The Spread of YouTube Misinformation Through Twitter

Anamika Gupta and Alisha Seghal
DSC 180B
March 9, 2022

Introduction

Millions of people use platforms such as YouTube, Facebook, Twitter, and other social media networks. While these platforms grew popular for their social aspects of connecting people, they have also become popular ways to share and consume news. Since these platforms are so accessible, information spreads rapidly and virally. One key issue is that social media can be a core source of misinformation as these platforms are often used to establish a narrative and conduct propaganda without verification or fact-checking. Over the past decade, the proliferation of misinformation has created concern in terms of social progress, politics, education, and national unification. Reports from the Pew Research Center show that 64% of Americans are confused about current events because of the rampant presence of fake news on social media and 23% have passed on misinformation to their contacts both intentionally and unintentionally [1]. Thus, it's clear that misinformation spreads very easily on social media platforms compared to other avenues of communication.

People are increasingly engaging with content that is often flashy and spreading misinformation (i.e conspiracy theories) but not engaging in fact-checking with the same fervor. Fact-checking and verification of online information is also a complicated task. Many accounts do not represent real people, posts can be sponsored, some users may be bots, and political affiliations are usually not disclosed. Sometimes it is impossible to differentiate between genuine content and content that is intended to manipulate opinions. This makes it difficult to validate information with the large volume of content churned out daily even for the most diligent and fact-checking individuals. As a result, many platforms have begun implementing more fact-checking to combat misinformation at a wider scale but the effectiveness of these initiatives is unknown.

Misinformation has been shown to mobilize people in dangerous ways and distract people from truthful cases of wrongdoing or public safety. Through this project, we dissect the spread of misinformation regarding public health over a time period in which the nation experienced major public health and safety issues such as the COVID-19 pandemic, mask mandates, uncertainties regarding medical treatments and development of new vaccines. This investigation seeks to understand how Twitter and YouTube's platforms interacted and aided the spread of misinformation by examining the video captions extracted from YouTube videos linked in tweets discussing anything health-related. The video captions and other YouTube and Twitter metadata were analyzed using NLP to identify false statements or misleading content. Ultimately, this work can help determine how to reduce the spread of misinformation by understanding effective policies against misinformation and creating better misinformation detection pipelines.

Literature Review

The spread of misinformation on social media platforms has been researched extensively by past projects. Both Twitter and YouTube are cognizant of the harmful effects of misinformation on their platforms and have taken steps to identify misleading content and limit its impact by either removing it or adding a warning label based on its propensity for harm [2], [3]. Yet, these methods are limited by the platforms' ability to accurately identify misinformation. With the large volume of content produced online, these companies must rely on automated detection of misinformation instead of manual efforts through their employees. Identifying misinformation first requires a clear consensus on fact-checking material like information from government organizations or research bodies. It can be difficult to reach agreement on what qualifies as misinformation and what does not. Additionally, social media platforms can be wary of taking an overly aggressive approach to removing content in an effort to maintain open communication and free speech. Thus, it is important to measure how well Twitter and YouTube are able to remove misinformation.

Furthermore, content is often exchanged between social media platforms making it important to study how misinformation might be propagated between Twitter and YouTube. One study found YouTube to have the strongest association with conspiracy beliefs [4]. As the second-largest social media platform, content from YouTube is shared or linked on other platforms like Facebook, Twitter, and Reddit. Knuutila et al. used posts from these platforms that linked to YouTube videos to measure how effective YouTube's policies were at removing misinformation [5]. This approach enabled the authors to discover which videos were removed and why, using the Wayback Machine, a digital archive of the internet. Additionally, this work explores how much misinformation transfers from one platform to another by looking at sharing statistics and other metadata.

Other works also revolve around methods to detect misinformation in YouTube videos. One study suggests a data-focused approach to identify content on social media through lexical and syntactical features from a document along with social context features to train a model based on fact-checking content and propagation [6]. This approach can be applied to any text like the captions of a video and metadata available through the platform like engagement statistics. Jagtap et al. focus upon extracting video captions from the YouTube API then applying NLP techniques to classify a video as misinformation or not [7]. Their findings show that training word embeddings on Google News and Wikipedia articles can result in classifiers with high F1 score and accuracy. For YouTube videos dealing with vaccines controversy, they found that a Support Vector Classifier had the best performance. Some studies also explored the role of comments in propagating misinformation on YouTube. One paper discusses analyzing user engagement through comments to help detect misleading content and determine if comments themselves are "inorganic" or coming from bot-like sources [8]. This paper looks at the behavior of commenters in multiple ways, including building video-commentator and commenter-comment networks, and sentiment analysis of top comments, in order to determine how comments can contribute to the spread of misinformation. This investigation combines these approaches to better evaluate how misinformation spreads between social media platforms and how effectively platforms can detect misinformation using automated approaches.

Methods and Data

Datasets

This work analyzes the video captions of YouTube videos linked in any tweets related to public health. This data was compiled by first gathering individual tweets. While Twitter does not provide access to all tweets directly, this data was acquired by compiling daily tweet ids from Panacea Lab's Covid-19 Twitter Chatter dataset spanning from March 22, 2020 to September 30, 2021 [10]. Using the Twitter API and the Twarc2 python library, tweet ids were sampled and hydrated into complete tweet objects which provided more complete data about the tweets including the text, author id, urls, date, and other useful information. These tweets were then filtered to check if the text content of the tweet contained any key terms relating to public health. Additionally, tweets were filtered to check if they contained any links that led to a YouTube video. Any such tweets were compiled into a dataset with the hydrated tweet objects and the YouTube video ids were stored separately.

The resulting dataset of video ids was then used to create YouTube data using the YouTube Data API and the YouTube Web Client. Special considerations were taken when fetching the video captions or subtitles. Firstly, not every video had captioning provided by the video creator. In that case, YouTube will often auto-generate captions which will not be completely accurate. Other times, the captions are in another language but the YouTube API provides the ability to translate captions into a specified language. In both situations, this dataset uses the available English captions whether they are autogenerated or translated. Additionally, the YouTube API can be used to find other information about a YouTube video using the video id. Our dataset contains relevant data including video title, date posted, like count, comment count, view count, description, video tags, and category that can be useful in determining if a video is contributing to the spread of misinformation.

One key consideration for the dataset was missingness. Between the date they were posted and then accessed, tweets can be removed by Twitter for violating the platform's policies or even deleted by the user. Twitter removes many tweets that the platform detects to have a high propensity for harm and contain misleading information and YouTube has similar platform policies. Thus, tweet and video missingness is not independent of misinformation and we will have to take this into account in our analysis.

Misinformation Detection

The focus of this research is to detect misinformation on YouTube which was accomplished by building a NLP model to categorize whether the text of YouTube video captions contained false and misleading information or not. Since our dataset of YouTube videos were not labeled, our model was trained on a pre-labeled dataset containing the text of articles with real and fake news from Kaggle [11]. We also created a text processing pipeline to prepare the caption texts for analysis. This included normalizing the case of the text, removing punctuation, removing stop words, tokenization, and lemmatization. Stop words are the most common words in language like "a", "the", "is", "are", etc. which do not add any context or information so removing them is important to ensuring the model focuses on the relevant terms. Tokenization is the process of splitting the text into individual words or sentences so that the model can work with smaller pieces of data that are still coherent and relevant to the context outside of the text. Lastly, lemmatization is the process of converting a word to its lemma, or returning an inflected word to its root word, in contrast to stemming which simply removes the suffix of a word.

Before this cleaned text data can easily be used to build a misinformation detection model, it has to be first transformed into a feature vector. For this investigation, we created a TF-IDF vector. TF-IDF stands for term-frequency inverse-document-frequency and it calculates how much a token appears in a specific document as compared to the entire text corpus. As a result of this calculation, TF-IDF is able to identify how important a word is, so in the vector the weight assigned to each token not only depends on its frequency in a document but also how recurrent that term is in the entire corpora.

After vectorization, we build the final model. We tried multiple classifiers including Logistic Regression, Naive Bayes, Decision Tree and finally found that sklearn's Passive Aggressive Classifier had the best performance on the training dataset which was then used to detect misinformation from the YouTube videos.

Topic Modeling and Sentiment Analysis

Topic modeling is a technique to extract the hidden topics from large volumes of text. As an unsupervised process, topic modeling is quite valuable as it is able to discover hidden semantic structures in text. We implemented the Latent Dirichlet Allocation (LDA) algorithm for topic modeling through gensim to determine the dominant topics found in misinformative YouTube videos. LDA builds a topic per document model and words per topic model, using Dirichlet distributions. The algorithm takes in a number of topics, then rearranges the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of topic-keywords distribution.

Next, we performed sentiment analysis on our YouTube videos to gauge if there was a difference in the engagement of positive and negative sentiment videos. We then wanted to see if there was a relation in the sentiment of the videos and whether the videos were spreading misinformation or not.

We trained a Logistic Regression model to predict the sentiment of videos. In order to label our videos for training, we researched what percentage of likes on a video deems it successful. On average, receiving likes that total 4-10% of the total views or above is seen as the baseline for a good video. We decided to take the upper limit of 10% and label all videos with a ratio of likes to views higher than 10% as positive. Any video with a ratio that fell below the threshold was labeled negative. We then fed this labeled data into the model with a 80/20 train test split.

We created two models following this labeling logic. The first model dealt with predicting sentiment based on video titles. The second model predicted sentiment using the video descriptions. Once we finished our models, we plotted our positive and negative sentiment videos to see how the distribution of misinformation spread compared.

Results and Conclusions

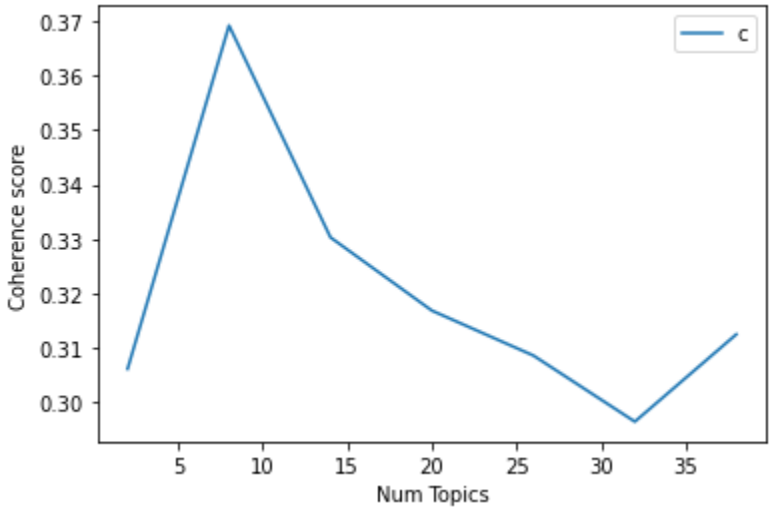
Model Evaluation

We considered several models using TF-IDF vectorization on our text which we evaluated through their accuracy score. We found that the Passive Aggressive Classifier from sklearn had the best accuracy score of 84.6%. This was the expected result as the Passive Aggressive Classifier is suited for online learning that deals with large sets of data and their loss function is passive when dealing with an outcome that has been correctly classified, and aggressive when a miscalculation takes place, thus the model is

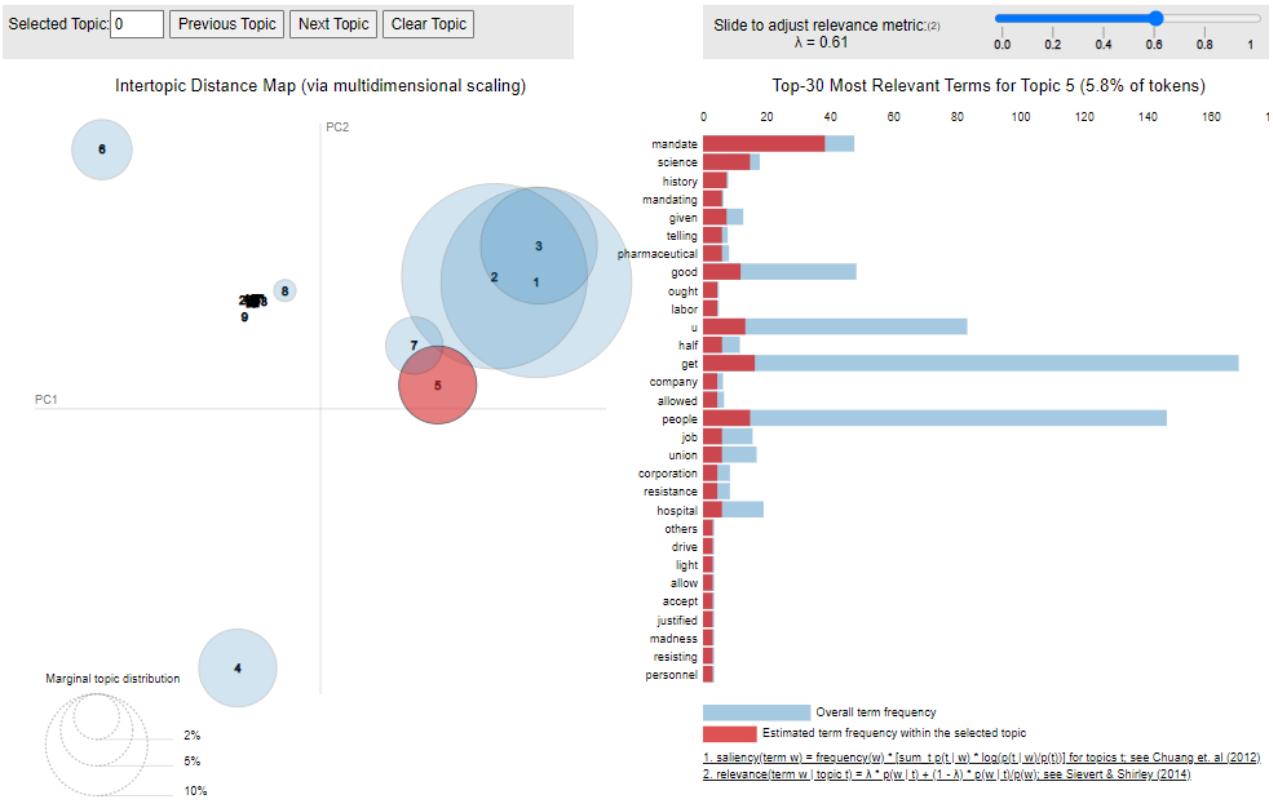
constantly self-updating and adjusting. Running this model on our dataset of YouTube captions resulted in 23% of the video captions being classified as misinformation.

Topic Modeling

The LDA model for topic modeling was built using cleaned captions from the YouTube videos collected from Twitter. It is important to find a good number of topics for the LDA model because it will provide meaningful and interpretable topics. Picking higher numbers can sometimes provide more granular sub-topics but if the number of topics is too high, the same keywords will be repeated in multiple topics. To find the best number of topics for the LDA model, we evaluate model performance using a coherence score. Coherence measures the degree of semantic similarity between high scoring words in the topic. These measurements help truly distinguish topics that are interpretable and understandable. We measured the coherence of LDA models built using 2, 8, 14, 20, 26, 32, 38 topics and plotted them below. As seen in the chart, the LDA model with 8 topics has the highest coherence with a score of 0.369. Thus, this is our optimal model for topic modeling.



The image below is a sample output from pyLDAvis which visualizes the topics produced by the optimal LDA model and associated keywords. Each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent is that topic. The highlighted bubble updates the words and bars on the right-hand side which are the salient keywords that form the selected topic. As seen in this example, topic 5 is focused on vaccine mandates in the workplace.



After discovering the dominant topics across the documents in the entire corpus, we also found the dominant topic in each sentence of the caption texts and most representative caption texts for each topic. Lastly, we looked at the topic distribution among video captions and found that topics 2 and 4 were the most dominant topics, appearing in 53% and 42% of video captions texts respectively. Topic 2 was focused on the origins of COVID-19 with keywords like coronavirus, China, Wuhan, and lab while Topic 4 was regarding COVID-19 vaccines with keywords such as vaccine, booster, and dose. Topic modeling provides a deeper insight into the topics covered in YouTube videos linked in Twitter. Clearly, contentious subjects that are often sources of misinformation are frequently spread on social media platforms as evidenced by the dominant topics we discovered..

YouTube Video Sentiment

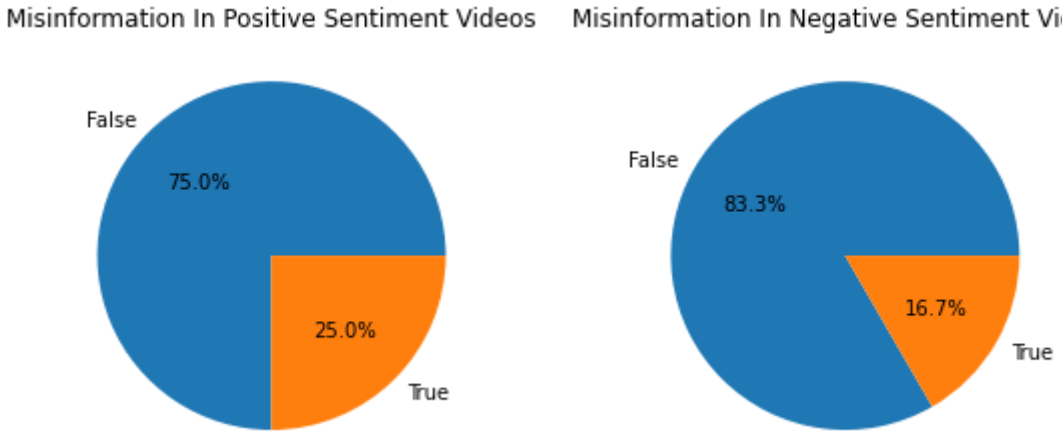
We created two logistic regression models to predict the sentiment of videos. The model based on video descriptions, with an accuracy of 0.83, performed better than the one run on video titles, which had 0.75 accuracy.

Looking into the video descriptions further, we were able to isolate the terms that appeared most frequently in both the positive and negative sentiment videos. The following are the most frequently found terms in positive videos and negative videos respectively:

Videos about general news channels seemed to have a more positive reaction from viewers than videos about specific platforms like Instagram, Twitter, Facebook, tv9kannada, or CTV news. China and omicron as terms have a higher appearance in videos that were taken negatively. African Diaspora News and non-English words appear in high amounts in positively received videos.

Sentiment Vs Misinformation

We found that the videos with positive sentiment had a slightly higher rate of misinformation (25%) vs negative videos containing misinformation (~17%).



These results raise an interesting question of how people react to videos. Are they reacting more positively when given specific types of information or are people being exposed to misinformation content that is more likely to yield a positive reaction?

Missingness

Missingness can be caused by several factors on both Twitter and YouTube. Oftentimes, users themselves can remove content, but the platforms themselves can also remove any content they believe violates their policies. With misinformation on the rise, social media platforms have developed policies to combat this and will sometimes remove content that contains false or misleading information. As a result, missing tweets and videos are not trivial and cannot be ignored in the conclusions we draw. After fetching the tweet objects from the Twitter API, we conducted some exploratory data analysis to gain an overall understanding of the available tweets. First, we calculated the number of tweets that were hydrated into tweet objects and found the number of tweet ids that could no longer fetch tweets because they had been

removed from the platform. To do so, we randomly sampled 10,000 tweet ids from the full tweet id dataset to build a 95% confidence interval of the proportion of missing tweets. This method resulted in an interval of (0.19, 0.37). As mentioned earlier, we are not able to verify the reason the tweet is unavailable but there is a higher chance that it was removed for violating Twitter's policies including promoting misinformation. Similarly, any YouTube videos linked in the missing tweets have a stronger likelihood of being related to misinformation.

Future Exploration

The distribution of misinformation vs fact-checked information between positive and negative sentiment videos shows us that there may be a higher rate of positive engagement among misinformation related videos. This could be due to how social media platforms' algorithms promote the misinformation videos or may be due to how viewers intake the information differently. Looking into this further will help us further find ways for social media platforms to better detect and reduce the spread of misinformation.

This work has the potential to expand beyond Twitter and apply to other social media platforms like Facebook, opening the door for comparative analysis to see how different communication forms affect the spread of misinformation. Most YouTube videos linked in tweets linked Instagram and Facebook accounts, websites, and subscription services for their viewers. A multidimensional analysis on these connections between platforms could be an interesting next step. This avenue of exploration will also help determine the most effective ways that these platforms can detect and remove misinformation before it has the chance to spread. One issue faced in this investigation was the sparse presence of YouTube videos linked in tweets. This could be resolved by searching for video ids on other platforms or expanding the domain beyond public health and including other controversial topics such as 9/11 conspiracy, election fraud, or Flat Earth theory. Increasing the size of the YouTube caption dataset would help improve the misinformation detection model. Another way to improve this model would be to train it on a subset of labeled caption data. Furthermore, this version of the model uses a TF-IDF vector as its embedding vector. This could be replaced by other word-to-vector embeddings such as GloVe or word2vec. Further improvements to automated misinformation detection, especially on autogenerated or translated text content as this investigation focused on will greatly help prevent the spread of harmful misinformation on social media platforms.

References

[1] Barthel M, Mitchell A, Holcomb J. Many Americans Believe Fake News Is Sowing Confusion; 2016. Available from: http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/.
[2] N. Mohan, "Perspective: Tackling Misinformation on YouTube," blog.youtube.com, Aug. 25, 2021. [Online]. Available: https://blog.youtube/inside-youtube/tackling-misinfo/. [Accessed: Dec. 4, 2021].
[3] Y. Roth, N. Pickles, "Updating our approach to misleading information," blog.twitter.com, May 11, 2020. [Online]. Available: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information. [Accessed: Dec. 4, 2021].
[4] D. Allington, B. Duffy, S. Wessely, N. Dhavan, J. Rubin, "Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency," Psychological Medicine, vol. 51, no. 10, pp. 1763–1769, 2021.
[5] A. Knuutila, A. Herasimenka, H. Au, J. Bright, R. Nielsen, P. Howard, "COVID-Related Misinformation on YouTube: The Spread of Misinformation Videos on Social Media and the Effectiveness of Platform Policies," COMPROP Data Memo, vol. 6, pp. 1-7, 2020.
[6] K. Shu, A. Sliva, S. Wang, J. Tang, & H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," Sigkdd Explorations, vol. 19, no. 1, pp. 22–36, 2017
[7] R. Jagtap, A. Kumar, R. Goel, S. Sharma, R. Sharma, C. George, "Misinformation Detection on YouTube Using Video Captions," 2021.
[8] M. N. Hussain, S. Tokdemir, N. Agarwal and S. Al-Khateeb, "Analyzing Disinformation and Crowd Manipulation Tactics on YouTube," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1092-1095, 2018.
[9] J. Pennington, R. Socher, C. Manning, "GloVe: Global Vectors for Word Representation," nlp.stanford.edu, 2014. [Online]. Available :https://nlp.stanford.edu/projects/glove/. [Accessed: Dec. 4, 2021].
[10] Banda, J, "A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration", 2021. Available from: https://doi.org/10.3390/epidemiologia2030024.
[11] Bisaillon, C, "Fake and real news dataset: Classifying the news", 2021. Available from: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset