COSC 419F
FINAL PROJECT REPORT
December 7th 2021

Analysis of TikTok using binary classification on metadata to predict for the presence of misinformation

Joe Gaspari
- Sentiment analysis, unigram and bigram creation

Hecheng Chen
- Research, work on reports, video editing

Patrick Mahler
- Data collection and formatting

Mason Plested
- Model training, analysis

1 Introduction

Information and the spread of misinformation is a growing concern in the social media sphere as an increasing number of typically younger individuals are becoming influenced by the conversations they see on platforms such as TikTok. We chose to examine the spread of misinformation surrounding the COVID-19 pandemic through the TikTok platform. There is a wealth of information that people share on TikTok and consequently, people are misled by false information shared from popular accounts. We selected Tik Tok because it “Has been downloaded more than a billion times, becoming the most downloaded non-gaming app in 2020”. (Cervi, 2021) We believe this large number of downloads represents a large sphere of influence.

On social media platforms, people can be exposed to misinformation which can have damaging effects on their well-being. This can also damage the well-being of our society as a whole if misinformation is not moderated. Misinformation can be attributed for some of the hesitancy we’ve seen in 2021 in people who are hesitant to get the vaccine (CBC, 2021). As more people continue to use social media as a part of their daily lives, it is possible they are exposed more readily to the damaging effects that misinformation can have on not only their well-being, but the well-being of society as a whole.

2 Objective

By looking at data from TikTok regarding the pandemic, we can gather insight into the source of misinformation on the platform. The goal of this study is to try and accurately identify posts on TikTok that may contain misinformation on the pandemic. This is the first step in understanding and addressing the spread of misinformation on TikTok.

Ideally, this information could be used in conjunction with more intensive analysis either performed by human moderators or advanced video content analysis to find and remove posts containing misinformation. We hope to use only easy to collect and compute metadata in order to make this process as efficient as possible while still remaining accurate. We hope this novel approach can speed up the process of identifying posts which may contain misinformation to drastically increase the speed posts which may mislead users can be identified and removed.

3 Methodology

To conduct our study, we needed to scrape data from TikTok before we could do our analysis. We explored a few different unofficial APIs, but we ended up using the Unofficial TikTok API in Python created by David Teather. The primary reason being that we wanted to use Python for our analysis due to the wealth of available libraries and analysis tools, so it made the most sense. This API allowed us to search for TikTok videos that contain specific hashtags in their descriptions, which allowed us to narrow down the results to videos that are related to the pandemic. This returns a JSON object for each TikTok containing all available metadata for each post, which we could filter down to only what was useful to us.

Hashtags we selected to use were: antivaxxer, antivax, vaccine, and coronavirus. We created a Python script to return 800 videos, 200 for each hashtag. We can only pull a limited number of videos at a time, which was one of the reasons we were limited to 800. We also needed to manually classify each video, which is another reason we chose to limit the number of posts. The attributes we chose to pull from the videos were the video ID, username of the person who posted the video, the description, the comment count, the play count, and the share count. Username and video ID were used to reconstruct the URL for each post.

To organise all this data before performing our analysis, we placed it into a CSV file. Five additional columns were also created based on calculations performed on the description. These columns represented the sentiment of the description, as well as the percentage match of known “misinformation flag” n-grams (These will be further explained in Data Collection). We then needed to know which posts in this dataset potentially included misinformation. We decided that the best way to do this was to manually go through the list and watch each post and assign it a score of either 0 or 1, representing if the post contained no misinformation or if we believed that it did.

A binary classification model was then trained using our data. The model uses decision trees to determine if a video may contain misinformation based on the data collected and created in the previous steps. New data was collected and manually classified. This was used to test the performance of the model, and a confusion matrix was created to display these results.

3.1 Data collection

As mentioned above, we collected the share count, play count, and comment count for each video. We also created 5 new columns containing sentiment and containment percentage. These are explained below:

3.1.1 Sentiment Analysis:

We postulate that posts which display more extreme polarity between positive and negative are more likely to contain misinformation. To generate the sentiment intensity of each post we utilised the nltk Sentiment Intensity Analyzer which generates a score for pos/neu/neg sentiment. We found that the analysis provided more variability when the textual data was scrubbed for special characters and other markers which cloud the algorithms output. We then combined the data tables to include these three new attributes.

3.1.2 Fake News N-grams:

In order to define a legitimate definition for COVID-19 misinformation we began by exploring a list of article titles provided by the Fake News infordemic research dataset (COVID-19-FNIR Dataset). This study found news stories related to the COVID-19 pandemic between the months of February to June of 2020. We generated source code to scrub the text data and extract significant words that would be used in the n-gram generating process.

We begin generating a list of unigrams to define the overall similarity of words used between post and all fake news titles. We then suggest that word phrases such as bi-grams, and tri-grams could be used to facilitate a stronger indication of fake news propagation within post descriptions. It was decided to only include unigrams and bigrams as they would show relative similarity to those article title topics.

3.1.3 Containment Comparison:

To strongly indicate any post may have misinformation we used a style of set containment between n-grams generated by each post and the fake news n-grams. We began by cleaning each post description, making note to remove non-unicode characters as many emojis are present. We then remove all stopwords and generate all uni-grams, and bi-grams for separate comparisons. Using set containment as a measure of similarity between the post and the fake news titles. This measure ranges from [0,1] indicating none to all overlap. We compute and assign each post a containment value for unigram and bigram matches separately, as we suspect the bigram score to show stronger indication of COVID-19 misinformation than that of single word matches.

3.2 System design

The model uses Sklearn's DecisionTreeClassifier which requires a Sklearn dataset. Due to prior knowledge of Pandas DataFrame, this required a bit of additional work to adapt the dataset to a proper format. After this, the model was trained using 95% of the model and tested using 5% to make sure the results were in the range of what we would expect. This test case is not included in this report as it was purely for a sanity check while testing the code.

Prediction works by taking in arguments and formatting them into the correct Sklearn dataset for the model. This dataset is passed into the model and the result is returned as either 0 (The evidence *does not* suggest the presence of misinformation) or 1 (The evidence *does* suggest the presence of misinformation).

3.3 Data analysis

Analysis of the model relied on collecting new data that the model had never seen before. We chose to run the collection script again, but added a check to make sure that we would not pull any TikTok posts which we had already used for training. We then, again, manually classified these videos as either potentially containing misinformation or not.

Using this information, we ran each item through the prediction model and estimated whether or not the post may contain misinformation. From this, we could perform some simple classification analysis to determine the efficacy of the model. We chose to keep the analysis simple in order not to complicate the model further. Only what is necessary was analyzed. This is further detailed in the Exploration section of this report.

According to Mohajon (2021), “[A] confusion matrix is a tabular way of visualising the performance of your prediction model.” A detailed structure of a confusion matrix can be seen in Fig.1.

In our model, positive class labels refer to misinformation. We’ve modeled the possible outcomes of our test set as TP, TN, FP and FN. These refer to true positive, true negative,false negative and false positive respectively (Mohajon, 2021). In the case that the model correctly predicted the content of the video, we assign it TP. TN, FP, and TN are similar.

We have used these to calculate accuracy, recall, precision, and F1 as defined in lecture. Before we talk about the result, it is better to go through the definition of some items aboved. Mohajon (2021) suggests that “ Recall tells you what fraction of all positive samples were correctly predicted as positive by the classifier. It is also known as True Positive Rate (TPR), Sensitivity, Probability of Detection.” In other words, the formula of calculating the recall is TP divided by the sum of TP and FN. Moreover, Mohajon (2021) argues that “Precision tells you what fraction of predictions as a positive class were actually positive. Shung (2020) argues “F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives)”. The result we got for the model will be shown in the next section, with the argument of which outcome is most important in our model.

4 Evaluation

True Positive(TP)	= 17	Accuracy	= 0.824
False Positive(FP)	= 17	Recall	= 0.515
True Negative(TN)	= 137	Precision	= 0.500
False Negative(FN)	= 16	F1 score	= 0.507

As mentioned above, we used new test data collected in the same manner as our training data to test the model. Our code specifically checked every new post to make sure that it was not included in our training set as it was collected. Once collected, we performed the same method of manual classification on it to get an “actual” value for every post in the test set. We then ran each item through the model to predict whether or not it may contain misinformation. Checking this predicted value against the true value resulted in the above breakdown that can be seen as a confusion matrix in fig.2.

Knowing this information, we were then able to calculate accuracy, recall, precision, and F1 for the test set. Accuracy isn’t of much use to us here as it basically shows what we already knew, which is that as seen in fig.3 and fig.4, approximately 17% of the data contains misinformation.

Here I would argue that precision is the most important score to consider. If we were to use this model to select TikTok posts for more expensive reviews, either by a more advanced process which can actually check the content of the video or by

human moderation, we want to reduce the amount of videos which are falsely selected. A precision of 0.500 isn't spectacular, but it's also promising. This shows that the model is capable of predicting misinformation in nearly half of its predictions. If this were implemented, it could help eliminate half of the misinformation surrounding COVID-19 and vaccination, which would be a noticeable decrease!

Potentially, we could see the model perform better should it be provided a much larger training set. Unfortunately, we were only able to train with 634 posts and test with 187. This was due entirely to the time we had available for this project, and could be remedied by simply allocating more time to data collection.

Furthermore, we plotted the metadata attributes for the data in the training set by their predicted or actual misinformation content. We hoped in doing so to locate any attributes in the dataset that were causing posts to be misclassified. The results of these plots can be seen in figures 5 through 10. These graphs show us that the model is mostly correct in it's predictions, with the most variability seemingly coming from the neutral and positive sentiment analysis scores as seen in figures 5 and 6. Overall, we believe this variability comes from the size of the training set, and would be mitigated should the model have been trained with more information.

5 Discussion

A number of changes were made from our initial proposal. Firstly, the number of posts was reduced from 1000 to 800. This choice was made in order to reduce the number of posts which had to be manually reviewed by our group. While we were reviewing these posts, we found that 160 were irrelevant and that some were not in english. These were removed and resulted in a total of 634 posts in our training set, down from 1000. The 160 posts which proved to be irrelevant were from the hashtag "Conspiracytheory" which contained no information about COVID or vaccination. Posts about vaccination which were not specifically about COVID were left in the data.

We also removed the perc_HashT_Match field from the analysis. This was meant to work similarly to unigram and bigram match, but would include how many of the hashtags a post had which matched a predefined list. This was removed due to time constraints, though we're not sure how effective it would have been.

Finally, our goal for the project changed slightly from the proposal. Initially we set out to gain a better understanding of misinformation on TikTok, which we found would be difficult using the data we had available. Shifting our goal to determine if we could predict misinformation proved to be more easily possible, though we still tried to lightly explore what caused this prediction, as shown in the Exploration section.

If we were to do this project again, the first change would be to increase the number of posts analyzed. Due to the limited time we had this semester, we weren't able to explore that many posts. Without this limitation we could have produced a more accurate model. As well, I think the confusion matrix was a bad method to

demonstrate the efficacy of the model. All it clearly demonstrates is that the model is much better at predicting videos which don't contain misinformation, which contributes so highly to the accuracy of the model. However, this accuracy result is almost meaningless. If the model had predicted every video didn't contain misinformation it would have had nearly the same accuracy. Obviously, the other metrics would all be 0, but accuracy would remain high just due to the nature of the data.

As well, we must acknowledge that any attempt to replicate or repreform this study would lead to different results. Not only are we performing this study at a single point in time, our biases were coded into the model. This model doesn't actually predict for the presence of misinformation, it predicts for the presence of what we believe to be misinformation. Attempting this same process a year from now with a different team may lead to drastically different results, and we can't know for sure how accurate our own analysis is without seeing a second study performed by another group at another time using the same methodology.

6 Conclusion

Ultimately we have shown that it is possible to predict the presence of misinformation in TikTok posts using metadata, though not with a high level of precision. If, like we suggested, this method was used for preliminary filtering, we would waste half of our effort on posts that don't contain misinformation.

However, this doesn't suggest a failure. What are the alternatives to this method? We would argue crowdsourcing via TikToks report system, or a much more complex system which analyzes the content of each video post. These systems both have their own advantages and disadvantages, and we believe a model like ours represents a middle ground. Our method was much faster and cheaper than complex video analysis with machine learning and much more predictable than crowdsourcing.

One of the biggest takeaways from our analysis is the quantity of misinformation on TikTok. Again, as seen in figures 3 and 4, TikTok contains a significant quantity of misinformation under the hashtags we searched. Despite this being against TikTok's community guidelines, some of these posts had been up for weeks or months, and had been seen by many thousands of people. Clearly, we believe methods for predicting the presence of misinformation are required soon lest we see scientific skepticism remain the issue it is today.

7 Appendix

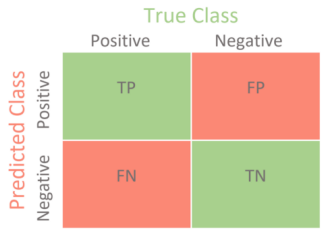

Fig. 1

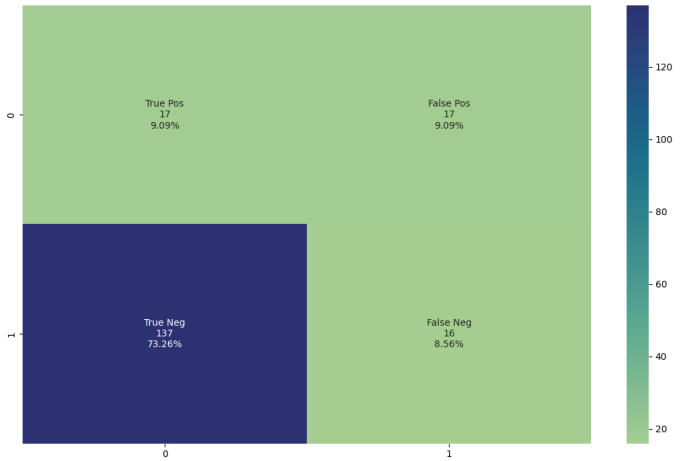

Fig. 2

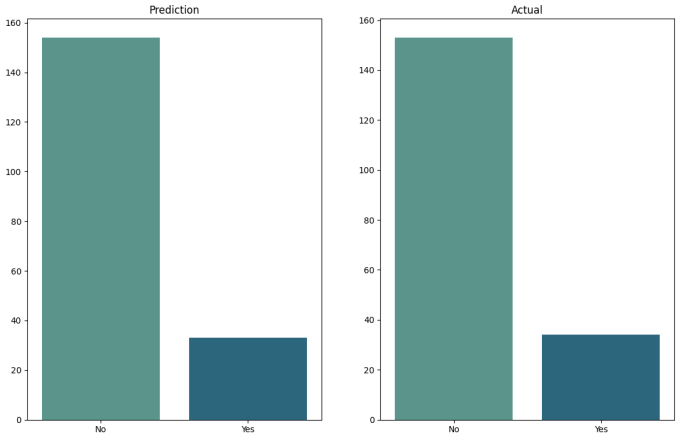

Fig. 3 (18.2% predicted, 17.6% actual)

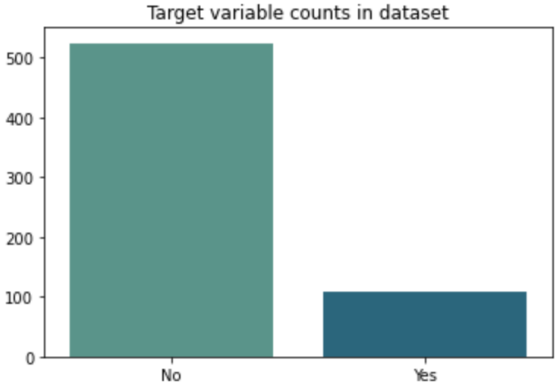

Fig. 4 (17.2% misinformation)

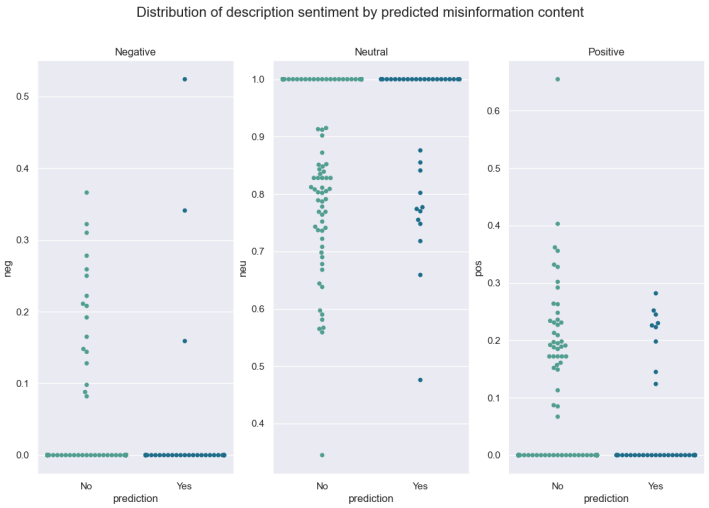

Fig. 5 (Predictions)

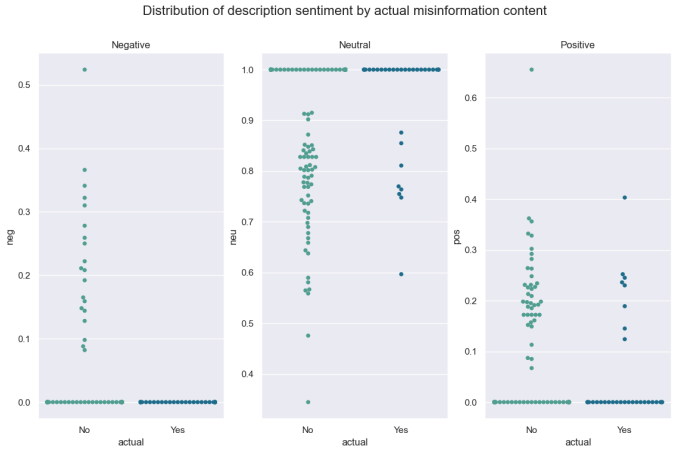

Fig. 6 (Actual content)

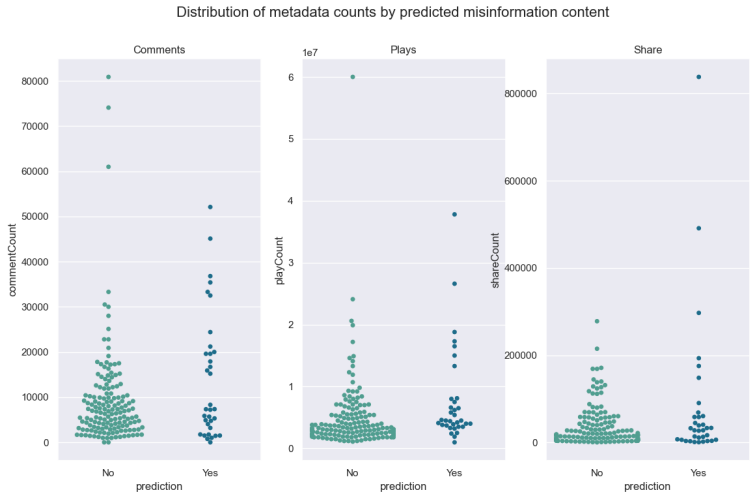

Fig. 7 (Predictions)

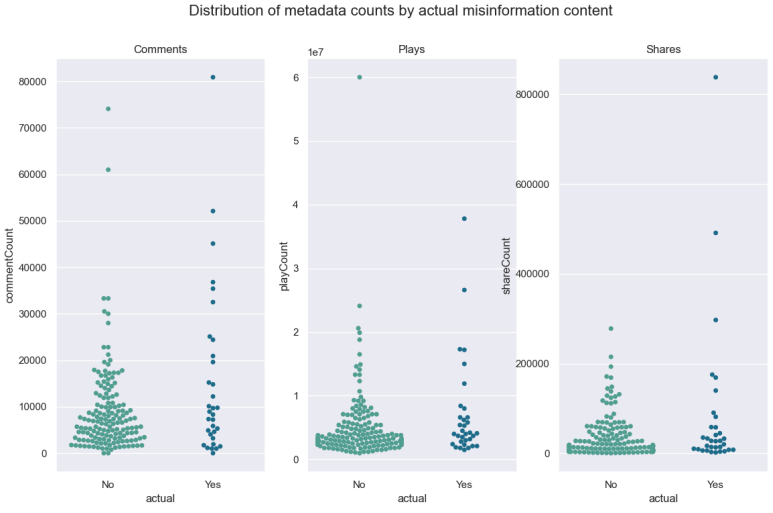

Fig. 8 (Actual content)

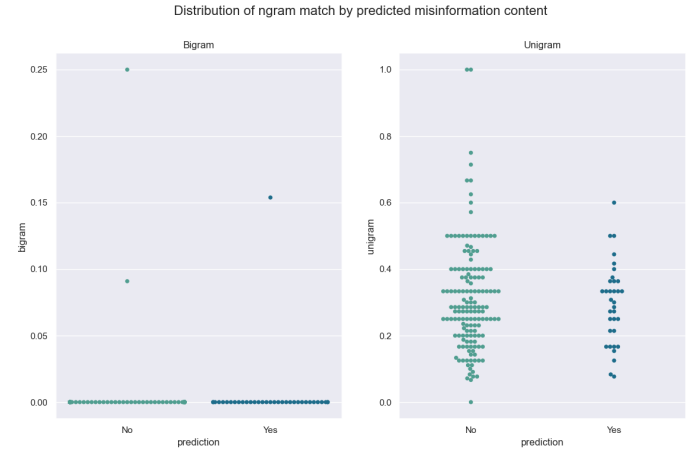

Fig. 9 (Predictions)

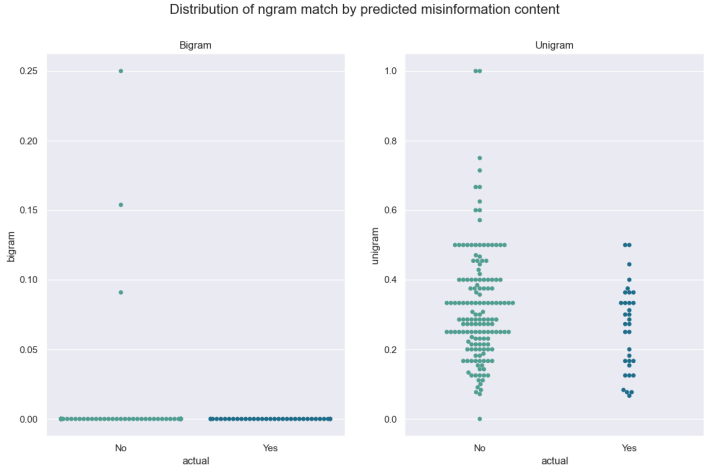

Fig. 10 (Actual content)

8 References

CBC/Radio Canada. (2021). Meet the unvaccinated: Why some Canadians still haven't had the shot CBC News https://www.cbc.ca/news/politics/meet-the-unvaccinated-why-some-canadians-haven-t-had-a-shot-1.6115270.

Julio A. Saenz, Sindhu Reddy Kalathur Gopal, Diksha Shukla, June 12, 2021, "Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset)", IEEE Dataport, doi: https://dx.doi.org/10.21227/b5bt-5244.

Mohajon, J. (2021, July 24). Confusion matrix for your multi-class machine learning model. Towards Data Science. Retrieved December 6, 2021, from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826.

Shung, K. P. (2020, April 10). Accuracy, precision, recall or F1? Medium. Retrieved December 7, 2021, from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.