GitHub:https://github.com/fieryashes/DSC180B_Misinformation_Project

Website:https://anaaamika.github.io/DSC180B-Misinformation/

Schedule:https://docs.google.com/document/d/1XDf9fIPS1Vmpk1xjQAxqhd93MKVWK-XwUrFiEUFFLwc/edit#heading=h.q9ym84d2qiu7

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 sensationalized content and spreading misinformation such as 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 threats. As one of the largest social media platforms,

Twitter is a major source of misinformation. It is very possible that the spreading of misinformation occurs in an organized effort with malicious intent. Clusters of tightly linked accounts play a large role in the spread of fake news and disinformation. Some of these accounts are not managed by real people and instead are automated bots. Even some high profile accounts circulate misinformation.Developing effective policies to tackle misinformation on social media platforms requires evaluating the magnitude of misinformation on Twitter and the avenues through which it spreads. As a result, we try to model a misinformation network on Twitter and understand the key characteristics regarding its participants and the flow of information.

What Problem We Are Considering

In our Capstone Project, we want to explore the spread of misinformation online. More specifically, we will look at the spread of misinformation across Twitter and YouTube because of the large role these two social media platforms play in the dissemination of news and information. Our main objectives are to understand how YouTube videos contribute to spreading misinformation on Twitter, evaluate how effectively YouTube is removing misinformation and if these policies also prevent users from engaging with misinformation. We will be taking a novel approach of analyzing tweets, YouTube video captions, comments, and archived videos using NLP to determine the presence of misinformation and investigate how individuals interact or spread misinformation. Our research will focus on the domain of public health as this is the subject of many conspiracies, varying opinions, and fake news.

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. Our investigation will be combining these approaches to better evaluate

how misinformation spreads between social media platforms and how effectively platforms can detect misinformation using automated approaches.

What We Hope to Learn

Beyond gaining a general understanding of the spread of misinformation on YouTube, we want to answer the following questions.

- How much public health misinformation spread on Twitter is from Youtube?
- How effectively does YouTube's platform detect public health misinformation?
- How do YouTube comments aid in spreading public health misinformation?

Description of Methods

Gathering Tweets

The first step we will be taking will be to gather tweets using the Twitter API. In Quarter 1, we collected all of the tweet ids spanning from March 22, 2020 to October 10, 2021. We will expand this set to include data from before the COVID-19 pandemic and add any tweets from January 1, 2020 to the dataset. We will rehydrate these tweet ids into full tweet objects using the Twarc python library. From there, we will select any tweets with hashtags and text that include health related keywords (using a software generated corpus) and a link to YouTube. We will be extracting the YouTube links and adding them to a dataset.

Extracting YouTube Links

From our dataset of selected tweets, we will be extracting the YouTube links and using the YouTube API to get data regarding the videos. We will be collecting the video captions, comments, comment metadata like user, likes, replies, and video metadata such as channel, views, likes, dislikes, date posted, video description. Since the focus of our project is to use NLP to determine misinformation in a video, we will not include any videos that do not have captions available in our final dataset.

Missing Videos

We also have to account for any broken links or links to YouTube for videos that are no longer on the platform. YouTube might remove videos for several reasons, including copyright violation, inappropriate content, harassment, hate speech, or misinformation. Additionally, a user might remove their own video. Thus, we cannot claim that all broken links lead to misinformation. However, to create a clear picture of why a video was removed, we will be recording the reason YouTube gave for the video's removal. From there, we will be using the Wayback Machine API to view the view counts, channel subscriber counts, full descriptions of the videos, and the video's creation date. Note, that not all the videos will be archived on the Wayback Machine since more popular and widely shared videos are more likely to be collected and saved in the archive thus we will have to account for that bias.

Analyzing Video Transcriptions

After we've collected the video caption texts from the YouTube videos, we will be using NLP to detect if the video propagates misinformation or not. We will then pre-process the texts by removing any special characters and removing texts with fewer than 500 characters to reduce noise in the dataset. From there, we will generate numerical representation vectors to represent the texts from the GloVe Wikipedia embedding [9]. We will then build a Support Vector Classifier to determine if video is misinformation based on its captions.

Analyzing Comments

We will also be considering the comments associated with a video to check the spread of misinformation. We want to determine if the comments are engaging in further misinformation, fact-checking or neither. To accomplish this, we will be building a network of commenters to see if the same users comment on multiple videos and conduct sentiment analysis on the text of the comment. This will help us determine if comments come from bot-like or spam accounts. Additionally, this will help us understand if comments can also manipulate users and promote disinformation.

Preliminary Results

Keywords and Tweet Filtering

In order to extract relevant tweets, we decided to create a corpus of health terms. At first, we tested our code on a manually created text file of health terms. However, we found that our limited terms yielded few to no tweet matches and decided that we needed a more comprehensive term list. We looked into doing a keyword extraction from several documents and articles related to health. However, we found an online tool called SketchEngine that would perform this for us, and we decided that using this tool would be the most effective and efficient way to move forward. Within the tool, we were able to input our initial corpus of terms and then choose articles that they automatically identified as relevant online. Then the tool ran through the document extracting words that related to our terms and health as a category in general. It gave us a few files at the end:

- Health Document Corpus - This file contained excerpts of significance from the articles and documents it curated about health.
- Health Term Corpus - This file consisted of key phrases and multi-word terms that were found in the health documents at a significant level (and thus were deemed relevant to health).
- Health Word Corpus - This file lists all single-word terms (aka words) that were found as significantly relevant in the health documents and articles.

Within the Health Term and Word Corpuses, we were given the strength of the terms and words as well. They provided the following values:

Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)	Score

The Frequency (focus) and the Frequency (reference) are the absolute counts of a word within the corpus. They cannot readily be used to compare a word's significance since the corpus consists of documents that differ in length. The Relative Frequency (focus) and Relative Frequency (reference) columns give us the number of times a word appears every million words. By describing the frequency in every million words, we are able to compare the frequencies of words across different length documents. SketchEngine also returns a score to show how significant a

word is within the corpus. This score is also known as the Keyness Score and is calculated via the following formula:

$$\{fpm\_\{rm\ focus\} + N\}/\{fpm\_\{rm\ ref\} + N\}$$

- **fpm_{rm focus}** is the normalized (per million) frequency of the word in the focus corpus
- **fpm_{rm ref}** is the normalized (per million) frequency of the word in the reference corpus
- **N** is the so-called smoothing parameter (N = 1 is the default value)

A high score correlates with a high frequency and a low score with a low frequency.

We used the Health Word Corpus as our new term list to extract relevant tweets from. We plan to next use the document corpus to do a bag of words and other NLP analysis on our youtube video transcript data. The Health Term Corpus could be an interesting additional tool to use when doing the n gram analysis on our transcripts.

Missingness

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 (0.21, 0.43). 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. These unknowns make it more difficult to do a comprehensive analysis of how YouTube content and other sources of misinformation are spread through Twitter. We are concerned that this missingness can introduce bias into our investigation as tweets that violate Twitter policies regarding COVID-19 and public health misinformation are more likely to be removed from the platform than tweets that do not have such violations. As a result, our dataset may become skewed towards tweets that do not contain misinformation.

YouTube Transcripts

We successfully translated the twitter links into inputs for the YouTube API which requires video IDs rather than links to return metadata and transcripts. To work around the ID requirement, we first researched whether the link to a youtube video contains the ID and found that the numbers at the end of the youtube video link consistently created the video link. We then used get requests on the twitter links, found the actual links within the tweets, extracted the links going to youtube, and appended all the end numbers as the video ids of interest into a list.

Within our code to extract the transcripts from youtube videos, we find and use the following information:

- The "videoid" option specifies the YouTube video ID that uniquely identifies the video for which the caption track will be uploaded.
- The "name" option specifies the name of the caption track to be used.
    - Defaults to "YouTube for Developers"
- The "file" option specifies the binary file to be uploaded as a caption track.
- The "language" option specifies the language of the caption track to be uploaded.
    - Defaults to “en”
- The "captionid" option specifies the ID of the caption track to be processed.
- The "action" option specifies the action to be processed.
    - Defaults to “all”

We found that some of the tweets had broken links in them. Their corresponding videos did not exist anymore and we wanted to make sure to catch these links before inputting them into the API. We wrote a simple try/except clause that would print out the status and content of the error whenever the links resulted in an HttpError.

Model For NLP

We have looked into a few models to proceed our Natural Language Processing analysis on the video transcripts and their comments. The following are some that we have identified as applicable to our needs of classifying videos by their transcripts and identifying the sentiment of comments.

MonkeyLearn

One potential route for text classification is to use MonkeyLearn. They have a strong model used by businesses in a wide array of industries for taking large texts and outputting the overall idea or key term based on predetermined tags. This option does have a limitation in terms of how many texts we can feed through in the free trial. If we are unable to process all the Youtube video transcripts via the free trial, we will likely explore more models on HuggingFace to classify our videos.

Otherwise, we hope to use the Text Classification with Machine Learning - Topic Modeling model outlined on their website. We would input the video transcripts to the model and the tags we hope to classify the videos as:

- Medical Emergency
- Vaccine
- Covid-19
- Testing
- Hospital Update
- Etc. (Health-related subcategories)

The online tool lets us manually train the model by tagging specific texts it extracts from our dataset and then learns the trend. We hope that our ability to extensively train the model and add onto its training easily will help us achieve more accurate results in classifying the videos.

HuggingFace

On this website, we found a sentiment analysis tool that returns the potential for the text inputted to be positive or negative (by %). We chose this model due to its accuracy of 91.3 on the development set and its ability to correctly classify complex phrases like “I want to hate you but I love you.” This phrase yielded 100% positive whereas other sentiment analysis tools we found had overall negative or fairly neutral results. The max length of inputs however is 128 and thus this will only be applicable for comments on youtube videos.

Using this tool, we will be able to determine the sentiment users have towards videos that are flagged to have misinformation vs those that are fact checked and accurate. Due to the nature of HuggingFace, we will be able to access the model with their hosted inference API.

—

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].

Project Proposal