Classifying Vaccine Misinformation in Online Social Media Videos using Natural Language Processing and Machine Learning

Sarah Schmidt[1], Brian Thoms[1], Evren Eryilmaz[2], Jason Isaacs[1]

[1]California State University, Channel Islands, [2]California State University, Sacramento

sarah.schmidt344@myci.csuci.edu, brian.thoms@csuci.edu, evren.eryilmaz@csus.edu, jason.isaacs@csuci.edu

Abstract

The spread of information through online social media videos is one of the most popular ways to share and obtain information, while at the same time the spread of misinformation across these same social spaces has become a significant concern affecting human well-being. Being able to detect this misinformation before it spreads is becoming more and more desirable for many social media platforms. This research focuses on exploring the accuracy of detecting misinformation across two social media platforms, YouTube and BitChute. This involves the classification of video data into two types: genuine information or misinformation. More specifically, this research generates additional metadata embedded within online videos related to the COVID-19 vaccination. Using natural language processing (NLP) we extract medical subject headings (MeSH) terms from video transcripts and classify videos using four machine learning techniques including naïve Bayes, random forest, support vector machine, and logistic regression. Implementation of each classifier is presented, and the accuracy of each technique is compared and discussed.

1. Introduction

In early 2020, the World Health Organization (WHO) declared a worldwide 'infodemic' to characterize the overabundance of largely false and misleading information [1]. WHO's declaration reinforces the dangers of misinformation across a wide array of health topics, including COVID-19 vaccine awareness and its efficacy. Moreso, misinformation may result in individuals being less likely to accept and follow public health guidelines critical in battling the pandemic [2] and be less likely to become vaccinated against the virus [3].

Exacerbating these challenges is the propagation of misinformation related to health information at an alarming rate on social media websites [4]. Consequently, the most popular social media companies such as Facebook, Alphabet and Twitter are faced with new challenges in mitigating the spread of misinformation online at the expense of placing limits on users' freedom of expression. As such, social media companies have relied on a combination of systems to manage such misinformation, including using human expert analysis and computer algorithms for flagging and filtering misinformation. Computer-based systems, such as work in [5], rely on natural language processing (NLP) to extract and arrange sentences in ways to aid in the classification of factual claims, while other solutions, such as research in [6] leverage advanced machine learning techniques that validate the credibility of information and aid in identifying what aspects of information a user should focus on.

While much recent research has focused attention on the mitigation of the spread of false information across social networking platforms, less research has been conducted on video-based sources. And, to the best of our knowledge, no research has focused on video content across social media websites where comments have been disabled. Research that has focused on online social media videos has focused primarily on the text-based aspects of the video such as the title, hashtags, and comments associated with the video as evinced in work by [7] as well as work by [8] which took comments into account.

In this research, we adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework and investigate the degree to which misinformation exists across online social media videos. More specifically, we aim to measure additional features found within online videos beyond its metadata, including video transcript sentiment and medical subject headings (MeSH) terms, to formulate a more complete subset of video attributes for which to analyze using NLP and machine learning techniques. Our research contributions are as follows. First, we identify a baseline dataset of social media videos for future research and further baseline comparison. Second, we identify a combination of novel features for identifying misinformation within this dataset. And third, we provide a detailed analysis of the performance of different machine learning algorithms on a subset of this dataset.

2. Background

2.1 Social Media

The general population continues to spend more time online each day consuming information. Some estimates state people in the U.S. spend an average of around 2.5 hours a day on social media [9]. According to a report by Shearer and Gottfried in, "News Use Across Social Media Platforms 2016", an estimated 62% of Americans get news on social media, with around 50% of this population having viewed this news on social media [10]. While interest in news and current events is generally encouraging, access to accurate

URI: https://hdl.handle.net/10125/106847
978-0-9981331-7-1

and reliable information remains challenging, particularly when it comes to critical information surrounding one's personal health. The challenge is only exacerbated by content generated by artificial intelligence (AI) and computer robots [11].

2.2 YouTube

YouTube was launched in 2006 and has since become one of the largest video stores on the planet with 2.2 billion users streaming 700,000 hours of video each minute [12, 13], numbers that continue to grow. Consequently, many organizations rely on YouTube to disseminate important health information, including WHO and the U.S. Centers for Disease Control (CDC). Consequently, as more users rely on streaming content within these social media platforms for information, other users leverage its wide access to disseminate misinformation. Recent research analyzing online videos on YouTube for vaccine misinformation discovered that as high as 65% of videos surveyed discouraged vaccine use for various reasons [14] and 23% of videos surveyed, spread misinformation related to pandemics [15].

In the past, YouTube has provided self-governing rules for nudity or sexual content, harmful or dangerous content, hateful content, violent or graphic content, harassment and cyberbullying [16], but left fact-checking to moderate itself. This changed in 2021 to combat the known issue of vaccine-related misinformation propagating across its network. Consequently, YouTube added new rules for posting videos related specifically to COVID-19 to its community, which states, "YouTube prevents uploading content that spreads medical misinformation that contradicts local health authorities' (LHA) or the World Health Organization's (WHO) medical information about COVID-19" [17]. To a large extent, this moderation significantly reduced the spread of misinformation across its services and resulted in the vast majority of information found on YouTube related to COVID-19 being factual.

2.3 BitChute

BitChute is an online social media hosting service launched in 2017 as one alternative to YouTube. The video-sharing platform joins an ecosystem of alternative, low content moderation platforms, including Gab, Voat, Minds, and 4chan [18]. While BitChute states that posted content must not violate the Prohibited Content or Platform Misuse guidelines [19], due to its low content moderation, extreme views and misinformation are able to propagate through its network unfettered. Some creators who use BitChute have even been banned from YouTube for violating its terms of service [18]. Consequently, BitChute offers a unique opportunity for researchers to study social media misinformation, since videos that spread misinformation, while banned from YouTube may continue to thrive and propagate through other online social media channels, providing for good baseline comparisons. As a result, videos linked to BitChute, remain unmoderated making it difficult for the average person to discern fact from fiction when only presented with an embedded video and not its source.

2.4 Natural Language Processing

Natural Language Processing (NLP) is a broad field that encompasses the use of computing systems and linguistics to achieve a better understanding of human language. NLP systems can greatly speed up the search and retrieval of embedded web content and determine whether a resource is relevant. Additionally, NLP can help to resolve ambiguity in language and can be used to preprocess and simplify data for downstream systems. NLP systems have been extensively researched in the area of misinformation detection across both social media and within the health domain. In Zhao et al. [20], research investigated behavioral patterns of health communities to mitigate the spread of misinformation online. More related to this research and COVID-19 misinformation within social media videos, Serrano et al. [9] used YouTube video comments to validate the content found within the video. Specifically in this research, NLP plays an integral role in analyzing transcripts of online videos and serves as a critical step in data preprocessing for social media videos where comments are disabled.

2.5 Machine Learning

In recent years, machine learning has become integral in leveraging computing power and efficiency for analyzing a wide variety of data across a wide variety of fields, including fake news detection. At a basic level, the field of machine learning is devoted to understanding and building methods that 'learn', i.e., methods that leverage data to improve performance for some specific task [21], such as classifying information as genuine or fake. Machine learning systems do so by constructing models using an initial set of training data to make predictions or decisions without being explicitly programmed to do so.

In this research, we implement four classical machine learning algorithms on a corpus of videos related to the COVID-19 vaccine. The machine learning classifiers implemented in this research were chosen because of their popularity across other misinformation detection research and, therefore, offer good baseline comparisons.

2.5.1 Naïve Bayes. Naïve Bayes is a probabilistic classifier based on applying Bayes' theorem with strong independence between data features. Naïve Bayes classifiers typically perform well across small data sets where data consists of numerous features. Bayes classifiers have been used in recent years in the detection of misinformation on online social networking platforms such as YouTube and Twitter. In [22], researchers implemented a naïve Bayes classifier on a Twitter dataset with 22 distinct features for fake news detection. And, more recently, research conducted in [19], incorporated a naïve Bayes classifier to assess

YouTube video misinformation related to COVID-19 using only video comments.

2.5.2 Random Forest. Random forest is a classification method that operates by constructing a multitude of decision trees during the training process. One reason random forest classifiers are popular is that they aim to reduce overfitting and variance which can improve the accuracy of a model. Random forest classifiers have been implemented in the identification of misinformation as well. In work by [23], researchers analyzed a random forest classifier on tweets related to COVID-19 during the early months of the pandemic. More recently, research in [24], implemented a random forest classifier on a corpus of 120 million COVID-19 tweets using sentiment and topic modeling as primary features.

2.5.3 Support Vector Machine. Support vector machine is a supervised learning model with associated learning algorithms that analyze data for classification and regression analysis. Research in [22], found support vector machine classifiers to be more effective for categorizing misinformation within the Twitter social network when compared to naïve Bayes, however, the classifier was less effective when compared to a random forest classifier [23].

2.5.4 Logistic Regression. Logistic regression is a classification technique that uses a logistic function to model the dependent variable. The dependent variable is dichotomous, i.e. there could only be two possible classes. In statistics, logistic regression can be used to model the probability of a certain outcome and is among the top 5 most widely used baseline models [25], largely chosen for its simplicity. This approach has become popular in machine learning comparisons such as work in [26], where researchers use logistic regression as a baseline for comparing other machine learning algorithms in fake news detection.

3. Research Methodology

3.1 CRISP-DM

The research framework implemented in this study adheres to rigor defined in the Cross Industry Standard Process for Data Mining (CRISP-DM), which is an open industry standard [27]. This standard was first introduced in the early 2000s to make the data mining process efficient and easily repeatable [28]. CRISP-DM remains an effective framework with applications in social networking and healthcare [29]. Illustrated in Figure 1 is an adapted CRISP-DM framework for guiding this research. The first phase involves data collection or gathering of social media, the second phase involves data pre-processing and metadata analysis, the third phase involves feature selection, the fourth phase is the model training and classification phase, and the fifth phase presents the output and discusses these results. The methodology allows us to focus on exploring the following high-level research questions:

1. What additional embedded metadata within online social media videos can NLP techniques help to uncover?
2. How will classical machine learning techniques perform and compare across these new sets of features?

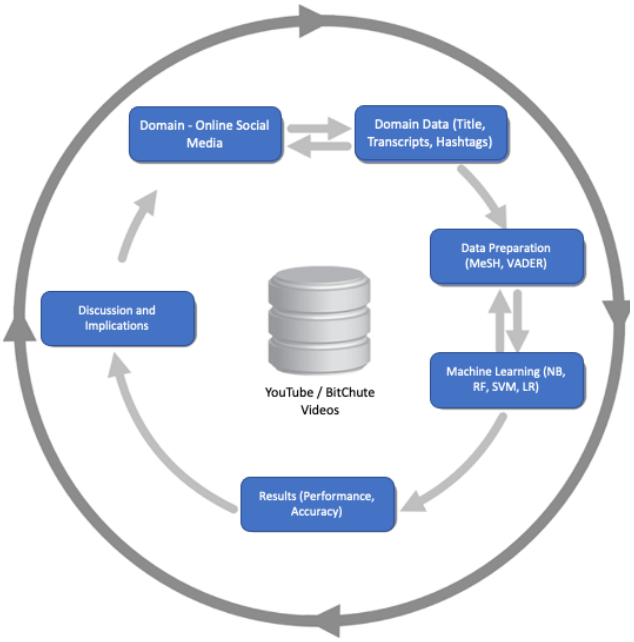

Figure 1 – Adapted CRISP-DM

3.2 Data Collection

The dataset analyzed in this research was collected from YouTube and BitChute social media platforms and consisted of 17 misinformed videos and 15 genuine videos with 14 (82%) of the misinformed videos, coming from BitChute and all 15 genuine videos analyzed in our dataset coming from YouTube. Videos were determined to be genuine based on their source from various government agencies such as the U.S. CDC and Federal Drug Administration (FDA). Videos containing misinformation were verified using a combination of human expert analysis and fact-checking websites such as Snopes.com and FactCheck.org.

3.3 Data Preprocessing

Videos were preprocessed for context using transcripts, as well metadata. This research focuses specifically on video transcripts of content. Video transcripts are an effective way to capture dialogue within a video, with the limitation that other aspects of the video are neglected such as tactic expressions and visual aids. Transcripts processed from YouTube leveraged the 'Show Transcript' feature of the software, while 17 video transcripts consisting of 555

minutes of video processed from BitChute were generated through a service called Happy Scribe, which transcribes the video into a plain text file. Happy Scribe's automatic transcription algorithm is self-reported to be 85% accurate when transcribing a video or audio file [30].

3.4 Feature Selection

Feature selection is a critical step in machine learning and allows researchers to select those features that contribute the most to the desired output. For our purposes, we aim to select features that may best classify a video as genuine or not. Feature selection not only helps to reduce overfitting and improve accuracy but also decreases the training time since there is less data to train. During this stage, we select which parts of the data we want to use from the data we collected. A total of five features were constructed for inclusion in our machine learning classifiers.

3.4.1 Transcript MeSH Terms. Since the information we are concerned with is medically related, the first features selected were terms extracted form the Medical Subject Headings (MeSH) thesaurus, which is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine (NLM). It is used for indexing, cataloging, and searching of biomedical and health-related information. MeSH has been an important way to classify medical publications since its first introduction in 1960 [31] and has been used in text classification machine learning algorithms within the medical domain, such as work in [32], which used machine learning to extract MeSH terms from unstructured medical notes. In the past decade, NLM has supported MeSH on Demand, a web-based platform for automatically generating MeSH terms from text [33]. The software provides a viable option for automatically extracting relevant keywords from online text, including video transcripts, as performed in this research. MeSH keyword count was thought to be a valuable feature because MeSH on Demand responds mainly to medical-related terms which should be used primarily by trustworthy or credible sources. Consequently, the more MeSH keywords a video has, the more scientific it may be perceived.

3.4.2 Video Length. The second feature focuses on video metadata and considers each video length, in seconds. Metadata remains a critical feature of social media, which allows content to be categorized. As such, metadata can be a critical component of machine learning and is commonly used in social media classification studies, such as work in [34], which used a combination of user profiles, metadata and textual content within tweets to create classification models for misinformation categorization. A lot of the misinformed videos were either from video podcasts which turned out to be lengthy, and there were a couple of videos that were short snapshots of things people were saying about the COVID-19 vaccine that was not true. Length was thought of as a weak feature but was included in the feature sets since the misinformed videos seemed to either be excessively long or very short, whereas a lot of the genuine videos selected were an average length in comparison.

3.4.3 Content Sentiment. The third feature selected focused on the sentiment of each video transcript. Due to the personal nature of healthcare, sentiment can play a crucial role in how health information is perceived. Sentiment analysis and persuasion has been garnered much attention across health misinformation research, such as work in [35] which discovered how typically tries to take advantage of viewers' fear and anxiety related to a particular health topic. In this study, video sentiment was extracted using the Natural Language Toolkit (NLTK) Sentiment Intensity Analyzer called Valence Aware Dictionary and Sentiment Reasoner (VADER) [36]. Results for each sentiment analysis were broken down into three categories: negative, neutral, and positive. Each category was decimal value having a maximum of three significant digits. There is a fourth separate category unrelated to the others called compound which can be any decimal value up to 4 significant digits between negative one and one.

3.4.4 Video Hashtags. The fourth feature selected was video metadata on the number of hashtags assigned to each video. This data was gathered manually by checking the hashtags on each of the videos at their original source. The reason this feature was featured here is that it is known that social media posts used hashtags to categorize and increase access and visibility of a post [35, 37]. For any video to reach large audiences, it must be easily accessed and viewed by many people in a short amount of time.

3.4.3 Title Sentiment. The final feature selected was the sentiment of the video titles. Using the same VADER library [36], sentiment for each video title was analyzed. Sentiment of the title is an important feature included in this study because even before the video is viewed, the title is the first thing that is seen by a social media user. Sentiment of titles was successfully used in [38] to achieve accuracy rates north of 85% and thus considered an important feature that has been used in fake news detection. This first look often determines if the user will eventually watch a video or not. It was thought that since misinformed media aim to capture the user by playing on their anxieties and fears [39] that a highly negative or even a highly positive title would be used to pique the user's interest. Our dataset identified genuine video titles to be sensible and simply reflected the date and primary topic of the meeting or seminar, nothing meant to be attention-grabbing, simply informational.

3.5 Feature Results

An overview of the features used in our analysis is provided in Table 1. The average number of MeSH terms for genuine videos was 78 compared to 88 for misinformed videos. The average hashtag count for genuine videos was 1 versus 2 for misinformed videos. The average length for genuine videos was 43 minutes versus 31 minutes for videos

with misinformation. The average VADER score for transcripts of genuine videos was 0.9 versus 0.2 for misinformed videos, while the average VADER score for the title of videos was 0.1 versus -0.1 for misinformed videos.

Table 1. Video Feature Overview

	Genuine	*Misinformation*
Video Count	15	17
Avg. MeSH Terms	78	88
Avg. Hashtag Count	1	2
Avg. Video Length (min)	43	31
Avg. Transcript VADER Score	0.9	0.2
Avg. Title VADER Score	0.1	-0.1

4. Implementation

Python along with the Natural Language Toolkit (NLTK) was used to preprocess data and implement the machine learning algorithms used in the analysis. Our aim was to test out the data in the dataset against several classifiers that NLTK has available for use in their library. A virtual environment for Python was set up to isolate and manage the python version used to run the algorithms. Preprocessing and machine learning model generation was performed on macOS operating system. The implementation of the classifiers involved using the NLTK built-in classifiers. Each classifier was set up similarly, using labeled feature sets for training purposes. The labels used were 'genuine' and 'misinformation'. Videos were labeled genuine if they were from a trustworthy or credible source and were verified to be true, and they were labeled misinformation if they were from questionable sources and were verified to be false. An example of how a labeled feature set would be structured is detailed as follows:

```
(
  {'mesh_keywords': 19,
   'hashtags': 4,
   'sentiment': 0.018,
   'title_sentiment': 0.4404,
   'length': 2580},
  'genuine'

)
```

All data was combined and labeled into a single dictionary, then the data was split into train and test sets with 80% in the test set and 20% in the train set. While a typical ratio is 80% training and 20% testing, we adopted less data to train. While the content generated by transcribing each video was large, our overall dataset was small and machine learning classifiers, such as Naïve Bayes, SVM [40] and logistic regression tend to perform well with smaller training sizes.

Implementing each classifier model required feature sets to be unlabeled. Consequently, the test feature set was stripped of the 'genuine' or 'misinformation' labels and then run through the classifier to test the accuracy and record the results in a text file with each line having the result for one video in the dataset. Since the algorithm was run several times with a change in how many features were defined in the feature set, the complete feature set was stored in a CSV file, but when it was parsed, only the features planned to be used in the classifier were included in the final feature set.

5. Results

Each classifier was measured for its performance across the combinations of each of the five features, resulting in 168 different combinations of features measured for each classifier. Table 2 provides an overview of the machine learning process broken down by performance across all feature tests and classifiers. Performance was measured using accuracy, which was defined as follows:

Accuracy Rate (%) = Number of Correctly Identified Videos / Total Number of Videos

Table 2. Classifier Performance

Accuracy	*Naïve Bayes*	*Random Forest*	*Support Vector Machine*	*Logistic Regression*
>= 90%	0%	1%	0%	0%
>= 80%	15%	5%	5%	8%
>= 70%	48%	16%	37%	35%
>= 60%	82%	30%	38%	42%
>= 50%	98%	65%	60%	64%
>= 40%	100%	97%	100%	100%
>= 30%	100%	100%	100%	100%

5.1 Naïve Bayes

The naïve Bayes classifier resulted in the highest levels of performance across the feature sets with 26 tests resulting in a performance of at least 80%. Detailed in Table 3 are the top three feature combinations that resulted in the best performance for the naïve Bayes Classifier.

Table 3. Naïve Bayes Performance

Feature Count	*Feature Set*	*Accuracy*
4	{Title Sentiment, MeSH Keyword Count, Hashtag Count, Video Length}	84%
4	{Title Sentiment, Transcript Sentiment, Hashtag Count, MeSH Keyword Count}	84%
4	{Title Sentiment, Transcript Sentiment, Hashtag Count, Video Length}	84%

The best performing combination of features for naïve Bayes was title sentiment, MeSH keyword count, number of hashtags and video length.

5.2 Random Forest

Random forest performance resulted in the highest accuracy levels. Detailed in Table 4 are the feature sets that contributed to the best performance for the random forest classifier.

Table 4. Random Forest Performance

Feature Count	*Feature Set*	*Accuracy*
2	{Transcript Sentiment, Hashtag Count}	92%
3	{Title Sentiment, Transcript Sentiment, Hashtag Count}	88%
3	{Title Sentiment, Transcript Sentiment, Hashtag Count}	88%
	{Title Sentiment, Transcript Sentiment, Hashtag Count}	84%

The best performing combination of features for random forest were transcript sentiment and number of hashtags.

5.3 Support Vector Machine (SVM)

SVM performance was lower than both Naïve Bayes and random forest with performance 80% or greater. Detailed in Table 5 are the feature set combinations for support vector machine.

Table 5. Support Vector Machine Performance

Feature Count	*Feature Set*	*Accuracy*
2	{Title Sentiment, Hashtag Count}	80%
3	{Title Sentiment, Transcript Sentiment, Hashtag Count}	80%

The best performing feature set was a combination of two features for support vector machine, title sentiment and number of hashtags.

5.4 Logistic Regression

Logistic regression performance was lower than both naïve Bayes and random forest but higher than support vector machine with performance of 80% or greater. Detailed in Table 6 are the top four feature set combinations for logistic regression.

Table 6. Logistic Regression Performance

Feature Count	*Feature Set*	*Accuracy*
3	{Title Sentiment, Transcript Sentiment, Hashtag Count}	84%
2	{Title Sentiment, Hashtag Count}	80%
3	{Transcript Sentiment, Hashtag Count, MeSH Count}	80%
3	{Transcript Sentiment, Hashtag Count, Video Length}	80%

The best performing combination of features for logistic regression were title sentiment, transcript sentiment and number of hashtags.

6. Discussion

This research explores an NLP approach to machine learning feature identification for social networking videos related to COVID-19 misinformation. Additionally, this work compares the performance of these features across multiple machine learning classifiers. Specifically, we aim to answer two research questions: what additional metadata embedded within online social media videos can NLP techniques help to discover and how will classical machine learning techniques perform and compare across these new sets of features? Based on the results of this study, the use of NLP techniques showed promise in enhancing feature selection and classification and should be considered in future studies.

6.1 Feature Performance

Feature selection plays a critical role in classical machine learning. To answer RQ1 we revisit how well each of our chosen features performed across our classification models. The general trend in each of the classifiers is that using four features in the feature set leads to lower standard deviations in accuracy compared to using fewer features in the feature set. Using four features in the feature set for three of the classifiers leads to higher average accuracy rates for naïve Bayes, SVM, and logistic regression, and higher maximum accuracy rate in one of the classifiers, naïve Bayes.

The features resulting in highest accuracy were title sentiment and number of hashtags, while the number of hashtags had the greatest effect in increasing accuracy rates across each classifier. During the data-gathering phase, it is worthwhile to note that each of the misinformed videos used hashtags whereas the genuine videos used very few hashtags, if at all. This is probably why this feature had such an impact on the classifiers' ability to detect misinformation. Misinformed social media posts are known for using hashtags to spread as quickly as possible to a larger audience [36]. Applying number of hashtags as a feature in a feature set for larger data sets will likely be a good feature to include to accurately detect misinformation, however, that trend could change in the future if more credible scientists try to leverage the use of hashtags to spread genuine information to people as fast as possible.

6.2 Naïve Bayes

Also in this research, we measure the performance of each classifier was also measured. To answer RQ2, we illustrate the performance of each machine learning classifier with respect to the number of features used.

Figure 2 illustrates a box and whisker plot of the accuracy rates compared to the number of features used in the naïve Bayes classifier. The highest accuracy rate occurred when four features were included in the feature set. Generally, for a naïve Bayes classifier, the more features used in a classifier, the better the results will be. The highest accuracy rate was 84% and involved the following four features: title sentiment, MeSH keyword count, hashtags count, and video length. These results provide insight into how naïve Bayes may be improved in the future to contain more features, i.e., more metadata surrounding each video. Overall, title sentiment, transcript sentiment, and hashtags count resulted in higher accuracy rates.

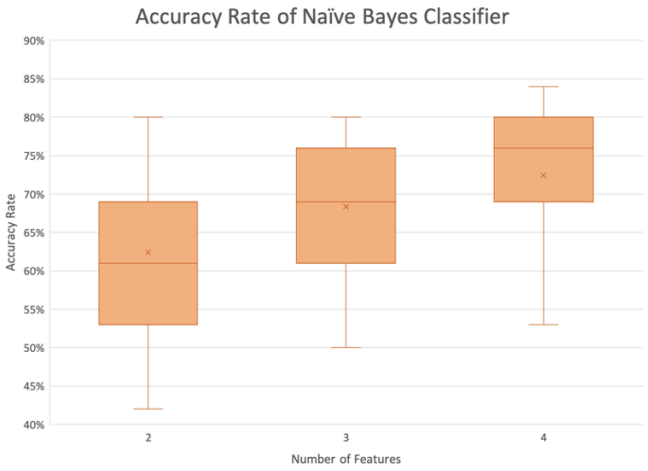

Figure 2. Box and Whisker Plot of Accuracy of Naive Bayes Classifier

6.3 Random Forest

Figure 3 illustrates a box and whisker plot of the accuracy rates compared to the number of features used in the random forest classifier. The highest accuracy rate used two features and resulted in 92% accuracy, which was also the highest across all classifiers. Features used to achieve these rates were transcript sentiment and hashtags count. There was not a significant correlation between the number of features and accuracy rate, but there was a slight increase in the highest accuracy rate when fewer features were used. This suggests that, unlike naïve Bayes, adding features will not necessarily improve results, although the variance is tighter with more features, different features may return higher accuracy rates. One additional finding was that the random forest classifier shared similar patterns with the naïve Bayes classifier in that when more features were used, lower standard deviations occurred, which indicates that more features can increase the precision of the classifier even if the accuracy rates are lower.

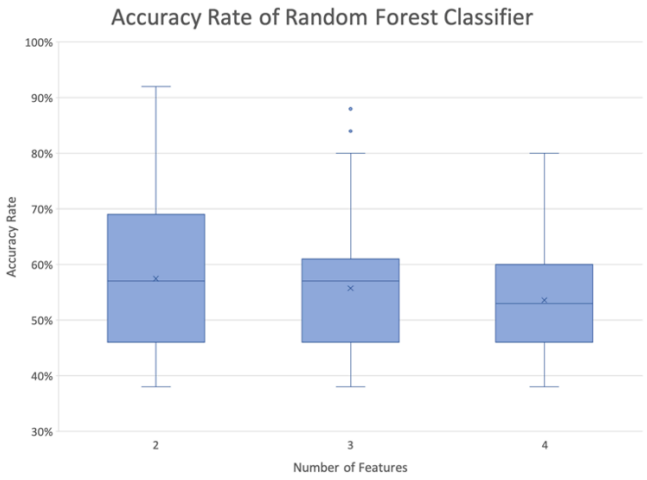

Figure 3. Box and Whisker Plot of Accuracy of Naive Random Forest Classifier

6.4 Support Vector Machine

Figure 4 illustrates a box and whisker plot of the accuracy rates compared to the number of features used in the support vector machine classifier. The highest accuracy rates were when two or three features were used. Overall, there was a slight decrease in accuracy rates when four features were used, but median and average accuracy rates increased. The highest rates for 2 or 3 features were 80% and involved the features: title sentiment, hashtag count and transcript sentiment. Overall, hashtag count, title sentiment and transcript sentiment resulted in the highest accuracy.

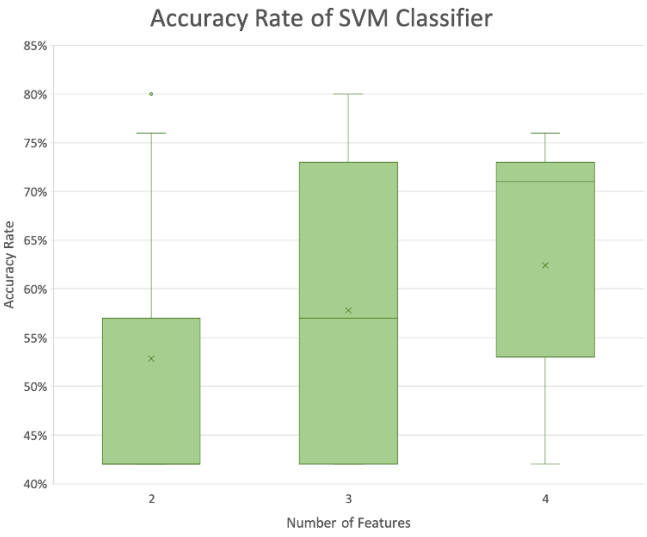

Figure 4. Box and Whisker Plot of Accuracy of Naive Support Vector Machine Classifier

6.5 Logistic Regression

Figure 5 illustrates a box and whisker plot of the accuracy rates compared to the number of features used in the Logistic regression classifier. The highest maximum accuracy rate was 84% and involved using three features. The features used that had the highest accuracy rate for this classifier were: title sentiment and hashtag count. Using a higher number of features in the feature set did result in slightly better average accuracy rates even though the highest score occurred when using three features.

Both the SVM and logistic regression classifiers show similar patterns of standard deviation being lowest when four features were used and highest when three features were used. This indicates that even though using three features in the feature set resulted in the highest maximum accuracy rate, it also decreased the precision of the classifier, whereas using four features increased the precision of the classifier.

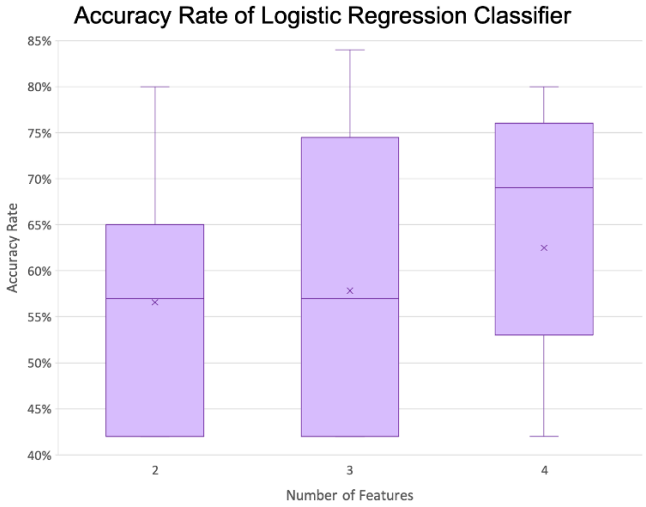

Figure 5. Box and Whisker Plot of Accuracy of Logistic Regression Classifier

7. Limitations and Next Steps

We acknowledge that there are limitations in this research. One limitation was the use of only two different data sources, YouTube and BitChute and the challenges of finding misinformation on YouTube. Due to the fact that YouTube began to moderate COVID-19 misinformation rather quickly, we relied on obtaining genuine information from YouTube and misinformation from BitChute due to its low moderation guidelines. A second limitation in this research concerns the relatively small sample size of 32 videos as well as the adoption of the 80/20 testing / training set. A source for further investigation, we discovered that models underperformed as the training set grew. Future goals would be to automate the process of obtaining transcripts and preprocessing MeSH terms and sentiment. This would allow for a much larger subset of videos to become available for testing and training.

Additionally, using classical machine learning algorithms, our highest accuracy rate of 92% performed worse than other fake news detectors with accuracy results above 95%. Other classifiers including neural networks working on larger data sets and additional data features may allow us to achieve these goals. Additionally, the average length of each video was long. A better approach may be to segment larger transcripts to create both larger datasets, and smaller video segments for analysis, which can identify 'genuine' information embedded within a misinformation video. Finally, more extensive research could utilize computer graphics to analyze visual cues from speakers and props, rather than to rely solely on video transcripts as performed in this research.

8. Conclusion

The spread of health-related misinformation is an important problem to resolve [35]. In this research, we break down transcripts from publicly available online social media videos related to the COVID-19 vaccines. In addition to available metadata surrounding each video, new metadata was generated using transcripts of each video, including transcript sentiment, title sentiment, and medical subject headings (MeSH) terms. Four classifiers were used to detect misinformation with varying levels of accuracy regarding the COVID-19 vaccine specifically. While random forest classifier resulted in the highest accuracy rates using two features of 92%, naïve Bayes performed the best with the greatest number of features across multiple feature sets (4 in total) and achieved accuracy rates of 84%. Logistic regression and SVM performed poorest across classifiers, achieving accuracy rates only around 80%. Additional discoveries found title sentiment to be a significant feature in all but the top-most performing classifiers.

9. References

[1] Zarocostas, J. (2020). "How to fight an infodemic," *The Lancet* 395(10225), pg. 676.

[2] Roozenbeek, J. Schneider, CR, Dryhurst, S. Kerr, J. Freeman, ALJ Recchia, G. van der Bles, AM, and van der Linden S. (2020). "Susceptibility to misinformation about COVID-19 around the world," *The Royal Society* v7(201199).

[3] Romer, D. and Jamieson, KH (2020). "Conspiracy theories as barriers to controlling the spread of COVID-19 in the US," *Social Science & Medicine*, v263.

[4] Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam MB, Karam, B, Adib, E., Zarka, J., Traboulsi, C. Akl, EW and Baddour, K. (2020). "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter," *Cureus*, v12(3).

[5] Hassan, N., Arslan, F., Li, C., and Tremayne, M. (2017). "Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster," *In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, pp. 1803–1812.

[6] Nguyen, TT, Weidlich, M., Yin, H., Zheng, B., Nguyen, QVH, and Stantic, B. (2019). "User guidance for efficient fact checking," *In Proceedings of the VLDB Endowment*. v12(8). April 2019, pp. 850–863.

[7] Sommariva S., Vamos, C., Mantzarlis, A., Uyên-Loan Đào L., and Martinez Tyson D. (2018). "Spreading the (Fake) News: Exploring Health Messages on Social Media and the Implications for Health Professionals Using a Case Study," *American Journal of Health Education, v49(4), pp 246-255.*

[8] Serrano, JCM, Papakyriakopoulos, O. and Hegelich, S. (2020). "NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube," *In Proceedings of the 1st Workshop on NLP for COVID-19* at ACL 2020, Online, July 2020. Association for Computational Linguistics.

[9] Tankovska, H. (2021). "Daily time spent on social networking by internet users worldwide from 2012 to 2020," *Statista*. Originally published on Feb 8, 2021.

[10] Shearer E. and Gottfried, J. (2016). "News use across social media platforms. News use across social media platforms," *Pew Research Center*, May 2016.

[11] Shao, C. Ciampaglia, GL, Varol, O. Yang, K. Flammini, A., and Menczer F. (2018). "The spread of low-credibility content by social bots," *Nature Communications*, 9(1):4787, December 2018.

[12] "Media Usage in an Internet Minute as of August 2021," *Statista*, 2023. Accessed 2023-06-04: https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/.

[13] "Forecast of the number of Youtube users in the World from 2017 to 2025," *Statista*, 2022. Accessed 2022-06-04: https://www.statista.com/forecasts/1144088/youtube-users-in-the-world.

[14] Basch CH, Zybert P., Reeves R., Basch CE (2017). "What do popular YouTube videos say about vaccines?" *Child Care Health Dev* 2017, v43(4) pp. 499-503.

[15] Bora K, Das D, Barman B, Borah P. "Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015-16 Zika virus pandemic," *Pathogens and Global Health,* 2018 Sep 29;112(6) pp320-328.

[16] "YouTube Terms of Service," *YouTube.com* (2023). Accessed 2023-06-04: https://www.youtube.com/static?template=terms.

[17] "YouTube COVID-19 Medical Misinformation Policy," *YouTube.com*, May 20, 2020. Accessed 2022-06-04: https://support.google.com/youtube/answer/9891785.

[18] Trujillo, M. Gruppi, M. Buntain, C. and Horne, BD (2020). "What is BitChute? What is BitChute? Characterizing the "Free Speech" Alternative to YouTube," *In Proceedings of the 31st ACM Conference on Hypertext and Social Media*. Association for Computing Machinery, New York, NY, USA, pp. 139–140.

[19] "BitChute Guidelines," *BitChute.com*, 2022. Accessed 2022-06-05: https://support.bitchute.com/policy/terms.

[20] Zhao, Y. Da, J. and Yan, J. (2021). "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches," *Information Processing & Management*, v58(1).

[21] Mitchell, T. (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7.

[22] Aphiwongsophon, S. and Chongstitvatana, P. (2018). "Detecting Fake News with Machine Learning Method," *In Proceedings of 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (ECTI-CON), pp. 528-531.

[23] Silva, M., Ceschin, F., Shrestha, P., Brant, C., Fernandes, J., Silva, C. S., Gregio, A., Oliveira, D., and Giovanini, L. (2020). "Predicting Misinformation and Engagement in COVID19 Twitter Discourse in the First Months of the Outbreak," arXiv:2012.02164.

[24] Gerts D., Shelley CD, Parikh N., Pitts T., Watson Ross C., Fairchild G., Vaquera Chavez, NY and Daughton, AR (2021). "Thought I'd Share First" and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study," *JMIR Public Health and Surveillance*, v7(4).

[25] Lin, W., Hu, Y. and Tsai, C. (2012). "Machine learning in financial crisis prediction: A survey," *IEEE Transactions on Systems, Man, and Cybernetics*, Part C (Applications and Reviews), v42(4), pp. 421-436.

[26] Tacchini, E., Ballarin, G. Della Vedova, ML, Moret, S., and de Alfaro, L (2017). "Some like it hoax: Automated fake news detection in social networks," arXiv preprint arXiv:1704.07506.

[27] Shearer C. (2000). "The CRISP-DM model: the new blueprint for data mining," *Journal of Data Warehousing*, pp. 13-22.

[28] Wirth, R. and Hipp, J. (2000). "CRISP-DM: Towards a Standard Process Model for Data Mining," *In Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39, 2000.

[29] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N. Ramírez-Quintana, MJ and Flach, P. (2021). "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, v33(8), pp. 3048-3061.

[30] "About Happy Scribe Automatic Transcript Software," *HappyScribe.com* (2022). Retrieved online on June 4, 2022 from: https://www.happyscribe.com/automatic-transcription-software.

[31] Mork, JG, Jimeno-Yepes, A., and Aronson, AR (2013). "The NLM medical text indexer system for indexing biomedical literature," In BioASQ@CLEF.

[32] Lu, H., Ehwerhemuepha, L. and Rakovski, C (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC Medical Research Methodology v22(181).

[33] Cho D., Mork JG, Aronson A., Demner-Fushman D., Schmidt S., Ozga D., Pash J., Kilbourne J. (2015). "MeSH on demand: an easy way to identify relevant MeSH terms and related articles from text [Poster]," *National Cancer Institute (NCI) Symposium* "RNA Biology 2015".

[34] Mourad, A. Srour, A., Harmanani, H. Jenainati, C. and Arafeh, M. (2020). "Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions," *IEEE Transactions on Network and Service Management*, v17(4), pp. 2145-2155.

[35] Zeng, J. and Chan, C. (2021). "A cross-national diagnosis of infodemics: comparing the topical and temporal features of misinformation around COVID-19 in China, India, the US, Germany and France," Online Information Review, v45(4), pp. 709– 728.

[36] Hutto, CJ and Gilbert, EE (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, (20 14). Ann Arbor, MI, June 2014. Available at https://www.nltk.org/api/nltk.sentiment.vader.html and https://www.nltk.org/howto/sentiment.html, Accessed: 2021-05-19.

[37] Rovetta, A. and Srikanth Bhagavathula, A. (2020). "Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags," *Journal of Medical Internet Research*, v22(8).

[38] Shariff, M. Thoms, B. and Isaacs J. (2022). "Approaches in Fake News Detection: An Evaluation of Natural Language Processing and Machine Learning Techniques on the Reddit Social Network," *In Proceedings of the 8th International Conference on Data Mining and Applications* (DMA 2022).

[39] Yammine, S. (2020). "Going viral: how to boost the spread of coronavirus science on social media," *Nature*, 581(7808), pp. 345–347.

[40] Toikka, T. (2017). *Feasibility of selected machine learning methods for failure forecasting of aeroplane flight control surfaces*, Master's thesis, Tampere University of Technology.