire was that these lists of words would help support why an article was classified as true or fake. There is no apparent rationale for the correlation between the terms in these lists and an article being true or fake news: these lists have been dropped from the dashboard.

**Fig 4.1**

Wordcloud of Word More Likely to be in the Body of Fake News Articles

Text

Description automatically generated

**Fig.4.2**

Wordcloud of Word More Likely to be in the Body of Real News Articles

**Text

Description automatically generated**

For the LIWC analysis of the body and title of the article, the user was shown the top features that indicated the article was either real or fake.

Below are graphs illustrating the relationship between equally spaced bins for a sample of highly influential features and fake news percent. These graphs show the distribution of a feature vs. the population of articles with any feature level and the proportion of news articles in each bucket of fake news.

**Fig 4.3**

Article Body LIWC Sample of Features vs. Fake News Percentage

**Graphical user interface, chart

Description automatically generated**

**Fig 4.4**

Article Title LIWC Sample of Features vs. Fake News Percentage

**Graphical user interface, chart

Description automatically generated**

4.1.2 Machine Learning Analysis

4.1.2.1 Article Body Term Analysis.

The Matrix type varied between TF\*IDF and Term Presence. The pre-processing varied between using full text and stemming the text. The most accurate algorithm with text preparation for the body of the article was Logistic Regression using full words and a term presence matrix (see table 4.1).

**Table 4.1**

Article Body Word Matrix ML Algorithm Results



Note. The MLP Classifier is for comparison purposes only.

4.1.2.2 Article Title Terms.

The most accurate algorithm with text preparation for the article’s title was Random Forest Classifier with full words and a TFIDF matrix (see table 4.2).

**Table 4.2**

Article Title Word Matrix ML Algorithm Results



4.1.2.3 Article Body LIWC.

The Scaled variable varied between using the output variable as they were, which was indicated by a “False” and scaling the values between 0 and 1, which was indicated by a “True”. The most accurate algorithm with LIWC preparation for the body of the article was Random Forest Classifier with Non-Scaled LIWC results (see table 4.3).

**Table 4.3**

Article Body LIWC ML Algorithm Results



4.1.2.4 Article Title LIWC.

The most accurate algorithm with LIWC preparation for the body of the article was Random Forest Classifier with Non-Scaled LIWC results (see table 4.4).

**Table 4.4**

Article Title LIWC ML Algorithm Results



4.2 Project Output

This project output consisted of two parts. The first was a process for a user to input one or more URLs containing news articles. The second was a dashboard that gives the user information about how reliable each news article was.

4.3 Algorithms

Five stages went into building the data pipeline and ML algorithms to support the project (see table 4.5). The first stage takes the three training sets with URLs and brings them into a single uniform data frame. The second stage scraps webpages for records with URLs. It then cleans and saves all title and body text as a file to be processed by the LIWC2015 tool. The third stage is pre-processing for text analysis. The pre-processing saves files to support Word-Presence and TFIDF ML. The fourth stage is ML training and test. The fifth and final stage is where a user can supply new URLs with news articles and have them analyzed to see the output dashboard details.

**Table 4.5**

Process Steps



4.4 Dashboard Graphs

The graphs are the heart of the dashboard; they are easy to understand and contain both the top-line result as to the article truthfulness and the constituent factors that were considered to make that conclusion.

As stated above, the first graph shows the degree of truthfulness of the article. There are five discrete levels and an analog pointer. The five levels of truthfulness are very true, true, unknown, fake, and very fake. The second and third graph shows how politically bias and factually accurate past news stores have been. The rest of the graphs use the same five discrete levels and an analog pointer to illustrate the degree of truthfulness of the article and show the ML results of the Linguistic Analysis for both the body and title and the ML results of the Word Matrix Analysis for both the body and title.

4.5 Dashboard Linguistic List

The final sections of the dashboard shown to the user are the four linguistic lists that show which linguistic features support the article being true and which support it being fake news. This is done for both the body and title of the article. The lists consist of the top 25 features of importance in the machine learning algorithm and are present in the article’s title or body, respectively. Almost all articles will have linguistic features indicative of both true and fake news. The decision to present both supporting and counter-supporting information is to be transparent. The goal of the dashboard isn’t to tell users what to read and believe; it is to give them information to help them make more informed decisions on what to read and accept.

Below are the actual lists based on user-supplied URLs. Each list was copied from a JoupiterLab Notebook cell.

4.5 Dashboard Examples

Below are the actual graphs and lists based on example user-supplied URLs (Figures 4.5 – Fig 4.14).

**Fig 4.5**

Dashboard for True News with Accurate Algorithm Score

**Diagram

Description automatically generated**

**Fig 4.6**

Dashboard list for True News with Accurate Algorithm Score

List of top linguistic features in the body that indicate the article is true news.

-----------------------------------------------------------------------------------

The low use of “Colons” indicates True news.

The high use of “Hear” indicates True news.

The low use of “Words/sentence” indicates True news.

The low use of “Common Adverbs” indicates True news.

The low use of “Question marks” indicates True news.

The high use of “Analytical thinking” indicates True news.

The low use of “See” indicates True news.

The low use of “Exclamation marks” indicates True news.

The high use of “Words > 6 letters” indicates True news.

The high use of “Work” indicates True news.

The low use of “2nd person” indicates True news.

The low use of “All Punctuation” indicates True news.

The high use of “Relativity” indicates True news.

The low use of “Swear words” indicates True news.

List of top linguistic features in the body that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Total pronouns” indicates Fake news.

The high use of “Other punctuation” indicates Fake news.

The high use of “Total function words” indicates Fake news.

The high use of “Present focus” indicates Fake news.

The high use of “Impersonal pronouns” indicates Fake news.

The high use of “Informal language” indicates Fake news.

The high use of “Personal pronouns” indicates Fake news.

The low use of “Past focus” indicates Fake news.

The high use of “Certainty” indicates Fake news.

The low use of “Periods” indicates Fake news.

The high use of “Netspeak” indicates Fake news.

List of top linguistic features in the title that indicate the article is true news.

-----------------------------------------------------------------------------------

The low use of “Colons” indicates True news.

The low use of “Words/sentence” indicates True news.

The low use of “Question marks” indicates True news.

The low use of “Total pronouns” indicates True news.

The low use of “Other punctuation” indicates True news.

The low use of “See” indicates True news.

The low use of “Exclamation marks” indicates True news.

The low use of “Total function words” indicates True news.

The low use of “Impersonal pronouns” indicates True news.

The low use of “Informal language” indicates True news.

The high use of “Words > 6 letters” indicates True news.

The low use of “Certainty” indicates True news.

The low use of “2nd person” indicates True news.

The low use of “All Punctuation” indicates True news.

The low use of “Netspeak” indicates True news.

The low use of “Swear words” indicates True news.

List of top linguistic features in the title that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Hear” indicates Fake news.

The low use of “Common Adverbs” indicates Fake news.

The high use of “Analytical thinking” indicates Fake news.

The low use of “Present focus” indicates Fake news.

The low use of “Personal pronouns” indicates Fake news.

The high use of “Past focus” indicates Fake news.

The high use of “Work” indicates Fake news.

The high use of “Periods” indicates Fake news.

The high use of “Relativity” indicates Fake news.

**Fig 4.7**

Dashboard for Fake News with Accurate Algorithm Score

**A picture containing shape

Description automatically generated**

**Fig 4.8**

Dashboard list for Fake News with Accurate Algorithm Score

List of top linguistic features in the body that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Colons” indicates Fake news.

The low use of “Hear” indicates Fake news.

The high use of “Common Adverbs” indicates Fake news.

The high use of “Question marks” indicates Fake news.

The high use of “Total function words” indicates Fake news.

The high use of “Present focus” indicates Fake news.

The high use of “Impersonal pronouns” indicates Fake news.

The low use of “Past focus” indicates Fake news.

The high use of “Certainty” indicates Fake news.

The low use of “Periods” indicates Fake news.

The high use of “2nd person” indicates Fake news.

The low use of “Relativity” indicates Fake news.

List of top linguistic features in the body that indicate the article is true news.

-----------------------------------------------------------------------------------

The low use of “Words/sentence” indicates True news.

The high use of “Analytical thinking” indicates True news.

The low use of “Total pronouns” indicates True news.

The low use of “Other punctuation” indicates True news.

The low use of “See” indicates True news.

The low use of “Exclamation marks” indicates True news.

The low use of “Informal language” indicates True news.

The high use of “Words > 6 letters” indicates True news.

The low use of “Personal pronouns” indicates True news.

The high use of “Work” indicates True news.

The low use of “All Punctuation” indicates True news.

The low use of “Netspeak” indicates True news.

The low use of “Swear words” indicates True news.

List of top linguistic features in the title that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Colons” indicates Fake news.

The low use of “Hear” indicates Fake news.

The low use of “Analytical thinking” indicates Fake news.

The high use of “Impersonal pronouns” indicates Fake news.

The low use of “Past focus” indicates Fake news.

The low use of “Periods” indicates Fake news.

The high use of “All Punctuation” indicates Fake news.

List of top linguistic features in the title that indicate the article is true news.

-----------------------------------------------------------------------------------

The high use of “Words/sentence” indicates True news.

The high use of “Common Adverbs” indicates True news.

The high use of “Question marks” indicates True news.

The high use of “Total pronouns” indicates True news.

The high use of “Other punctuation” indicates True news.

The high use of “See” indicates True news.

The high use of “Exclamation marks” indicates True news.

The high use of “Total function words” indicates True news.

The high use of “Present focus” indicates True news.

The high use of “Informal language” indicates True news.

The low use of “Words > 6 letters” indicates True news.

The high use of “Personal pronouns” indicates True news.

The high use of “Certainty” indicates True news.

The low use of “Work” indicates True news.

The high use of “2nd person” indicates True news.

The low use of “Relativity” indicates True news.

The high use of “Netspeak” indicates True news.

The high use of “Swear words” indicates True news.

**Fig 4.9**

Dashboard for True News with AmbiguousAlgorithm Score

**Diagram

Description automatically generated**

**Fig 4.10**

Dashboard list for True News with Inaccurate Algorithm Score

List of top linguistic features in the body that indicate the article is fake news.

-----------------------------------------------------------------------------------

The low use of “Hear” indicates Fake news.

The high use of “Common Adverbs” indicates Fake news.

The low use of “Analytical thinking” indicates Fake news.

The high use of “Total pronouns” indicates Fake news.

The high use of “Other punctuation” indicates Fake news.

The high use of “Exclamation marks” indicates Fake news.

The high use of “Total function words” indicates Fake news.

The high use of “Present focus” indicates Fake news.

The high use of “Impersonal pronouns” indicates Fake news.

The high use of “Personal pronouns” indicates Fake news.

The low use of “Work” indicates Fake news.

The low use of “Periods” indicates Fake news.

The high use of “2nd person” indicates Fake news.

The high use of “All Punctuation” indicates Fake news.

The low use of “Relativity” indicates Fake news.

List of top linguistic features in the body that indicate the article is true news.

-----------------------------------------------------------------------------------

The low use of “Colons” indicates True news.

The low use of “Words/sentence” indicates True news.

The low use of “Question marks” indicates True news.

The low use of “See” indicates True news.

The low use of “Informal language” indicates True news.

The high use of “Words > 6 letters” indicates True news.

The high use of “Past focus” indicates True news.

The low use of “Certainty” indicates True news.

The low use of “Netspeak” indicates True news.

The low use of “Swear words” indicates True news.

List of top linguistic features in the title that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Colons” indicates Fake news.

The low use of “Hear” indicates Fake news.

The low use of “Analytical thinking” indicates Fake news.

The high use of “Present focus” indicates Fake news.

The low use of “Past focus” indicates Fake news.

The low use of “Periods” indicates Fake news.

The low use of “Relativity” indicates Fake news.

List of top linguistic features in the title that indicate the article is true news.

-----------------------------------------------------------------------------------

The high use of “Words/sentence” indicates True news.

The high use of “Common Adverbs” indicates True news.

The high use of “Question marks” indicates True news.

The high use of “Total pronouns” indicates True news.

The high use of “Other punctuation” indicates True news.

The high use of “See” indicates True news.

The high use of “Exclamation marks” indicates True news.

The high use of “Total function words” indicates True news.

The high use of “Impersonal pronouns” indicates True news.

The high use of “Informal language” indicates True news.

The low use of “Words > 6 letters” indicates True news.

The high use of “Personal pronouns” indicates True news.

The high use of “Certainty” indicates True news.

The low use of “Work” indicates True news.

The high use of “2nd person” indicates True news.

The high use of “All Punctuation” indicates True news.

The high use of “Netspeak” indicates True news.

The high use of “Swear words” indicates True news.

**Fig 4.11**

Dashboard for Fake News with Ambiguous Algorithm Score

**Diagram

Description automatically generated**

**Fig 4.12**

Dashboard list for Fake News with Neutral Algorithm Score

List of top linguistic features in the body that indicate the article is fake news.

-----------------------------------------------------------------------------------

The low use of “Analytical thinking” indicates Fake news.

The high use of “Total pronouns” indicates Fake news.

The high use of “See” indicates Fake news.

The high use of “Total function words” indicates Fake news.

The high use of “Present focus” indicates Fake news.

The high use of “Impersonal pronouns” indicates Fake news.

The low use of “Words > 6 letters” indicates Fake news.

The high use of “Personal pronouns” indicates Fake news.

The high use of “Certainty” indicates Fake news.

The low use of “Periods” indicates Fake news.

The low use of “Relativity” indicates Fake news.

List of top linguistic features in the body that indicate the article is true news.

-----------------------------------------------------------------------------------

The low use of “Colons” indicates True news.

The high use of “Hear” indicates True news.

The low use of “Words/sentence” indicates True news.

The low use of “Common Adverbs” indicates True news.

The low use of “Question marks” indicates True news.

The low use of “Other punctuation” indicates True news.

The low use of “Exclamation marks” indicates True news.

The low use of “Informal language” indicates True news.

The high use of “Past focus” indicates True news.

The high use of “Work” indicates True news.

The low use of “2nd person” indicates True news.

The low use of “All Punctuation” indicates True news.

The low use of “Netspeak” indicates True news.

The low use of “Swear words” indicates True news.

List of top linguistic features in the title that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Colons” indicates Fake news.

The low use of “Hear” indicates Fake news.

The low use of “Words > 6 letters” indicates Fake news.

The low use of “Past focus” indicates Fake news.

The low use of “Work” indicates Fake news.

The low use of “Periods” indicates Fake news.

The high use of “All Punctuation” indicates Fake news.

List of top linguistic features in the title that indicate the article is true news.

-----------------------------------------------------------------------------------

The high use of “Words/sentence” indicates True news.

The high use of “Common Adverbs” indicates True news.

The high use of “Question marks” indicates True news.

The low use of “Analytical thinking” indicates True news.

The high use of “Total pronouns” indicates True news.

The high use of “Other punctuation” indicates True news.

The high use of “See” indicates True news.

The high use of “Exclamation marks” indicates True news.

The high use of “Total function words” indicates True news.

The high use of “Present focus” indicates True news.

The high use of “Impersonal pronouns” indicates True news.

The high use of “Informal language” indicates True news.

The high use of “Personal pronouns” indicates True news.

The high use of “Certainty” indicates True news.

The high use of “2nd person” indicates True news.

The low use of “Relativity” indicates True news.

The high use of “Netspeak” indicates True news.

The high use of “Swear words” indicates True news.

**Fig 4.13**

Dashboard for Fake News with Accurate Algorithm Score And Missing Source Information

**A picture containing text

Description automatically generated**

**Fig 4.14**

Dashboard list for Fake News with Accurate Algorithm Score And Missing Source Information

List of top linguistic features in the body that indicate the article is fake news.

-----------------------------------------------------------------------------------

The high use of “Colons” indicates Fake news.

The high use of “Words/sentence” indicates Fake news.

The low use of “Hear” indicates Fake news.

The high use of “Common Adverbs” indicates Fake news.

The high use of “Other punctuation” indicates Fake news.

The high use of “Question marks” indicates Fake news.

The high use of “See” indicates Fake news.

The high use of “Informal language” indicates Fake news.

The low use of “Past focus” indicates Fake news.

The low use of “Work” indicates Fake news.

The low use of “Power” indicates Fake news.

The high use of “All Punctuation” indicates Fake news.

The low use of “Perceptual processes” indicates Fake news.

The low use of “Relativity” indicates Fake news.

List of top linguistic features in the body that indicate the article is true news.

-----------------------------------------------------------------------------------

The high use of “Analytical thinking” indicates True news.

The low use of “Total pronouns” indicates True news.

The low use of “Total function words” indicates True news.

The low use of “Personal pronouns” indicates True news.

The low use of “Exclamation marks” indicates True news.

The high use of “Words > 6 letters” indicates True news.

The low use of “Present focus” indicates True news.

The low use of “Impersonal pronouns” indicates True news.

The low use of “2nd person” indicates True news.

The low use of “Certainty” indicates True news.

The high use of “Periods” indicates True news.

List of top linguistic features in the title that indicate the article is fake news.

-----------------------------------------------------------------------------------

The low use of “Hear” indicates Fake news.

The high use of “Common Adverbs” indicates Fake news.

The high use of “Total function words” indicates Fake news.

The low use of “Words > 6 letters” indicates Fake news.

The low use of “Work” indicates Fake news.

The low use of “Power” indicates Fake news.

The high use of “All Punctuation” indicates Fake news.

The low use of “Perceptual processes” indicates Fake news.

The low use of “Relativity” indicates Fake news.

List of top linguistic features in the title that indicate the article is true news.

-----------------------------------------------------------------------------------

The high use of “Colons” indicates True news.

The high use of “Words/sentence” indicates True news.

The low use of “Analytical thinking” indicates True news.

The high use of “Other punctuation” indicates True news.

The high use of “Question marks” indicates True news.

The high use of “Total pronouns” indicates True news.

The high use of “Personal pronouns” indicates True news.

The high use of “Exclamation marks” indicates True news.

The high use of “See” indicates True news.

The high use of “Present focus” indicates True news.

The high use of “Informal language” indicates True news.

The low use of “Past focus” indicates True news.

The high use of “Impersonal pronouns” indicates True news.

The high use of “2nd person” indicates True news.

The high use of “Certainty” indicates True news.

The low use of “Periods” indicates True news.

CHAPTER 5

CONCLUSION

5.1 Conclusion

Today with such easy access to the news, it is important that people can distinguish between true and fake news. Because news is so easy to create and access, fake news is hard to avoid. This online fake news propagates “significantly farther, faster, deeper, and more broadly than the truth in all categories of information (Vosoughi, Roy, & Aral, 2018).” This project served as a proof of concept that machine learning can identify fake news and help users understand the supporting details of that conclusion. This result was obtained through a multistep process. The first step was creating a data pipeline that incorporated URL scraping, data cleaning, and pre-processing. The next step was in training and testing machine learning algorithms to find the most accurate ones. The final step was in giving the user a way to supply news article URLs and then view a results dashboard.

The final product incorporates four machine learning algorithms that use different data extracted from the news articles to rate the reliability of the news. The accuracy of these four algorithms ranges between 93% and 96%.

5.2 Future work

This tool is not the end of what can or should be done to help society identify fake news; this was an example of how we can use ML to give people information to make better decisions.

Below are a few ideas on future enhancements and areas of study.

An area of tool improvement would be to enhance the usability and accessibility of this process. In its current configuration, the project is a collection of Python programs on a local laptop; the logic could be moved to a web page with an intake form for URLs and an automated output dashboard.

An area of investigation to improve accuracy and give the user more supporting information about true and fake news would be to analyze more data about the website and the content on the website itself. The website and web domain contain information that may be beneficial in identifying fake news. This information could include details about other articles and authors, the number and type of advertisements, and an analysis of the types of media like audio, pictures, and video. Each of these aspects of the website may contain information about how truthful or fake the information on the site is likely to be.

A second area of investigation would be in domains outside of machine learning which could provide novel machine learning approaches to differentiate true vs. fake news. Three examples of other areas of study that might provide insight into fake news identification would be psychology, political science, and economics.

A third area of academic research would be to investigate other methods to pre-process the data. For example, this project used stemming to reduce variations of the same word, but Lemmmatization might have been better. One benefit of lemmatization would be that the output will always resolve to actual words. Therefore, it may lend itself to helping the user understand why an article is categorized as either true or fake news.

A fourth area of academic research would be to expand the number of categories to reporting mistakes and satire as separate levels in addition to the real and fake news levels. Both reporting mistakes and satire are neither real nor fake news examples.

A fifth area of academic research would be identifying ways to create a feedback loop to turn this supervised learning process into a semi-supervised or even un-supervised learning process and allow the algorithms to learn as they are fed more information.

# References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives, 31*, pp. 211-236. doi:10.1257/jep.31.2.211

Beri, A. (2020). Stemming vs Lemmatization. *Toward Data Science*. Retrieved from https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221

Bisaillon, C. (2020). Retrieved from Kaggle: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Brownlee, J. (2019). A Gentle Introduction to the Bag-of-Words Model. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Dattatreya, S. (2019). Understanding Text Complexity withReadability Formulas. *Analytics Vidya*. Retrieved from https://medium.com/analytics-vidhya/visualising-text-complexity-with-readability-formulas-c86474efc730

Helstowski, C. (2021). *Newspaper3k – A Python Library For Fast Web Scraping.* Retrieved from Finxter: https://blog.finxter.com/newspaper3k-a-python-library-for-fast-web-scraping/

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media. (Elsevier, Ed.) *Computers in Human Behavior, 80*, pp. 295-302. doi:10.1016/j.chb.2017.11.034

Kaggle. (2017). Retrieved from https://www.kaggle.com/jruvika/fake-news-detection

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings. *Journal of Management Information Systems , 36*(3), pp. 931-968. doi:10.1080/07421222.2019.1628921

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*. doi:10.1186/s41235-020-00252-3

Milligan, A. (2021). Retrieved from https://blog.frac.tl/online-media-bias-and-accuracy

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015.* University of Texas at Austin, Austin, TX.

Rachiele, G. (2018). Tokenization and Parts of Speech(POS) Tagging in Python’s NLTK library. *Medium.com*. Retrieved from https://medium.com/@gianpaul.r/tokenization-and-parts-of-speech-pos-tagging-in-pythons-nltk-library-2d30f70af13b

Sample, C., Jensen, M. J., Scott, K., McAlaney, J., Fitchpatrick, S., Brockinton, A., . . . Ormrod, A. (2020). Interdisciplinary Lessons Learned While Researching Fake News. *Frontiers in Psychology, 11*. doi:10.3389/fpsyg.2020.537612

Sen, O. (2020). *Emotion Detection from an Input Text Using Python.* Retrieved from My DataScience Notebook: https://oindrilasen.com/2020/08/emotion-detection-from-an-input-text-using-python/

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Lie, H. (2018). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286, 8*(3).

Shu, K., Sliva, A., Wang, S., tang, J., & Liu, H. (2018). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36.

Shu, K., Wang, S., & Liu, H. (2017). Exploiting Tri-Relationship for Fake News Detection. *arXiv preprint arXiv:1712.07709*.

Tripathi, M. (2018). How to process textual data using TF-IDF in Python. *Free Code Camp*. Retrieved from https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/

Viviani, M., & Pasi, G. (2017). Credibility in social media: opinions, news, and health information—a survey. *WIREs Data Mining and Knowledge Discovery, 7*(5). doi:10.1002/widm.1209

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 356*(6380), pp. 1146-1151. doi:10.1126/science.aap9559

Wales, J. (2019). Retrieved from https://towardsdatascience.com/full-pipeline-project-python-ai-for-detecting-fake-news-with-nlp-bbb1eec4936d

Wales, J. (2019). Full Pipeline Project: Python AI for detecting. *Towards Data Science*.

Wei, S., & LiZhen, L. (2020). Representation learning in discourse parsing: A survey. *Science China. Technological Sciences, 63*(10), pp. 1921-1946. doi:10.1007/s11431-020-1685-2

Wright, D., & Augenstein, I. (2020). Claim Check-Worthiness Detection as Positive Unlabelled Learning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 476-488.

Zafra, M. F. (2019). Web Scraping news articles in Python. *Towards Data Science*. Retrieved from https://towardsdatascience.com/web-scraping-news-articles-in-python-9dd605799558

Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass, 15*(10). doi:10.1111/lnc3.12438

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys, 53*(5), pp. 1 - 40. doi:10.1145/3395046

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys, 53*(5), pp. 1-40. doi:10.1145/3395046

Zhou, X., Jain, A., Phoha, V. V., & Zafarani, R. (2020). Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice, 1*(2), pp. 1-25. doi:10.1145/3377478

APPENDIX A

|  |  |  |
| --- | --- | --- |
| **LIWC Features** | | |
| Word count | Negative emotion | Time orientations |
| Analytical thinking | Anxiety | Past focus |
| Clout | Anger | Present focus |
| Authentic | Sadness | Future focus |
| Emotional tone | Social processes | Relativity |
| Words/sentence | Family | Motion |
| Words > 6 letters | Friends | Space |
| Dictionary words | Female references | Time |
| Total function words | Male references | Work |
| Total pronouns | Cognitive processes | Leisure |
| Personal pronouns | Insight | Home |
| 1st pers singular | Causation | Money |
| 1st pers plural | Discrepancy | Religion |
| 2nd person | Tentative | Death |
| 3rd pers singular | Certainty | Informal language |
| 3rd pers plural | Differentiation | Swear words |
| Impersonal pronouns | Perceptual processes | Netspeak |
| Articles | See | Assent |
| Prepositions prep | Hear | Nonfluencies |
| Auxiliary verbs | Feel | Fillers |
| Common Adverbs | Biological processes | All Punctuation |
| Conjunctions | Body | Periods |
| Negations | Health | Commas |
| Common verbs | Sexual | Colons |
| Common adjectives | Ingestion | Semicolons |
| Comparisons | Drives | Question marks |
| Interrogatives | Affiliation | Exclamation marks |
| Numbers | Achievement | Dashes |
| Quantifiers | Power | Quotation marks |
| Affective processes | Reward | Apostrophes |
| Positive emotion | Risk | Preentheses pairs |
|  |  | Other punctuation |