Misinformation: The Disease Borne out of the COVID-19 Pandemic

Eliana Mugar
emugar@bu.edu

Samantha Rigor
srigor@bu.edu

Abstract

This paper describes the usage of text classification for misinformation models on the Covid Fake News (Patwa et al., 2020) and Covid Fact Checked Polifact (Bui, 2022) datasets from HuggingFace. Using the datasets, we ran two different misinformation models. Model 1 consisted of using Spencer Gable-Cook's "covid19-misinformation-detector" (Gable-Cook, 2022), which is a modified version of the "bert-base-uncased" transformer model (Devlin et al., 2018), and Model 2 consisted of using Roupen Minassian's "TwHIN-BERT-Misinformation-Classifier" model (Minassian, 2023), which is a fine-tuned version of Twitter's "twhin-bert-large" model on an unknown dataset (Zhang et al., 2022). While Model 1 generally performed worse than Model 2, we found that the randomized datasets we used had issues with balanced distribution of real versus fake news as well as the combined grouping of both "false" and "misleading" posts as misinformation. The distribution of the Tweets in the datasets explain the underlying mechanisms behind the models, as well as the metrics drawn from the experiment. From investigating the data from the models, results show that both Model 1 and Model 2 have glaring issues in its labeling of misinformation. We explore the quality of the datasets, noticing the difficulties of annotating and labeling the misinformation to their respective tweets based on the annotators.

1 Introduction

In recent years, younger people have been using social media as their main form of obtaining news. In acknowledging this, though, it is important for us as Internet users to understand how the way news is represented online can affect our views and reactions by fueling polarization or feeding us biased or untruthful content. On the other hand, though, social media provides us with an opportunity for us to express ourselves and connect with others all over the world.

During the height of the COVID-19 pandemic, social media was the only way for people to communicate with one another when they were not able to go outside or travel to see others in person. As a result, sites like Facebook, Twitter, and Instagram became places for people to vent their frustrations with the pandemic, post their opinions on public health regulations, and share other types of content about COVID-19. However, not all of these users may be mindful of the content they post online, and it is even less likely that they are fact-checking the statements they make. As a result, whether or not this misinformation arose on purpose, the pandemic created a new infodemic in which false and misleading posts can thrive. This issue only proliferated once vaccines became available to the public— people voiced their support or disapproval of getting vaccinated against COVID-19.

In response, health officials and social media companies have fought back against the ongoing surge of misleading posts. However, between language generation models and actual human users, it is still difficult to mitigate the dangerous amounts of COVID-19 misinformation that have accumulated over the past three years. It is imperative that language models are trained enough to detect false information. Though human review is the most accurate way of determining whether a post is misleading or not, developing automated tools may be the most immediate way to counteract the growing infodemic. Thus, this paper investigates whether existing text-classification models are able to identify misinformation in social media posts about COVID-19.

2 Task/Dataset Creation

2.1 Task Description

Through this project, we wanted to determine whether or not existing language models are able to correctly identify if social media posts contain

misinformation regarding COVID-19. We chose one model specifically trained on COVID-19 misinformation, and then we chose a model that was trained more generally on misinformation.

2.2 Task Format and Metrics

In this project, we ran two models on two datasets, resulting in four different sets of metrics. In each set of metrics, we will record the accuracy, precision, recall, and F1. The accuracy metric will describe how many cases the model or classifier got correct out of the total number of cases. In other words, the accuracy measures how many times the model or classifier made a correct prediction across the entire dataset. The precision metric will describe how many "misinformation" cases the model or classifier got correct. In other words, precision measures how many "positive" predictions made by the model or classifier were correct. The recall metric will describe how many cases the model or classifier classified as "misinformation" out of the total number of "misinformation" cases. In other words, the recall measures how many of the positive class samples present in the dataset were correctly identified by the model. The F1 metric combines the precision and recall scores.

One set of metrics will be the results of running Spencer Gable-Cook's "covid19-misinformation-detector" model on the "Covid Fake News" dataset. The second set of metrics will be the results of running Spencer Gable-Cook's "covid19-misinformation detector" model on the "Covid Fact Checked Polifact" dataset. The third set of metrics will be the results of running Roupen Minassian's "TwHIN-BERT-Misinformation-Classifier" model on the "Covid Fake News" dataset. The fourth set of metrics will be the results of running Roupen Minassian's "TwHIN-BERT-Misinformation Classifier" model on the "Covid Fact Checked Polifact" dataset.

2.3 Dataset Description and Sources

For this project, we will focus on any Tweets that mention COVID-19 in some way. That being said, a variety of keywords can be used for this dataset, such as "nCov," "Pfizer," "vaccine," "corona," and "Wuhan." This set of Tweets should also span from the onset of the pandemic in December 2019 to the present day. We have chosen to modify two existing datasets from HuggingFace: "nanyy1025/covid_fake_news" and "justinqbui/covid_fact_checked_polifact".

The first dataset, "Covid Fake News," consists of 10,700 social media posts and articles on COVID-19. In the "train" dataset, there are 3 columns and 6,420 rows. The columns represent "id", "tweet", and "label". The "id" column has data entries of 64-bit integers representing the ordering of Tweets in numerical order, the "tweet" column has data entries of type string representing the original Tweet texts regarding COVID-19, and the "label" column has data entries of type string representing whether the Tweet is "real" or "fake" news.

In the "Covid Fact Checked Polifact" dataset, there are 4 columns and 1,186 rows. The "Unnamed: 0" column has data entries of 64-bit integers representing the ordering of Tweets in numerical order. The "claim" column has data entries of type string representing the original Tweet texts regarding COVID-19. The "rating" column has data entries of type string representing how true or false the Tweet is based on the Politifact scale ("false", "pants-fire", "full-flop", "barely-true", "half-true", "mostly-true", "true"). The "adjusted rating" column has data entries of type string representing how true or false the Tweet is based on a simplified version of the Politifact scale ("false", "misleading", "true"). The adjusted rating was created by mapping the raw rating given by Politifact. "True" and "mostly-true" are categorized as "true", "half-true" and "barely-true" are categorized as "misleading", and "false", "pants-fire", and "full-flop" are categorized as "false."

2.4 Task/Dataset Limitations

Since detecting misinformation is currently a developing subject, datasets that were created to study misinformation detection using Tweets can be prone to a temporary lifespan. For some datasets such as ANTi-Vax and CoAID, Tweet IDs were used to label the Tweets instead of the actual text itself. The issue is that Tweets can be deleted at any time, so if the Tweet IDs were recorded in the dataset instead of the original text, then there is no way to retrieve that information once the Tweet is deleted. This leads to the problem of outdated datasets when having to trace back the Tweet IDs.

Also due to misinformation detection being a recent subject, there is a lack of documentation and citations for newly developed models and datasets. This made the process for dataset selections difficult, since it became difficult to code with a lack of documentation on certain datasets. Working with

models and datasets with little detail on documentation led to extra time being spent investigating the models and datasets. Some information that were missing were explanations and meanings of labels and data types, which were essential to writing the code to explore the data. Additionally, since there were rarely any citations to previous work or current research, it was hard to gather previous knowledge to further investigate the area of interest.

Another problem we face is determining the accuracy of the annotations in comparison to the Tweet claims. Because we are not medical professionals or health officials, our knowledge of the pandemic and COVID-19 safety precautions is not much higher than the average person. That being said, we may not have the expertise or resources to be able to verify whether the existing dataset annotations are accurate. On the other hand, it can be difficult to verify dataset annotations due to the aforementioned issue of deleted Tweets. As stated previously, many Twitter datasets on HuggingFace consist of Tweet IDs and labels but do not include the Tweets' content. If we cannot access the original text, there is no way for us to determine why the dataset annotators may have marked a Tweet as misinformation.

Some of the datasets were fairly large, so we had to be wary of processing the data through the resources we had available. Using virtual computer spaces such as Jupyter's Notebook, Boston University's Shared Computing Cluster, and Google's Colaboratory proved useful since we could allocate resources, but due to the nature of collaboration, Google's Colaboratory seemed to make the most sense for this project. The other limitations regarding Google's Colaboratory are the monetary limitations to access more "compute units," faster GPUs, and more memory, which is only available through certain Colaboratory paid plans.

2.5 Related Work

Researchers have been developing datasets to train machine learning models to detect misinformation about COVID-19. Some datasets that were developed include "ANTi-Vax" (Hayawi et al., 2022) and "CoAID" (Cui and Lee, 2020). Current models, such as Spencer Gable-Cook's ""covid19-misinformation-detector" which has been fine-tuned on the ANTi-Vax and CoAID datasets, have fairly high classification accuracies, but still needs to be trained on a wider range of social media posts, which is limited by access to labels for each post (Gable-Cook, 2022). To further this research, some have suggested employing *active learning* to improve the model, where the training process can choose the data from which it learns so the only "chosen" data points need to be labeled. This would allow the model to continue learning even with the limited labels and learning from the most beneficial data points (Settles, 2009).

Outside of the research world, Twitter, among many other social media companies, flagged posts containing COVID-19 misinformation during the pandemic. They began taking down misleading accounts and posts beginning in December 2020, but they stopped tracking COVID-19 misinformation in November 2022 (Safety). In the beginning, this process relied heavy on human review teams, but the company gradually used the human annotations to train automated machine learning models to flag misinformation independently. Once this content was marked as misinformation, Twitter then chose to challenge accounts and/or remove misleading posts. These policies led to 11.72 million accounts challenged, 11,230 accounts suspended, and 97,674 posts removed (Twitter). In spite of these statistics, though, Twitter has not posted the metrics of how accurate their misinformation classification tools were in identifying misleading posts about COVID-19.

3 Model Evaluation

3.1 Description of Experiments

For each of the datasets, we wrote code to randomly choose N=1,000 entries, which created a total of 2,000 Tweets to run with both models. We then standardized the dataset annotation systems to tag Tweets as 0 for factual information or 1 for misinformation. For the "Covid Fake News" dataset, "real" posts were marked 0, and "fake" posts were marked 1. For the "Covid Fact Checked Polifact" dataset, any Tweets that had "adjusted ratings" of "misleading" or "false" were marked 1. This means that only Tweets marked as "true" or "mostly-true" in the "rating" column were marked 0 for factual information.

Model 1 (Spencer Gable-Cook's "covid19-misinformation-detector") has two labels: "LABEL_0" means that no misinformation was detected in the post, while "LABEL_1" means that the post is misinformation. Model 2 (Roupen Minas-

sian's "TwHIN-BERT-Misinformation-Classifier") also has two labels: "factual" and "misinformation". To maintain consistency, we changed these labels such that "LABEL_0" and "factual" were changed to 0, and "LABEL_1" and "misinformation" were both changed to 1.

We also ensured that once we randomly chose the 1,000 entries for both datasets, those subsets were consistent for both models to run.

Once we ran the initial model tests on our datasets, we calculated percentages to compare how the models marked the various degrees of misinformation in the Polifact dataset. We also calculated other percentages to evaluate the models' classification biases and to determine the distribution of real versus fake information in our randomized datasets.

3.2 Results

3.2.1 Model 1: spencer-gable-cook/COVID-19_Misinformation_Detector

After running the model on the "Covid Fake News" dataset, the code reported that the accuracy is 0.574, the precision is 0.627, the recall is 0.574, and the F1 is 0.523.

After running the model on the "Covid Fact Checked Polifact" dataset, the code reported the accuracy is 0.307, the precision is 0.863, the recall is 0.307, and the F1 is 0.362.

Below is a table representing the metrics for this model:

	Fake News	Polifact
Accuracy	0.574	0.307
Precision	0.627	0.863
Recall	0.574	0.307
F1	0.523	0.362

3.2.2 Model 2: roupenminassian/TwHIN-BERT-Misinformation-Classifier

After running the model on the "Covid Fake News" dataset, the code reported that the accuracy is 0.510, the precision is 0.537, and the recall is 0.510, and the F1 is 0.409.

After running the model on the "Covid Fact Checked Polifact" dataset, the code reported the accuracy is 0.794, the precision is 0.838, and the recall is 0.794, and the F1 is 0.814.

Below is a table representing the metrics for this model:

	Fake News	Polifact
Accuracy	0.510	0.794
Precision	0.537	0.838
Recall	0.510	0.794
F1	0.409	0.814

3.3 Error Analysis

After running code to gather the metrics, we ran code on the percentage makeup of each dataset for context of the metrics. In the random subset of 1,000 Tweets in the "Covid Fake News" dataset, 52.1% were real news and 47.9% were fake news. In the random subset of 1,000 tweets in the "Covid Fact Checked Polifact" dataset, 10.9% were real news, 65.8% were fake news, and 23.3% were misleading news.

In the subset of 1,000 tweets for the "Covid Fact Checked Polifact" dataset, the percentages are notably higher for the fake and misleading news, as opposed to the real news. Due to the unbalanced nature of the "Covid Fact Checked Polifact" dataset, the higher percentages of metrics for both models on this dataset may be dependent on whether or not the model is more inclined to mark a tweet as "misinformation."

The Polifact dataset has a "rating" column which uses the Politifact scale ("true", "mostly-true", "half-true", "barely-true", "false", "pants-fire", "full-flop"). To give more context on the metrics regarding the extent of misinformation bias marking, we ran code to determine the percentages of how much each rating is marked as misinformation.

After running Model 1 on the subset of 1,000 random Tweets in the Polifact dataset, 9.5% of "true" Tweets were marked as misinformation, 11.5% of "mostly-true" Tweets were marked as misinformation, 8.4% of "half-true" Tweets were marked as misinformation, 22.6% of "barely-true" Tweets were marked as misinformation, 23.1% of "false" Tweets were marked as misinformation, 29.8% of "pants-fire" Tweets were marked as misinformation, and 0.0% of "full-flop" Tweets were marked as misinformation.

After running Model 2 on the subset of 1,000 random Tweets in the Polifact dataset, 67.4% of "true" Tweets were marked as misinformation, 71.4% of "mostly-true" Tweets were marked as misinformation, 77.1% of "half-true" Tweets were marked as misinformation, 76.0% of "barely-true" Tweets were marked as misinformation, 87.5% of "false"

Tweets were marked as misinformation, 85.6% of "pants-fire" Tweets were marked as misinformation, and 100.0% of "full-flop" Tweets were marked as misinformation.

Below is a table representing the misinformation marking results for each Politifact rating:

	Model 1	Model 2
True	9.5%	67.4%
Mostly-true	11.5%	71.4%
Half-true	8.4%	77.1%
Barely-true	22.6%	76.0%
False	23.1%	87.5%
Pants-fire	29.8%	85.6%
Full-flop	0.0%	100.0%

The high percentage markings towards "misinformation" for Model 2 may be due to a misinformation bias marking, meaning the model may lean towards defaulting a Tweet as "misinformation" rather than "not misinformation". Some other issues in the percentages may be due to the distribution of how many Tweets there are for each rating in the random subset of 1,000 Tweets. For example, 100% of the "full-flop" Tweets were marked as misinformation, but about 0.1% of the random 1,000 Tweets subset are "full-flop" Tweets.

To confirm this hypothesis, we ran code to determine the percentages of each rating that exist in the Polifact dataset. Below is a table representing the distribution percentage of the ratings in the Polifact dataset of 1,000 random Tweets:

	Distribution of Tweets
True	4.3%
Mostly-true	6.6%
Half-true	8.3%
Barely-true	15.0%
False	51.1%
Pants-fire	14.6%
Full-flop	0.1%

3.4 Takeaways

The metrics of both models vary for both datasets. In Model 1, we can see that its accuracy, recall, and F1 scores are low for the Polifact dataset. Since we know that the Polifact dataset is made up of mostly fake and misleading news Tweets, seeing these low percentages in these metrics show that Model 1 poorly identifies whether a Tweet is misinformation or not. In fact, these numbers show it misses misinformation. The low accuracy but high precision scores show the model either consistently marks "misinformation" or "not misinformation," but is not very accurate. Due to the high precision number and low recall number in the Polifact dataset, it is evident the model misses many misinformation cases, but when it does flag the case as "misinformation," it is likely to be misinformation. In other words, the model is unlikely to believe a case is misinformation, but when the model flags the case misinformation, the model most likely drew a conclusion that it is misinformation due to evidence from training. To check this, we ran code on how much Model 1 marks each dataset as "misinformation," regardless of whether or not the prediction is actually correct. Model 1 marked 14.1% of the "Covid Fake News" dataset as misinformation and 21.5% of the Polifact dataset as misinformation. This further supports the conclusions drawn from Model 1's metrics, showing that Model 1 marks misinformation at a low rate. With only 21.5% of the Polifact dataset being marked as misinformation, this displays Model 1's low accuracy, recall, and F1 scores on the Polifact dataset since we know the dataset is about 89.1% misinformation.

Though Model 1 on the "Covid Fake News" dataset has a lower precision score than the Polifact dataset, it seems to have higher accuracy, recall, and F1 scores. This may be due to the more balanced nature of the "Covid Fake News" dataset. The model marked 14.1% of the Fake News dataset as "misinformation" regardless of its correctness out of a 47.9% make-up of fake news, which is a better ratio than the Polifact dataset, which marks 21.5% out of a 89.1% make-up of fake news. This would contribute to the better metrics for Model 1 on the Fake News dataset versus the Polifact dataset.

Model 2's metrics have generally higher scores than Model 1's metrics, though this does not prove Model 2 is better than Model 1. Since we ran code on how much Model 1 marks each dataset as "misinformation," we ran the same test on Model 2. Overall, Model 2 marked 91.5% of the "Covid Fake News" dataset as "misinformation," regardless of its correctness and 83% of the Polifact dataset as "misinformation," regardless of its correctness. Since the Fake News dataset is fairly evenly distributed between real and fake news, the accuracy (51.0%), precision (53.7%), and recall (51.0%) scores confirm the dataset make-up. Since 91.5% of the markings are "misinformation" on the

overall dataset, the model is bound to get around 47.9% of the misinformation markings correct.

Model 2's metrics on the Polifact dataset are very high scores. Since we know that Model 2 leans heavily towards a "misinformation" bias marking, it is unsurprising that the metrics would be high for the Polifact dataset, which has 89.1% fake news in the dataset. Yet, it is surprising that Model 2 marks less misinformation in the Polifact dataset than the Fake News dataset, even though the Polifact dataset objectively has more misinformation.

In the "Covid Fact Checked Polifact" dataset, we initially grouped the "misleading" and "false" categorizations of the Tweets into one "misinformation" category, since intuitively misleading and false news should be categorized as misinformation. We initially took issue with this grouping because of the low scores for Model 1 and the inflated scores for Model 2. However, further investigating this disparity helped us reveal some of the mechanisms behind the models. Grouping the misleading and false categorizations revealed that most of the Polifact dataset was misinformation. This discovery then showed us how Model 2 has a heavy bias towards marking Tweets as misinformation based on the metrics drawn.

4 Conclusion

In light of society's growing dependence on social media, especially during the COVID-19 pandemic, we sought to develop a project that would evaluate the performance of text classification models on misinformation. We modified two existing datasets of Tweets about COVID-19 that were annotated with different systems of misinformation classification. We then ran Spencer Gable-Cook's "covid19-misinformation-detector" model and Roupen Minassian's "TwHIN-BERT-Misinformation-Classifier" model on both datasets to determine how these models classified misinformation of varying degrees.

While the over-labeling of misinformation can lead to accidental censorship, the under-labeling of misinformation can lead to the spread of deception. We ran 2 datasets on both models, proving that the "covid19-misinformation-detector" model tends to under-label misinformation and the "TwHIN-BERT-Misinformation-Classifier" model tends to over-label misinformation. Additionally, while our second dataset used a variety of labels to grade the severity of misinformation in a given post, we did not see a relationship between the degree of misinformation and the rate at which the models flagged it as misinformation. Rather, these rates were more reflective of the widespread distribution of false or misleading tweets in the dataset— only 10.9% of the Tweets in the randomized datasets had an adjusted rating of "true."

Future directions could see how the classification models perform with regard to varying types of misinformation, such as rumors and conspiracies, anti-Asian stereotypes, and unproven medical advice. In addition, it would be interesting to see whether the models would be able to distinguish jokes and sarcasm about COVID-19 from posts that are intentionally spreading false, harmful, or misleading information.

Even though Twitter has stopped flagging accounts and posts for COVID-19 misinformation, the fight against the COVID-19 infodemic is not over. The development of classification models for misinformation has implications far past the pandemic, especially as we continue to rely heavily on social media platforms for our news. If language models come with risks and malicious uses, it is vital that we also develop models to counteract these harms and ensure that people can access truthful and reliable information.

References

Justin Bui. 2022. Covid fact checked polifact.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Spencer Gable-Cook. 2022. Covid-19 misinformation detector: Technical report. *EECS 605 Final Project Report*.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203.

Roupen Minassian. 2023. Twhin bert misinformation classifier.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: Covid-19 fake news dataset.

Twitter Safety. Updates to our work on covid-19 vaccine misinformation.

Burr Settles. 2009. Active learning literature survey. *University of Wisconsin-Madison, Computer Sciences Technical Report 1648*.

Twitter. Covid-19 misinformation.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.