Final Project:

Identifying Misinformation in the Media

James Hardaway

College of Education, NC State University

ECI 587: Machine Learning in Education

Dr. Jiang

December 5, 2021

Identifying Misinformation in the Media

James Hardaway

December 5, 2021

Abstract

Whether a student or a working professional, safely navigating digital media is a requirement to understand and engage with the world around us. This paper explores the power of machine learning to assist in weeding out information designed to deceive or mislead. Using traditional data science methods, this project proposes a support vector machine (SVM) learning model that classifies text as either true or false with an accuracy approaching 70%. Through exploring multiple machine learning models, I used contextual filtering methods to train an algorithm against a real-world media dataset. While not full proof, this experiment shows the potential to use automated means to identify misinformation. This effort adds to the growing body of research on how to better trust what we experience online.

1 Introduction

Discriminating between truth and fiction is becoming increasingly difficult in the age of ubiquitous digital information. From national politics to the current COVID-19 pandemic, understanding who or what to believe impacts our lives daily. While I can still remember the days of hardcopy periodicals, books, and the library's card catalog, today's students and young professionals have grown up in an almost exclusively digital world. The digital generation is quite comfortable engaging online as much if not more than in the real world, but their ability to recognize the quality of information on their screens leaves much to be desired.

In the 18 months leading up to the 2016 presidential election, Stanford University commissioned a study to evaluate U.S. students' ability to distinguish between factual media articles and politically biased or manufactured stories (Wineberg, et al., 2016). The results were not flattering. Stanford followed up that study with a national survey that culminated in 2019 with equally dismal results. Most students struggled, with close to 90% failing four of the six survey tasks (Breakstone, et al., 2019). The study concluded, unsurprisingly, that technology is evolving faster than our educational institutions can adapt.

An inability to distinguish between truth and fiction is only getting more difficult as gigs of data are created by the hour that are never reviewed by an editor, proofreader, or certified publisher. It's easy to see how one can get overwhelmed by the sheer volume of information we wade through in a day online. A key recommendation out of the Stanford study is for there to be a fundamental shift to increase digital literacy instruction at all levels of education. This problem is the driving force behind the key question this experiment is attempting to address. Specifically, can technology automate the process of recognizing misinformation in online media?

Understanding how software applications can do this could be extremely relevant for educators as well as anyone working in the field of digital media production. In addition to the technical implications, this research has the potential to shed light on how news media is

categorized, titled, and promoted to either highlight or hide its true character.

This paper is organized according to the broad phases of data analytics: prepare, wrangle, analyze, model, and communicate. Section 2 will introduce other studies and experiments related to my research. These works were essential in understanding how to organize an experiment to adequately answer the research question. Section 3 describes the data and how it was collected, pre-processed, and formatted so it could be properly analyzed. Section 4 outlines the exploratory analysis I conducted to better understand unique aspects of the data, uncovering potential avenues of investigation. In Section 5, the various machine learning models used to classify media articles are discussed to include the process to identify my primary model, support vector machine (SVM). Section 6 focuses on communicating the results of each model as well as efforts to tune the models to achieve increased accuracy. Finally, this section concludes the paper with the experiment's key insights and limitations.

2 Preparation

My military background in the domain of misinformation led me to studies and experiments on how software applications are being applied to the broader data analytics problem of text mining. The analytic challenge that text presents is primarily due to its unstructured nature. Websites, emails, chat conversations, and social media are just a few of the data sources that produce millions of lines of freeform characters.

Unlike tables of data containing values easily parsed and analyzed mathematically, written communications contain many more potential variables based on who's writing, the topic, the medium, and even the language. The goal of a text mining is to break down the long strings of characters into smaller pieces that can be analyzed to gain insights into patterns of communication that can be useful for a variety of applications (Witten et al., 2017, p. 515). Sales figures may indicate how many widgets you sell during the holidays, but they can't tell you why a customer may or may not recommend your product to their neighbor. Text analysis also gives an additional avenue into understanding trends. From breaking news to stock prices to celebrity fashion, people talk about all these on the open internet constantly and understanding these trends give businesses, the government, and consumers a potential advantage.

In developing my experiment plan, I had to determine which text mining methods best support identifying misinformation. The two categories most closely aligned to my research question are information extraction and document classification. Extraction seeks to identify key pieces of information, such as locations, dates, and addresses that can by analyzed as structured data. Alternatively, classification seeks to categorize a document or piece of unstructured text based on word choice and can be used to analyze sentiment or discussion topics. Sentiment analysis attempts to understand the attitude of a given text towards a specific topic. This method works well to determine bias in phrasing. Topic detection attempts to identify subjects or themes in a body of text. Since my challenge is to identify truth versus fiction, I decided to explore both sentiment and topic analysis in the experiment.

Studies on how to identify and mitigate the threat of inaccurate media gained steam in the wake of the 2016 U.S. Presidential Election as the term "fake news" began trending. In December of that year, an entrepreneur/ professor (Dean Pomerleau) and an artificial intelligence researchers (Delip Rao) organized a competition to develop software tools to aid factcheckers in identifying online hoaxes and misinformation (fakenewschallenge.org). The event eventually brought together more than 100 volunteers and 71 teams globally vying for

a nominal cash prize. Their contest highlighted the difficulty with applying broad labels of true and false to nuanced news media, so they instead settled on a "stance detection" methodology, that aimed to identify how varioius media sources reported on specific topics or news stories. These "stances" assisted fact checkers (and the public) in understanding which media sources were more reliable, a key first step in mitigating the spread of misinformation.

FNC-1, as that competition was titled, opened the flood gates for similar experiments and studies across the data science landscape. Two voluminous sources in my research were Kaggle (kaggle.com) and Github (github.com), repositories of data and analysis for a wide array of data science projects. Through Github, I identified another key contributor to this effort, Kai Shu, an assistant professor at the Illinois Institute of Technology. His site (FakeNewsNet) hosts data and analysis of the media misinformation problem set that can be updated with real-time queries. This source is often cited as a start point for fake news experiments such as one completed by data scientist Kumud Chauhan in 2019. Her work focused heavily on exploratory feature analysis to include media source, key word/phrase usage, and length of title and the articles themselves. She ran multiple machine learning models against the data and landed on a Random Forest Classifier as the most accurate at 80%. Her results spanned from 50% - 80% effective depending on the model used and highlighted the variability that can be introduced into ML models based on how the text is decomposed into core features. Features are those elements of text used as the building blocks for pattern analysis.

A final key component to shaping my methodology was determining the type of machine learning to apply to my problem set. Did I want to use "ground truth" to guide my model (supervised) or did I prefer the model to determine the inherent nature of the data (unsupervised)? In this case, I planned to use data that was already categorized as either real or fake, so I began exploring which supervised models were best suited to text classification. Three models rose to the top in my literature review: Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM). NB classifiers often serve as a default start point due to their simplicity and speed. They give a great initial estimate, but struggle when faced with large or complex datasets. LR models work well in binary classification problems where there's a relationship between variables or features. Again, this model struggles with complexity between features, but is probably more suited to our problem than NB. Lastly, SVM models can be used for classifying multiple categories while attempting to identify the best solution that fits numerous variables or features. SVM works best with more complex feature space but can sometimes take a while to process all that data. My plan is to experiment with all three and then progress with fine tuning once the most accurate model is identified.

3 Data Wrangling

As the field of fake news detection grows, finding publicly available data to analyze is not difficult. With numerous sets to choose from, I settled on one that I felt was large enough to ensure I could generate a complex feature space allowing me maximum area to test various theories and learn. To that end, I settled on data initially curated to support a fake news detection study out of Canada (Traore, et al., 2017).

Their data originated in news articles collected from Reuters.com and Politifact.com, arranged in .csv files labeled as either real or fake. The articles are political in nature and occurred in the same general time span between 2015 and 2018. The initial data was separated into two files: a fake news file containing ~23,500 instances and a real news file

containing ~21,400 instances. Both files were structured similarly with columns (variables) for article title, text, subject, and date of publication. The text column contained the actual article and would become the focus for the bulk of my modeling and analysis. The subject column was interesting as it indicated a broad categorization of the article according to where it was originally published. Those categories were eventually weeded out as a key indicator, as there seemed to be little recognizable definition as to the meaning of the categories.

As the data was downloaded in two separate files, none of the news articles were labeled according to their category, so my first step in the wrangling process was to create a 'class' column in each file labeling instances as either real or fake (see Appendix, Figures 3 & 4). This would prove useful when I would need to combine data files prior to employing the ML algorithms during the modeling phase.

Now that I had two adequately large and properly labeled sets of news articles, I did some basic tidying of the data to ensure as little variance between the files as possible. The first edit I made was based on time frame. The fake dataset ranged from March 2015 – February 2019 while the real dataset was limited to 2016 – 2017. I reduced the fake data to match the real data's time frame. I then looked at duplicate, missing and corrupt instances. I removed hundreds of rows with either empty cells or data that did not fit into the basic schema of the five columns. Often when pulling data from secondary sites, some corruption of the files occurs. Again, this dataset was large enough that removing these instances should do little to impact the final outcome.

Lastly, I needed to do a final cut to ensure my analytic software could handle the amount of data for the trials. My primary tools to ingest the data and apply ML models are LightSIDE and WEKA. LightSIDE is a text mining tool created and maintained by Carnegie Mellon University. The Waikato Environment for Knowledge Analysis (WEKA) is an open-source software suite hosting numerous machine learning models that can be used in tandem with LightSIDE. As featurization in text mining can create incredibly large data tables, I had to ensure I stayed within the capacity of these two programs. Some experience has shown that they tend to slow down or crash once you get above 10,000 instances per datafile. To that end, I targeted a final data set with that number of news articles.

4 Exploratory Data Analysis

My initial data exploration was conducted on a set of 10,000 articles that were relatively balanced between real (~4800) and fake (~5200) stories. Using WEKA to run a default Naïve Bayes model returned an accuracy rating of 98.5% and a Kappa coefficient (data consistency) of 0.9703.

This initial run was much more accurate than expected, returning only ~ 150 erroneous classifications. Relooking the data, I recognized there was more cleaning to do due to stray special characters that were missed in the initial data preparation. Additionally, I was concerned that the unbalanced nature of the data (more fake stories than real) may influence the model to place a higher weight on that class. To mitigate this outcome, I modified the data set to be 100% balanced. The refined and cleaned dataset was comprise of 9,000 instances, with 4500 being classified as real news articles and 4500 classified as fake.

I segregated the data into developmental (6000 instances) and test (3000 instances) datasets, with the intent to focus my exploration on the developmental set. The objective during this phase is to identify unique characteristics of the data that could inform model selection or tuning. The news article text column was converted to unigram features using LightSIDE. For this baseline

approach, I did not apply any special filters for the unigrams. I ran another default Naïve Bayes model using this refined dataset in WEKA:

Table 1 - Baseline Model, Naïve Bayes

Dataset	Features	Accuracy	Kappa	Errors
DEV	16,508	97.18%	0.9437	169
TEST	11,438	98.07%	0.9613	58

The results were slightly less accurate than my original trial but still very high for a baseline model with no tuning. While the additional data cleansing had some effect, applying this baseline model produced an extremely high accuracy rating of over 98% for the test dataset.

Those results drove me to employ cluster analysis to see if I could identify any themes in the data that could be biasing the results. I loaded the training dataset (6000 instances) into WEKA and configured the cluster analysis using a simple K-Means algorithm. Generally, this model is best suited for unlabeled data to assist in finding groups of similar instances. In the case of my data, I filtered the algorithm to ignore the class label in the analysis. I then experimented with varying numbers of clusters to see if characteristics other than 'real' or 'fake' are influencing the data:

Table 2 - Development Data Cluster Analysis

Cluster	Instances by Cluster			
0	4571	4475	4470	4465
1	1429	1429	1417	1402
2		96	96	96
3			17	17
4				20

The results clearly show a strong affinity for the instances in cluster #0 and cluster #1. As I add an additional cluster each trial, cluster #0 changes minimally, while cluster#1 produces the bulk of the instances for the new clusters. Looking at the instances for each cluster also produces some interesting themes. Cluster #0 contains numerous articles focused on politics or politicians in the news. Cluster #1 articles, however, are more clearly focused on international issues and security. As the number of clusters grow, the new themes are more specifically focused on security or political issues outside the United States.

The clustering algorithm clearly indicates two primary groupings, though they fail to match up neatly to our known labels of 'real' and 'fake.' I selected four random instances from each of the first two clusters and saw that the subject variable for each instance was similar to the other random instances in that cluster. For Cluster #0, all four instances had either 'politics' or 'politicsNews' as the category. This cluster was comprised of nearly all the fake instances and about half of the real instances. For Cluster #1, all the randomly selected instances were categorized as 'worldnews' in the subject column. This cluster was comprised of almost exclusively real news stories.

This exercise demonstrates a relationship between the real data and international stories while the political theme spans a mix of stories with primacy going to the fake instances. I did one final cluster run ignoring both 'class' and 'subject' variables resulting in over 5900 instances in the first cluster. From this, I'm inclined to conclude the 'subject' variable may be overly influential for the real articles.

My next round of exploratory analysis focused on understanding key words and sentiment of the instances to see if there may be a relationship to article veracity. For these exercises, I employed the RStudio integrated development environment (IDE) which employs the coding language 'R' to assist in understanding and visualizing statistical relationships. After importing the full dataset (9000 instances), I took a random sampling of 10% of both fake and real news instances resulting in a balanced data frame of 900 articles.

To measure the sentiment of these articles, I applied the {vader} package in R which gives a sentiment "score" to each article based on combining values assigned to each word in that article. Those values are derived from a set of sentiment lexicons (or dictionaries) resident in the {tidytext} package in R. The {vader} package computes sentiment scores for each instance by averaging values according to the different dictionaries. In this case I was able to apply these tools to derive sentiment scores for individual as well as groups of instances. The sentiment analysis for the fake and real data frames returned the following summary:

Table 3 - Instance Sentiment by Data Frame

	Positive	Negative	Neutral	N/A
Fake	217	229	4	0
Real	96	128	4	222

For the fake dataset, there is little difference in sentimentality across the articles. Roughly half are seen as positive and half as negative. For the real articles, there is a little more discrepancy, but there are also quite a few instances that were not categorized (see Appendix, Figures 5 & 6). Upon further review, those instances contained strange characters that were indecipherable by the program. My only surprise in this data is that I would have expected the fake news articles to be much more negative in tone.

Finally, I tokenized the 900 articles so RStudio could better treat words as individual features to analyze. These features were then filtered to better understand key words and themes. The initial tokenization resulted in a data frame containing ~382K features. This list includes common "stop words" like "the", "to", "and", "in" that don't carry much meaning by themselves. Upon removing those terms, we have a data frame containing ~180K features, much better for our task. The top tokens in each data frame could give us some insight into discrete themes or commonalities that are exploitable during model development. A query of top tokens in each list returned the following:

Table 4 - Top Fake Tokens

	Word	n
1	trump	1865
2	people	546
3	president	546
4	donald	409
5	obama	349
6	white	311
7	clinton	305
8	time	300
9	news	298
10	image	245

Table 5 - Top Real Tokens

	Word	n
1	trump	1248
2	u.s	875
3	reuters	624
4	president	582
5	government	434
6	republican	406
7	people	395
8	house	385
9	united	371
10	election	322

Key to note are the terms with high counts (n) common to both lists: 'trump,' 'people,' and 'president.' This may indicate some challenges in discriminating between instances that contain similar word choice or phrasing. Also, while it appears in both lists, 'trump' is more than three times as prevalent as the next highest term for fake news. To complete this phase of analysis, I used R to construct some word clouds capturing the top 50 tokens in each data frame:

Figure 1 - Fake News Word Cloud

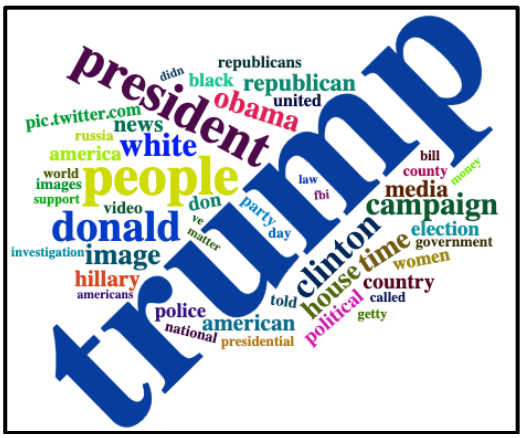
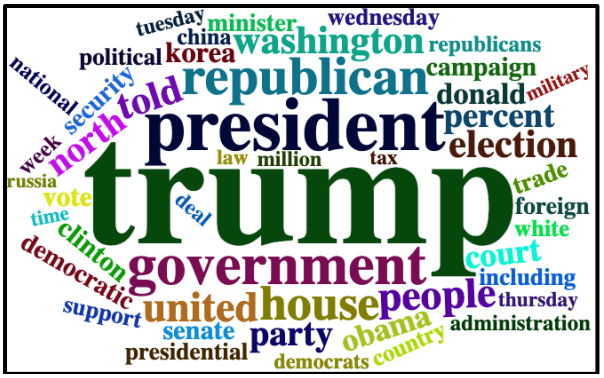


Figure 2 - Real News Word Cloud



Each visual highlights the usage of the top 50 terms by word size. This view emphasizes how much the word 'trump' dominates the fake news articles. The real news articles are also heavily laden with that term, which is to be expected during an election year. That said, there are a few more terms, like 'government' and 'united,' that are more relevant to the real news instances.

To summarize, key insights from the exploratory analysis include a lack of sentimentality, key word (unigram) sharing, and potential metadata and subject bias. Prior to modeling, some additional data cleaning was executed to address some of these issues. Specifically, I removed the source data attached to the 'real' instances. This removed "Reuters" as a key indicator for real news. Additionally, I removed the story location that was also associated with these same instances.

5 Modeling

The process to select a ML model began with comparing how multiple models performed using default parameters against a wide feature space. A key insight carried over from exploratory analysis is the overly influential impact of specific key words, punctuation, and stop words. To guard against this skewing the results, I'm using bigrams (2-word groupings) as my basic feature structure. This does two things for my analysis. First, it decreases the influence of single words (unigrams). Second, it enables context to play a larger role as more can be inferred from word pairings: "white house" versus "white" and "house." Additionally, I'm removing punctuation and stop words from consideration during feature selection. This feature space was then applied to multiple default ML models:

Table 5 - Default Model Comparison

Model	Accuracy	Kappa	# Errors
Naïve Bayes	95.47%	0.9093	272
Simple Logistic Regression	98.45%	0.969	93
Support Vector Machine	98.33%	0.9667	100
Random Forest	95.67%	0.9133	260

To review, based on the strengths/weakness of various classification models, we've chose the SVM for this experiment. This chart reinforces its baseline results. While LR was slightly more accurate, the time to run the model (dataset complexity) and a lack of dependency between features mitigate those minimal gains.

With the SVM selected, the next step was to shape the feature space to best support discrimination between the two classifications in our experiment. While basic bigrams improve performance over other broad feature types, I wanted to dive a little deeper into context to ensure my model has the most optimum chance for applicability to a wider range to text datasets. To do this, I'm adding 'patterns of speech' as an additional filter to the bigram feature space. This gives me two feature tables to work with as I tune the model. The table below compares these two tables to the Unigrams features used in baseline testing:

Table 6 - Feature Space Comparison

Configuration	# Features
Unigrams	16,372
Bigrams	61,927
Bigrams/POS	1142

With model and feature tables identified, I ran a baseline test using the developmental and test datasets. This final baseline will be used to conduct error analysis and fine tuning. The SVM test trial with Bigram features resulted in the following:

Accuracy – 98.1%
Kappa – 0.9620
Errors – 57

Table 7 - Baseline Confusion Matrix

		Predicted	
		Real	Fake
Actual	Real	1465	35
	Fake	22	1478

My error analysis began by investigating samples of the 35 instances that were predicted as 'real' in error. A key theme is inflammatory language in the article that references lying but may or may not actually be untrue. This type of language seems to confuse the algorithm. Another issue is that the vast majority of an article has numerous factual statements, but begins or ends with an inflammatory, biased remark or contains a single opinion that may not reflect factual events. Again, the algorithm seems to struggle sifting through the bigrams to discriminate multiple claims in an article. Finally, I've spotted numerous punctuation errors in the data. While the feature table is designed to filter out specific punctuation, contractions that are missing their apostrophes are treated as two words. For example, "can't" is in the data as "can t" due to a missing apostrophe. That is translated by the feature table as "can_t," a two-word bigram. Those instances are few but may still impact testing results. Unfortunately, not much can be done to the data to fix these items.

Such high initial accuracy is often indicative of a model that is 'overfitted' to the data. This would make sense based on some of the characteristics observed during exploratory analysis. Challenges with the metadata and overuse of key words and phrases could be biasing the results. If true, then this model will only work this well on this data and is less generalizable to new datasets. To test this theory, I obtained another news dataset from a completely different source and formatted it to match (save variables/columns) as the test data. This sample is smaller at 2000 instances but contains the same basic style of online news articles. Applying the baseline model to this "validation" dataset returned the following results:

Accuracy – 61.2%
Kappa – 0.2270
Errors – 773

This seems to confirm that the model is overfitted to my specific dataset. The next step, fine tuning, will aim to address variables within the model that could mitigate the sources of this overfitting.

A popular way to combat overfitting is to reduce features. Our feature table is large at over 62K features. There is a chance that many features get diluted as "noise" and are drowned out by some key words and phrases in the data. LightSIDE has a way to limit the feature table to a specific number of features while training the model. I chose three feature selection limits to see if that changes our results: top 100, top 1000, and top 2000 features. None of these filters changed our results significantly. The validation set only improved to 61.35%.

Another way to reduce feature space is to investigate part-of-speech (POS) tagging to reinforce context. POS tagging labels each word as a part of speech (noun, adjective, etc) and has been found to improve natural language processing task in more advanced ML models. Sticking with bigrams, I added the POS filter and extracted the features from the developmental dataset. This returned a feature table of 1142 features, drastically smaller than the previous sets. Running the SVM model against this feature table produced the following for the test dataset:

Accuracy – 95%
Kappa – 0.8993
Errors – 151

While less accurate than the bigram feature table, 95% is still pretty good. Keep in mind my goal is not to max out the test data but rather to improved results on the validation dataset. Using validation data in this POS filtered model, my results did improve (if only by 8%):

Accuracy – 64.5%
Kappa – 0.2900
Errors – 710

With such a large decrease in features, I expected the model to perform much worse than it did. As a last tuning measure, I experimented with an alternative feature filter than maintains the model and feature table while focusing exclusively on only the most important features. Iteratively, I filtered this model between low and high feature numbers, until I reached what I believe to be the optimum features. This measure significantly reduces the noise presented by much lower weighted features:

Table 8 - POS Bigrams Filtered by Feature

Features	Accuracy	Kappa	Errors
1142 (all)	64.50%	0.2900	710
Top 100	64.65%	0.2930	707
Top 300	67.45%	0.3490	651
Top 350	67.65%	0.3530	647
Top 500	66.65%	0.3330	667
Top 750	63.50%	0.2700	730

My bracketing showed the most improvement at the top 350 features, with a validation set accuracy of 67.65%, a more respectable 16% increase relative to the baseline results.

6 Discussion

For a human to classify text manually, specific domain expertise is required. Expecting a person to maintain this type of knowledge across multiple domains in the digital age is nearly impossible. For the challenge of monitoring online news media for misinformation, the tasks include collecting the articles, ingesting their content, recognizing patterns in the data, and finally classifying the articles into categories. This paper demonstrates that all of these can be accomplished with current software and machine learning algorithms.

This study demonstrated how various textual features could be extracted through LightSIDE and then used to build algorithms that learned to recognize fake news. This was accomplished on a dataset comprised of 6000 news articles. With additional computing power, much larger datasets could yield even greater learning power.

Lessons learned from data wrangling drove this project. To build a model able to recognize misinformation across a broad spectrum of topics requires data from multiple sources. Independent journalists, private citizens, and news organizations each have varying styles for publishing online material. While identifying sourcing may indicate misinformation potential, this project was oriented on article text alone. My dataset, unfortunately, was contaminated with metadata that indicated sourcing and article location. As a result, the trained algorithms were less generalizable against data outside these primary sources. The simplicity of the chosen data also proved problematic during the exploratory analysis phase. Data with additional variables may have provided many more avenues to exploit for potential insights on recognizing fact versus fiction.

Fake news detection contains many areas that can be explored that I didn't address. Pictures, video, and audio can be doctored or artificially produced depicting what could be interpreted as factual content. These areas are ripe for deep learning algorithms that can process large multi-media datasets. Another vector for misinformation is the producers

themselves. Whether attributable or anonymous (including bots, burner, spam), the accounts or sites that host inaccurate content continue to multiply. Identifying and flagging this content would take an army of humans or a couple of smart algorithms.

This class project was designed to explore the potential for automated machine learning tools to recognize misinformation in online news media. These results and those of other similar efforts show that current technology is absolutely up to the task. While the threat of misinformation continues to increase, thankfully the number of individuals and organizations jumping into the fight against it is also growing.

7 References

[1] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, *2020*, 1–11. https://doi.org/10.1155/2020/8885861

[2] Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using N-gram analysis and machine learning techniques. *Lecture Notes in Computer Science*, *10618*, 127–138. https://doi.org/10.1007/978-3-319-69155-8_9

[3] Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., & Saavedra, A. (2019). Students' civic online reasoning: A national portrait. Stanford History Education Group & Gibson Consulting. https://purl.stanford.edu/gf151tb4868

[4] Butrymowicz, S., & Salman, J. (2021). *The in-school push to fight misinformation from the Outside World*. The Hechinger Report. Retrieved September 30, 2021, from https://hechingerreport.org/the-in-school-push-to-fight-misinformation-from-the-outside-world/.

[5] Chauhan, K. (2019). *Fake news analysis and classification*. Kaggle. Retrieved October 5, 2021, from https://www.kaggle.com/kumudchauhan/fake-news-analysis-and-classification.

[6] Li, S. (2018). *Multi-class text classification model comparison and selection*. Medium. Retrieved December 1, 2021, from https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568.

[7] Paialunga, P. (2021). *Fake news detection with machine learning, using python*. Medium. Retrieved October 14, 2021, from https://towardsdatascience.com/fake-news-detection-with-machine-learning-using-python-3347d9899ad1

[8] Pomerleau, D., & Rao, D. (2017). *Fake news challenge stage 1 (FNC-I): Stance detection*. Fake News Challenge. Retrieved October 1, 2021, from http://www.fakenewschallenge.org/.

[9] Sengupta, P. (2021). *Fake news detector: Eda & Prediction(99+%)*. Kaggle. Retrieved October 14, 2021, from https://www.kaggle.com/paramarthasengupta/fake-news-detector-eda-prediction-99

[10] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, *8*(3), 171–188. https://doi.org/10.1089/big.2020.0062

[11] Wineburg, Sam and McGrew, Sarah and Breakstone, Joel and Ortega, Teresa. (2016). Evaluating Information: The Cornerstone of Civic Online Reasoning. Stanford Digital Repository. http://purl.stanford.edu/fv751yt5934

[12] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

8 Appendix

Figure 3 - Fake Dataset Example

	A	B	C	D	E
1	**title**	**text**	**subject**	**date**	**class**
2	CORRUPT CLINTON AIDE Rigged T	The big question in all this corruption is why these people aren t in jail! The collusion be	Government	29-Jun-16	Fake
3	FITTING END FOR COMMUNIST D	As FoxNews.com reportedThe Russianmade jeep ferrying Castro s ashes broke down an	politics	4-Dec-16	Fake
4	U.S. DEPARTMENT OF EDUCATIOI	Does anyone in recent history remember the US Department of Education asking educat	politics	15-Feb-16	Fake
5	This Clip Of Obama OBLITERATIN	Over the last few days Republican presidential frontrunner Donald Trump is finally getti	News	1-Mar-16	Fake
6	BOTTOM FEEDERS PAINT TEAR IT	HERE WE GO yet another statue of a military hero is reportedly under fire. This figure a	politics	18-Aug-17	Fake
7	HYSTERICAL! HERE'S WHY TRUMI	Read moreWT	politics	23-Mar-17	Fake
8	WATCH Ben Stein LOSES IT Over	Conservative economist Ben Stein just let everyone know how much he hates Republica	News	4-Apr-16	Fake
9	WATCH Veteran Rip Trump A Ne	It s not just white American men who fight and die for this country. Women Muslims ar	News	2-Jun-16	Fake
10	Texas GOP'er SHREDDED For INS.	Republicans are hilarious when they try to point out Democratic hypocrisy. They re even	News	12-Jan-17	Fake
11	MARINE CORPS GENERAL WARN	While the mainstream media remains transfixed on a phony TrumpRussian collusion sti	politics	23-Dec-17	Fake
12	Trump's MaraLago Trips Force Se	Despite repeatedly claiming that he will be tough on crime Donald Trump s weekend es	News	22-Mar-17	Fake
13	White House Staff Hiding From R	Word has gone out through the White House that anyone caught speaking with reporter	News	18-May-17	Fake
14	Racist Donald Trump Actually Has	Monday August 29 happens to be the 53rd anniversary of the Civil Rights March on Was	News	29-Aug-16	Fake
15	MIKE ROWE SENDS A BRUTAL ME	An electrical contractor wrote to the 54yearold host of Dirty Jobs to say that he finds it	politics	28-Aug-16	Fake
16	BREAKING...VP PENCE and Wife K	Mike Pence tweeted about how he was looking forward to attending a Colts game with	politics	8-Oct-17	Fake
17	CAN'TMISS Bernie And Hillary MC	You can t say that both parties are the same when the Republican nominee avoids payir	News	2-Oct-16	Fake
18	CAUGHT ON CAMERA! ANTITRUN	The antiTrump protesters stoop to a new low and when challenged by the cameraman t	politics	29-May-16	Fake
19	Happy Tears Young Girl Heartbrol	For many of us electing President Obama into office was an amazing thing to be a part	News	30-Mar-16	Fake
20	SPECTRE OF BENGHAZI DOJ Drops	21st Century Wire says The US Department of Justice (DOJ) is dropping all charges agai	USNews	5-Oct-16	Fake

Figure 4 - Real Dataset Example

	A	B	C	D	E
1	**title**	**text**	**subject**	**date**	**class**
2	Obama to consider proportional re	ABOARD AIR FORCE ONE (Reuters) U.S. President Barack Obama will consider a variet	politicsNews	11-Oct-16	Real
3	Trump backers disparate hopes co	WASHINGTON (Reuters) The pomp and circumstance were like any big Washington cel	politicsNews	20-Jan-17	Real
4	Trump says he would push univers	CHESTER TOWNSHIP Pa. (Reuters) Republican presidential nominee Donald Trump saic	politicsNews	22-Sep-16	Real
5	Exclusive Skeptical Trump says wo	NEW YORK (Reuters) Republican presidential contender Donald Trump said on Tuesday	politicsNews	17-May-16	Real
6	China warned North Korea of sanc	WASHINGTON (Reuters) China has told the United States that it warned Pyongyang it v	politicsNews	27-Apr-17	Real
7	Shots fired at DMZ as North Korea	SEOUL (Reuters) South Korean guards fired warning shots across the heavily militarize	worldnews	21-Dec-17	Real
8	The growth of Syrias humanitaria	LONDON (Reuters) While 2017 saw the tide of the military conflict in Syria seemingly t	worldnews	14-Dec-17	Real
9	Buffett says success of Trump tax	(Reuters) Warren Buffett is closely monitoring whether U.S. President Donald Trump ca	politicsNews	3-Oct-17	Real
10	Election stirs debate about Feds h	WASHINGTON (Reuters) Donald Trump says the Federal Reserve has stoked asset bubl	politicsNews	7-Apr-16	Real
11	U.S. Labor Department rescinds O	(Reuters) The U.S. Labor Department on Wednesday said it was rescinding the Obama	politicsNews	7-Jun-17	Real
12	Senate rejects new U.S. retiremer	WASHINGTON (Reuters) The U.S. Senate voted along party lines on Tuesday to repeal a	politicsNews	24-May-16	Real
13	Cruz picks up backing of Family Re	WASHINGTON (Reuters) The leader of an influential Christian conservative lobbying gr	politicsNews	27-Jan-16	Real
14	Trump says Puerto Rico has throw	WASHINGTON (Reuters) U.S. President Donald Trump praised federal officials and the	worldnews	3-Oct-17	Real
15	Socalled judge derided by Trump l	(Corrects first paragraph to make it jurist" instead of justice" as Robart is not a justice)	politicsNews	5-Feb-17	Real
16	NZ kingmaker calls for inquiry ove	(Reuters) A populist politician whose party could emerge as a kingmaker at this month	worldnews	14-Sep-17	Real
17	U.S. Nordic nations call on Russia	WASHINGTON (Reuters) Five Nordic countries and the United States on Friday called o	politicsNews	13-May-16	Real
18	China rules out military force as o	BEIJING (Reuters) The use of military force to resolve issues on the Korean peninsula is	worldnews	5-Sep-17	Real
19	New Senate healthcare plan keep	WASHINGTON (Reuters) The revised U.S. Senate healthcare plan will keep in place two	politicsNews	13-Jul-17	Real
20	U.S. adjusts military support to pa	WASHINGTON (Reuters) U.S. President Donald Trump said that he had informed Turke	worldnews	24-Nov-17	Real

Figure 5 - {vader} Scoring, Fake Articles

	text	word_scores	compound	pos	neu	neg	but_count
1	Watch the video and you be the judge. Was Piers Mor…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	−0.916	0.054	0.863	0.083	2
2	If there s one thing that has become abundantly clear…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 1.6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…	0.959	0.117	0.805	0.078	0
3	IMAGE: Globalist scribe for the American branch of int…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	0.965	0.110	0.845	0.045	0
4	According to the Journal of Blacks in Higher Educatio…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.1, 0, 0,…	−0.991	0.018	0.860	0.122	0
5	House Republicans leaders on Monday embraced a le…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2.2, 0, 0, 0, 0, 0, 0,…	0.969	0.095	0.842	0.063	0
6	The plans for the Republican convention have suffere…	{0, 0, 0, 0, 0, 0, 0, −1.1, 0, −0.8, 0, 0, 0, 0, 0, 0, 0, 0, …	0.978	0.150	0.761	0.090	4
7	Tucker Carlson debated a New Jersey Democratic ope…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, −0.45, 0, 0, 0, 0, 0, 0, 0,…	−0.980	0.038	0.746	0.217	1
8	The Republican Party s establishment candidates are …	{0, 0, 0.85, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …	0.943	0.121	0.782	0.097	2
9	21st Century Wire asks The sensational reporting by …	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, −0.6035, 0, 0, 0, 0,…	−0.829	0.074	0.760	0.166	1
10	Wow! Former Bears coach Mike Ditka is coming under…	{1.4, 0, 0, 0, 0, 0, 0, 0, 0, −0.7, 0, 0, 0, 0, 0, 0, 0, 0.0…	0.995	0.121	0.820	0.059	3
11	With their puppet firmly secured in the White House R…	{0, 0, 0, 0, 0.85, 0, 0, 0, 0, 0, 0, 0, 0.5, 0, 0, 0, 0.5, 0,…	−0.934	0.092	0.793	0.116	1
12	IT S INTERESTING THAT A LIGHTHEARTED MOMENT LI…	{0, 0, 1.2165, 0, 0, 1.2665, 0, 1.1165, 0, 0, 0, 0, 0, 0,…	−0.932	0.079	0.812	0.109	1
13	Tune in to the Alternate Current Radio Network (ACR) …	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	0.930	0.099	0.878	0.023	0
14	Washington Post: Is this the woman that Michelle Oba…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2.3, 0, 0,…	0.937	0.070	0.888	0.042	0
15	It has become customary for a president upon leaving…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, −0.…	0.982	0.134	0.782	0.084	2
16	By Vin ArmaniHillary Clinton continues to blame Russi…	{0, 0, 0, 0, 0, 0, −1.4, 0, 0, 0, 0, −1.4, 0, 0, 0, 0, 0, 0.…	−0.361	0.071	0.818	0.111	0
17	Documentary filmmaker Ken Burns knows a thing or t…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	−0.568	0.052	0.892	0.056	5
18	Despite the fact that many election experts have been…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…	0.966	0.141	0.754	0.105	2

Figure 6 - {vader} Scoring, Real Articles

	text	word_scores	compound	pos	neu	neg	but_count
1	(Reuters) - Maryland's House of Delegates on Mond…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	WASHINGTON (Reuters) - U.S. President-elect Donald …	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
3	WASHINGTON (Reuters) - The United States has sancti…	{0, 0, 0, 0, 1.8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…	0.818	0.132	0.868	0.000	0
4	LONDON (Reuters) - London s police force urged peo…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	0.077	0.016	0.984	0.000	0
5	LONDON (Reuters) - Back from Brussels with a hard-f…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	0.945	0.084	0.843	0.073	7
6	BERLIN (Reuters) - Largely unperturbed by Angela Mer…	{0, 0, 0, 0, 0, 0, 0, 0, 0, −1.15, 0, 0, 0, 0, 0, 0, 0, 0, 0,…	−0.973	0.110	0.758	0.132	8
7	WASHINGTON (Reuters) - Foreigners aiming for temp…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
8	WASHINGTON (Reuters) - U.S. senators have been inf…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	0.103	0.017	0.969	0.014	1
9	The following statements were posted to the verifie…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
10	WELLINGTON (Reuters) - A New Zealand Labour gover…	{0, 0, 0, 0, 0, 0, 0, 0, 0, 0.15, 0, 0, 0, 0, 0, −0.4, 0, …	0.992	0.163	0.763	0.074	4
11	BERLIN (Reuters) - Chancellor Angela Merkel said Ger…	{0, 0, 0, 0, 0, 0, 0, 0, 0.85, 0, 0, −0.75, 0, 0, 0, 0, 0, 0…	0.802	0.140	0.803	0.057	2
12	BEIRUT (Reuters) - Two suicide bombers struck a poli…	{0, 0, 0, 0, −1.75, 0, −0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…	−0.998	0.040	0.743	0.217	3
13	WASHINGTON (Reuters) - The White House said on Th…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
14	WASHINGTON (Reuters) - Supreme Court nominee Me…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
15	WASHINGTON (Reuters) - Several Democratic senators…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
16	ALBANY N.Y. (Reuters) - New York lawmakers on Mon…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
17	NEW YORK (Reuters) - Vermont Senator Bernie Sander…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
18	WASHINGTON (Reuters) - The Obama administration…	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*