DETECTING AND DETERMINING THE INTENTIONALITY OF COVID-19 MISINFORMATION ON SOCIAL MEDIA

Joshua Peterson
University of North Carolina at Charlotte

ABSTRACT

Throughout the COVID-19 pandemic, social media platforms were and continue to be a prevalent method for spreading misinformation related to the pandemic. The need for misinformation detection models and methods was and is necessary to combat this misinformation spread. The purpose of this research is to develop models that are able to detect misinformation on social media platforms. In addition to detecting misinformation, we also explore methods for determining the intent of that misinformation so the maliciousness and severity of the detected misinformation can be better understood. This natural language processing task will employ both machine learning and deep learning based models and methods as well as exploring the use of zero-shot classification methods to determine intentionality. Through this research, high performing models for detecting COVID-19 misinformation were identified to be a random forest model and a transformer based neural network model. The random forest model obtained a test accuracy of 94.9% and the transformer based neural network obtained a test accuracy of 95%. A recommended method for determining the intent of misinformation is to utilize a zero-shot classification method to assign pre-defined intent labels to social media posts.

INTRODUCTION

Since the onset of the COVID-19 global health emergency, the prevalence and spread of COVID-19 related misinformation on social media has been a consistent issue that has led to the obfuscation of accurate and useful information. The importance of being able to deliver timely and accurate information to a populace during a health emergency was uniquely evident during the early developments of the COVID-19 pandemic and is still an issue to this day. This is why being able to detect misinformation on social media is so vital as it helps identify the sources and spread of that misinformation. Through intent detection, the motive behind that COVID-19 misinformation on social media could start to be better understood. Additionally, this would help prioritize the most egregious spreaders of misinformation as being able to determine intent would help researchers or organizations be able to focus on those spreading COVID-19 related misinformation with purposeful intent.

In the years since 2020, there has been research conducted related to the detection of COVID-19 misinformation on social media. However, the aspect of involving intent detection is somewhat novel within this field. Current research is focused on developing the best methods for detecting misinformation. This research project endeavors to identify how being able to identify the intent of misinformation can inform those COVID-19 misinformation detection methods and models. Therefore, this research project and the subsequent research paper would add a novel perspective within the field of research.

In this paper, a method is developed to detect COVID-19 misinformation on social media and determine the intent of that misinformation. Building upon prior research, current natural language processing (NLP) techniques and using primary knowledge of the subject; these methods and models were developed. Machine learning models were explored to see if these more straightforward and lightweight models would provide a good method for detecting misinformation. Additionally, various types of neural network models were explored to see if these would offer the best method for detecting social media misinformation. To determine intent of the misinformation, zero-shot classification methods were explored to understand if they could label posts with intent using language models.

REVIEW OF LITURATURE

From a review of the relevant current literature regarding the detection of COVID-19 misinformation on social media, it is evident that at this point most high performing models and methods are random forest models or various types of neural network constructions. Checkovid is a COVID-19 misinformation detection system that achieved an f1-score of 94% using a long short-term memory (LSTM) based neural network [3]. Additionally, research from Kolluri, Liu and Murthy developed bidirectional LSTM models that achieved an f1-score of 93% [6]. Mu-Yen Chen, Yi-Wei Lai and Jiunn-Woei Lian conducted research that recommends the use of a bidirectional LSTM model to detect fake news related to COVID-19 [1]. Research by Pramukh Vasist and M.P. Sebastian states that ensemble based methods outperform other methods by achieving accuracies of over 98% and 95% [11]. Therefore, based on current research, it is seen that often deep learning or ensemble models and methods are often the best performing models for COVID-19 misinformation detection.

For intent detection, much of the research revolves around dialogue systems that need to understand the intent of human speech input in dialogue systems. For example, researchers have recommended using a Label-Aware BERT Attention Network to perform zero-shot multi-intent detection to derive the intent of human dialogue [13]. LSTM neural networks have also been recommended to classify human queries with unseen intents [12]. Intent detection typically requires human labeling of training data to train models to detect intent. Therefore, the focus of this research was concerned with how to efficiently label training data using zero-shot classification methods to build those training datasets or provide intent labels for specific examples. Zero-shot classification is an emergent feature of large language models allowing text input to be classified with unseen labels, which could be useful when building these datasets [14].

METHODS

Data Collection

Due to the time available to conduct this research and the need to utilize as much COVID-19 related tweet data as possible, several current open twitter datasets comprised of labeled COVID-19 misinformation tweet data were explored and utilized. CoAid is a healthcare misinformation dataset that contains identified fake news articles and true news articles from websites and social media [2]. The CMU-MisCov19 dataset contains twitter data collected using keywords and hashtags in conjunction with “coronavirus” and “covid” then those tweets were classified using 17 identified categories via human annotation [7]. The COVID-19 Misinformation on Twitter dataset is comprised of human annotated tweets containing COVID-19 information [8].

Once the datasets were identified for use, the labeled tweet datasets had to be rehydrated. Each of these datasets only provided the tweet ID of the labeled tweets; therefore, rehydrating the datasets meant using the tweet IDs to collect the actual text of each of the tweets from Twitter. To do this, a social media scraper was utilized called snscrape. Using snscrape, the tweet IDs were used to scrape the appropriate tweet information from Twitter. Once the datasets were rehydrated, the datasets were joined together into one overall dataset for use. For the purposes of this research, the goal was to only predict if the content of a tweet was misinformative or informative. To this end, the labels of either *fake news* or *true news* were applied to misinformative or informative tweets respectively based on the previous labeling conducted for the open datasets (please see Table 1).

Dataset	**Original Labels**	**Label Conversion**
CoAid[5]	true news article, fake news article	true news = [true news article] fake news = [fake news article]
CMU-MisCov19[6]	calling out or correction, conspiracy, politics, sarcasm or satire, false fact or prevention, true prevention, true public health response, ambiguous or hard to classify, fake cure, irrelevant, news, panic buying, commercial activity or promotion, fake treatment, emergency, false public health response	true news = [calling out or correction, true prevention, true public health response] fake news = [conspiracy, false fact or prevention, fake cure, fake treatment, false public health response]
COVID-19 Misinformation on Twitter Dataset[7]	false, partially false, true, others, unproven	true news = [true] fake news = [false, partially false]

Table 1: Overview of Label Conversion for Assembled COVID-19 Misinformation Dataset

In addition to these rehydrated datasets, additional COVID-19 related tweets and comments were collected from Twitter and Reddit to use when testing the developed models. The additional tweets and comments were used to see how well the developed models and methods could handle new inputs from Twitter as well as text input from other social media platforms such as Reddit. Once this data was collected and collated, there were 9,004 true news tweets and 7,395 fake news tweets.

Data Preprocessing and Analysis

Once the dataset for this research was assembled, it was time to preprocess and analyze the collected data to understand it better. The text of the tweets were first cleaned so any unwanted elements were removed. The following preprocessing methods were applied to the collected tweet data:

- All hyperlinks were removed from tweets. Regular expression methods were utilized to remove the hyperlinks.
- Any characters that are not either alphabetical or numerical were removed. This was done to remove emojis, special characters, etc. Regular expression methods were utilized to remove these characters.
- Hashtags and usernames were removed so only the essential text of each tweet remained. Regular expression methods were utilized to remove these elements.
- All the text was converted to lowercase to further standardize the text. The lower function method available in Python was utilized to convert the text to lowercase.
- Punctuation and stopwords were removed to eliminate irrelevant or unwanted elements from the text of the tweets. The string module was used to remove punctuation via regular expression methods. SpaCy was used to remove stop words.
- Lemmatization was applied to the text of the tweets. SpaCy was used to perform the lemmatization. A SpaCy English language model was initiated and utilized to convert each word within the tweet text to their base form.

After cleaning the tweet text, text analysis was performed to understand the tweet dataset. First, sentiment analysis was conducted utilizing the cleaned tweets. For the sentiment analysis, the TextBlob python package was utilized to generate polarity and subjectivity scores of the tweet data. Polarity is a value between -1 and 1 that indicates how positive a sentiment is with 1 being most positive. Subjectivity is a value between 0 and 1 that indicates how opinionated or subjective an input is with 1 being the most subjective or opinionated. This sentiment information was then used to perform sentiment analysis on the tweet data. From this analysis, it was evident that *fake news* is more negative and less positive overall (please see Figure 1).

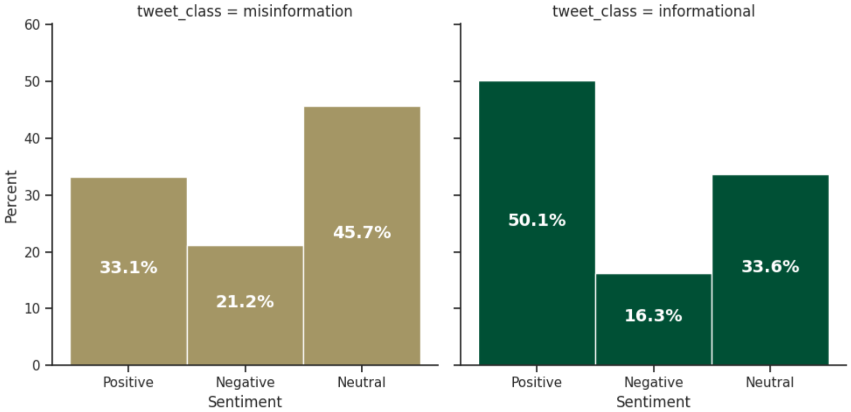

Figure 1: Sentiment distribution of fake news and true news tweets

Once sentiment analysis was performed, word clouds were utilized to better understand what common elements or terms are present in the tweet data (please see Figure 2). These word clouds showed a clear difference between the content in tweets classified as *true news* and those classified as *fake news*. *True news* tweets contained terms such as “covid vaccine”, “vaccine candidate”, “early stage”, “shows promise” and more (please see Figure 2). *Fake news* tweets contained terms such as “vitamin c”, “bill gate”, “artificially created”, “created bioweapon” and more (please see Figure 2). Therefore, a clear differentiation starts to emerge from this initial text analysis.

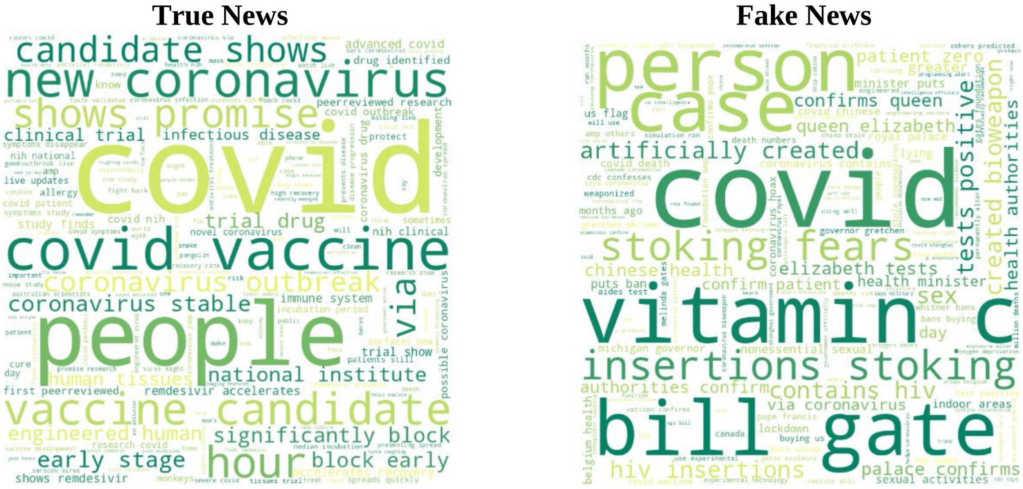

Figure 2: Word clouds of tweets either classified as “true news” or “fake news”

Topic modeling was then conducted. Latent Dirichlet Allocation (LDA) was utilized to identify topics within the tweet data using unigram, bigram and trigram representations of the data. This topic modeling can be used to identify areas of interest within the tweet data. Additionally, it can be used to inform the pre-defined intent labels that will be utilized to determine intent. To find the optimal number of *K* topics, the coherence score was referenced to identify what number of topics produced the most coherent topics. Achieving a coherence score of 0.45 for both the *fake news* and *true news* topic modeling conducted. This topic modeling was conducted on the entire dataset of tweets but was also conducted on just the *fake news* tweets and just the *true news* tweets. For example, one topic in the *fakes news* tweets topic modeling showed top terms of "wearing_masks", "man_made", "like", "health", "april", "americans", "dangerous", "wearing", "world" and "days". One topic in the *true news* tweets topic modeling showed top terms of "vaccine", "study", "scientists", "candidate", "enzyme", "viral", "published", "developed", "candidates" and "mice". The *fake news* topic indicates the misinformation spread about wearing masks as well as the idea the virus was man-made. The *true news* topic shows the information that was being provided regarding the development of a vaccine in response to the virus. This topic modeling helps us develop a better understanding of the *fake news* and *true news* tweets.

It was now time to prepare the tweet data for modeling. To prepare the twitter data for modeling, Term Frequency – Inverse Document Frequency (TF-IDF) vectorization was conducted; which calculates the TF-IDF score of the words within the tweet dataset. This score is then used within the modeling. A max of 7,000 features was allowed during this vectorization, which would allow only the top 7,000 terms/words from the tweet data. After this was done, one hot encoding was utilized to transform the tweet classes of *fake news* and *true news* to 0 and 1 respectively to use as the dependent variable to be predicted.

Modeling for the Detection of COVID-19 Misinformation

Once text preprocessing and analysis was done, a method for detecting COVID-19 misinformation was developed. Before modeling, the data was split into a train set comprising 70% of the tweet data and the remaining 30% was used for the test set. The split was stratified based on the *fake news* and *true news* labels to make sure that the proportion of *fake news* and *true news* classes present in the train and test datasets is equivalent to the overall proportion of those classes in the dataset.

A naive bayes classifier was first tested for the COVID-19 misinformation text classification task. Naïve bayes is a more simple and less processing intensive machine learning (ML) based model. Therefore, it is a good starting point from which to build off of with future tested models. Additionally, the MultinomialNB function does not really have hyperparameters to be tuned. This is why optimization of this model was not explored. For predicting *fake news*, the naïve bayes model had a precision of 89%, recall of 95% and f1-score of 92%. The overall accuracy of the model on the test dataset was 93%.

Model	**Tested Parameters**	**Optimal Parameters**

Support Vector Machine	Gamma: [0.1, 1.0, 10, 100, scale]
	C: [0.1, 1.0, 10, 100]
		Gamma = 1.0
		C = 100

Random Forest	Max Features: [none, auto, sqrt] Criterion: [gini, entropy]	Max Features = auto Criterion = entropy
XGBoost Classifier	Max Depth: [5, 10, 15, 20] Learning Rate: [0.1, 0.2, 0.3, 0.4, 0.5] N Estimators: [100, 500, 1000]	Max Depth = 5 Learning Rate = 0.1 N Estimators = 1000

Table 2: Overview of Parameter Tuning for Models

A support vector machine (SVM) was then tested for this text classification task. The goal of SVM is to find the optimal hyperplane that is able to make the appropriate classifications. For the SVM, the “gamma” and “C” parameters were tuned to assist with finding the most optimal version of the SVM model (please see Table 2). Additionally, k-fold cross validation was performed to identify the performance of these parameters across several different data splits. For predicting *fake news*, the SVM model had a precision of 93%, recall of 93% and f1-score of 93%. The overall accuracy of the model was 93%.

A random forest classifier was then explored. A random forest combines the output of many decision trees to find the optimal classification model. These trees are built independently of each other. The max_depth, max_features and criterion parameters of this model were tuned to find the optimal version of the random forest model (please see Table 2). K-fold cross validation was utilized to ensure the best performance of the model was achieved. For predicting *fake news*, the random forest model had a precision of 94%, recall of 94% and f1-score of 94%. The overall accuracy of the model was 94.94% or approximately 95%.

The last of the machine learning based models to be tested was an XGBoost classifier. XGBoost uses gradient-boosted decision trees to make classifications. In contrast to random forests, XGBoost builds each decision tree on top of the last so the previous tree can be used to inform the next. The max_depth, learning_rate, and n_estimators parameters were tuned to find the optimal version of the model (please see Table 2). K-fold cross validation was performed as well. For predicting *fake news*, the model had a precision of 94%, recall of 94% and f1-score of 94%. The overall accuracy of the model was 94.63% or approximately 95%.

After exploring ML based methods for detecting COVID-19 misinformation, deep learning based methods and models such as using neural networks was explored. The first neural network built was a convolutional neural network. However, additional preprocessing was conducted on each tweet before the neural networks were tested. Each tweet was converted to vector sequences with a max length of 200 words for each sequence. GloVe was utilized to create word embeddings based on the tweet data. These embeddings were then used within the neural networks. Once that embedding layer was constructed, the hidden layers of the convolutional neural network (CNN) were constructed. The CNN utilized the following structure after the input and embedding layers were initiated:

- A SpatialDropout1D layer was used with a rate of 0.2.
- A Conv1D layer with 64 filters, a kernel size of 5 and a *relu* activation.
- A Bidirectional and Long Short-Term Memory layer was used with 64 units, a dropout rate of 0.3 and recurrent dropout rate of 0.3.
- A Dense layer with 512 units and a *relu* activation.
- A Dropout layer with a rate of 0.2.
- A Dense layer with 512 units and a *relu* activation.
- An output Dense layer with 1 unit and a *sigmoid* activation.

This CNN was trained over 20 epochs and started to overfit after about 4 epochs (please see figure 3). The classification report of this CNN showed that the model was able to predict *fake news* with a precision of 94%, recall of 93% and f1-score of 94%. The overall accuracy of the model is 94%.

After the CNN was created and evaluated, a neural network utilizing a transformer was built and evaluated. A class was constructed that included a MultiHeadAttention layer, which is the basis of constructing a transformer as transformers are distinguished by their utilization of self-attention. Self-attention takes the inputs and endeavors to identify those elements of the input that it should “pay attention” to within the input tweet text. Therefore, this model is endeavoring to identify the most important elements for making predictions. This self-attention can be performed through pre-trained models available from platforms such as Hugging Face or via a MultiHeadAttention layer which was utilized in this case. The following is the structure of the transformer based neural network after the embedding layer and transformer block layers were initiated:

- A GlobalAveragePooling1D layer
- A Dropout layer with a rate of 0.1
- A Dense layer with 128 units and a *relu* activation
- A Dropout layer with a rate of 0.1
- An output Dense layer with 1 unit and a *sigmoid* activation

This model was trained over 20 epochs and began to overfit after around 2 epochs (please see Figure 4). For predicting *fake news*, this transformer based neural network has a precision of 93%, recall of 95% and f1-score of 94%. The overall accuracy of this model is 95%.

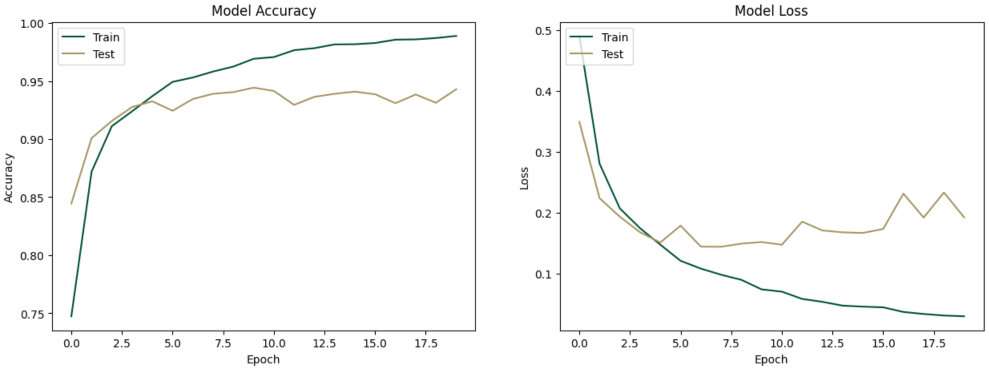

Figure 3: Accuracy and Loss of Convolutional Neural Network

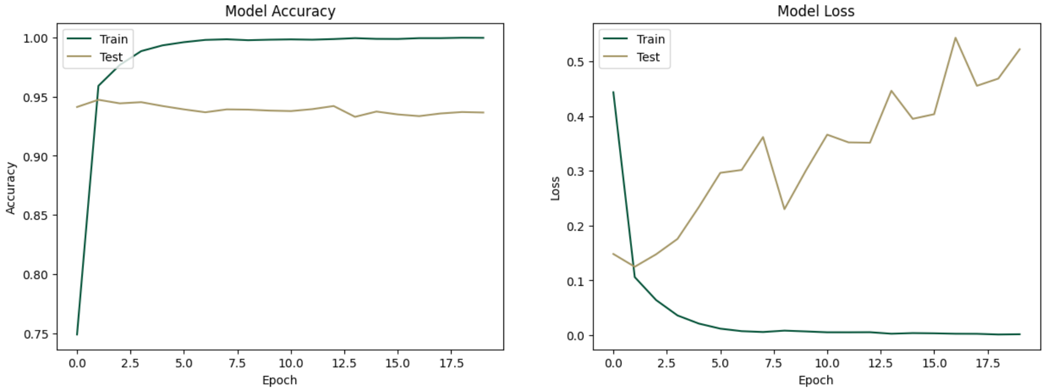

Figure 4: Accuracy and Loss of Transformer Based Neural Network Model

Determining the Intent of COVID-19 Misinformation

Intent determination is a text classification task similar to detecting COVID-19 misinformation. The difficulty lies in being able to label large amounts of tweet data with intent to develop a training dataset to develop models that would be able to detect intent. Furthermore, there were no available open datasets found that had already been labeled with intent classifications. To develop these datasets, it would typically be done with human labeling; however, this can be time consuming and difficult when labeling large amounts of data. This is where zero-shot or few-shot classification is useful. Zero-shot classification is a method that is able to make class predictions without the model having to have seen those classes prior to making the predictions [14]. The concern of using this method would be the accuracy of the unsupervised labeling process. However, human review of the zero-shot generated labels would help determine how well the zero-shot classification method worked.

The zero-shot classifier takes as an input the text to be classified and a set of pre-defined classes to use when making the classifications. The intent labels to be used as classes for the zero-shot classifier were determined to be “question information”, “provide information”, “suggest a political motive”, “promote alternative”, “promote prevention”, and “provide support”. These labels were selected as they would capture many aspects concerning the intent of COVID-19 information or misinformation related tweets. Additionally, the topic modeling conducted helped inform what the intent labels should be for this task. Does the misinformation contain political language; therefore, the intent being to suggest a political motive? Is the tweet promoting something; therefore, it is promoting an alternative or prevention? Does the misinformation tweet make a supportive statement; therefore, intending to provide support? These intent labels would work in conjunction with the detection model to apply intent labels to tweets that are either labeled as *fake news* or *true news*.

The zero-shot method uses an embedding model to create word embeddings of the pre-defined labels and the input text. Once the zero-shot classifier takes those embeddings as inputs, the classifier calculates the similarity between each of the pre-defined labels and the input text to be classified with one of the pre-defined labels. The classifier provides a probability distribution for the pre-defined labels indicating the likelihood that those labels are similar to the input text. A selected pre-trained model is used to score the likelihood that a certain label from the aforementioned pre-defined intent labels applies to a certain tweet. Additionally, a hypothesis template can be used to help the pre-trained language model score the predicted label. In this case, the hypothesis template used was “The intent of this tweet is to”. This template would help the language model determine which intent class to assign to a tweet. The language model used for this intent labeling was the “bart-large-mnli” model. This is a checkpoint from bart-large that was trained on the MultiNLI dataset [4]. The zero-shot classification method was then applied to a subset of 1,500 tweets from the assembled dataset to label each of those 1,500 tweets with one or more of the pre-defined intent labels.

The limitations of using this method are that it can be difficult to develop a comprehensive set of pre-defined intent labels that would account for all types of intent that could be represented within COVID-19 related tweets. There would be out-of-domain (OOD) intents that this method does not account for when labeling tweets. It would be difficult to account for this issue through just expanding the number of labels or increasing the specificity of the pre-defined labels as that can make it more difficult for the zero-shot method to score the similarity of a pre-defined intent label and the input text. Therefore, all this needs to be accounted for when utilizing the zero-shot classification method for intent labeling.

RESULTS

Results of Modeling for Misinformation Detection

Based on the ML models explored, the random forest classifier performed the best overall with an accuracy of 94.94%. The confusion matrix showed that this classifier was able to appropriately classify the majority of COVID-19 tweets as being either misinformative or informative (please see Figure 5).

Of the deep learning methods explored, the transformer based neural network is the best performing model with an accuracy of 95%. The confusion matrix shows that this model was able to classify the majority of COVID-19 tweets in the test dataset appropriately (please see Figure 5).

Both models are able to be used to detect COVID-19 misinformation on social media with a comparable level of accuracy. The random forest model is somewhat more interpretable and lightweight than the deep learning model but there is a higher risk of the model overfitting. The transformer based neural network is more of a black box but it is easier to identify which version of the model is able to generalize best.

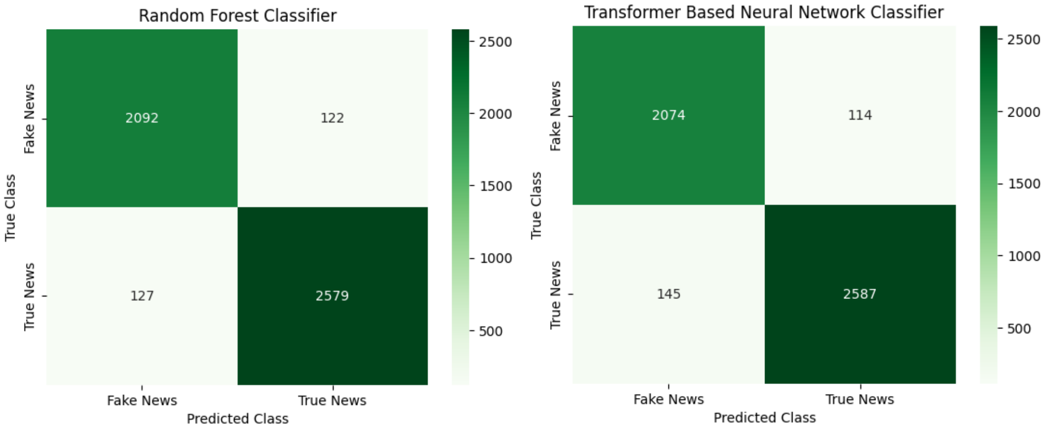

Figure 5: Confusion Matrices of Random Forest Model and Transformer Based Neural Network

Results of Determining the Intent of COVID-19 Misinformation

When using the zero-shot classification method to classify the intent(s) of a single tweet, all intent labels with a score greater than or equal to 75% were applied to an individual tweet. This is because a tweet can have multiple intentions or labels, but it is important to set a threshold for the confidence score so only the most likely intent labels are applied to a given input text. For example, for the tweet, “In today media briefing on #COVID19 I stressed that we are continuing to recommend that all countries make containment their highest priority to find, test, isolate and care for every case and to trace every contact" [5] the zero-shot classification model stated that the “provide information” and “promote prevention” intent labels have a greater than or equal to 75% confidence score in relation to the input text with respective confidence scores of 0.96 and 0.96. This means that the tweet was multi-labeled as “provide information” and “promote prevention”. A sample of 100 tweets with these multiple intent classification label(s) were reviewed manually to see how many of the assigned intent label(s) the reviewer agreed with in the dataset. The reviewer agreed with 85% of the multi-labeled intentions.

Implementation of this involves passing a tweet to the zero-shot classification pipeline, which would then label the tweet with one or more of the intent labels. Further examples of this are tweets such as, “If this is true....we have been duped big time. Any fact checkers out there?” [9]; which was assigned the “question information” label with a confidence score of 0.93. No other intent labels had a confidence score greater than or equal to 75%. Another example is the tweet, “In Uganda, we have contributed more than $15 million, or UGX 56 billion, to the COVID-19 response....we are in this together, #HandInHandWithUganda. Please see video message from our Chargé d’Affaires Chris Krafft on US support to #Uganda for #COVID19” [10], which was assigned the “provide support” and “provide information” intent labels with respective confidence scores of 0.94 and 0.93. These examples show how zero-shot classification can be utilized to start identifying the intent of COVID-19 related social media posts.

To test using this intent labeling method to develop a training dataset with intent labeled tweets, the zero-shot classification method was utilized to label 1,500 of the tweets in the dataset with one of the 6 pre-defined intent labels. In this instance, only the intent label with the highest score was applied to each of the

1,500 tweets to limit the number of classes present within this training dataset. This training dataset could be utilized to train a machine learning or deep learning model to detect the intent of COVID-19 related tweets or other social media posts.

CONCLUSIONS AND FUTURE WORK

In conclusion, this research can be utilized to add to the body of knowledge within the field of COVID-19 misinformation identification on social media. Specifically, the aspect of using zero-shot classification methods to classify COVID-19 related tweets with intent is a novel way of informing COVID-19 misinformation detection. This is because intent can inform the actions taken after detection. In the end, a random forest model or a transformer based neural network model were seen to be high performing methods for detecting COVID-19 misinformation on social media due the approximately 95% test accuracy of these models. A zero-shot classification method is recommended as a possible method to assign intent labels that could be used to understand the intent of that COVID-19 misinformation.

For future work, it would be important to continue to develop and refine the zero-shot classification method for determining the intent of tweets. This could involve reevaluating or refining the pre-defined intent labels to see if the chosen language model is better at making intent classifications with different pre-determined labels or even fine-tuning the language models to use a few-shot method to label the tweets with intent. The use of different language models could be further explored as well to see if a yet unexplored language model is able to make those intent classifications better. Furthermore, the misinformation detection models could continue to be explored to see if their performance can meet the benchmarks of the most state-of-the-art models.

Additionally, it would be advantageous to test performing the zero-shot classification on an input tweet before determining if that tweet is misinformative. The intent of the tweet could inform and provide context to the misinformation prediction. For example, if we are able to determine that the intent of the tweet is sarcastic in nature that would inform if the tweet is misinformative. Therefore, further exploration is needed regarding the interaction between the misinformation detection and intent classification models and methods to see how each classification can impact the other. All of this can lead to the development of machine learning or deep learning models that could detect the intent of COVID-19 related social media posts.

REFERENCES

[1] Chen, M.-Y., Lai, Y.-W., & Lian, J.-W. (2022). Using deep learning models to detect fake news about COVID-19. ACM Transactions on Internet Technology. https://doi.org/10.1145/3533431

[2] Cui, L., & Lee, D. (2020). CoAID: COVID-19 Healthcare Misinformation Dataset. Retrieved from http://arxiv.org/abs/2006.00885

[3] Dadgar S, Ghatee M. Checkovid: A COVID-19 misinformation detection system on Twitter using network and content mining perspectives [Internet]. arXiv [cs.LG]. 2021 [cited 2023 May 7]. Available from: http://arxiv.org/abs/2107.09768

[4] Facebook/bart-large-mnli · hugging face. (n.d.). Retrieved April 28, 2023, from Huggingface.co website: https://huggingface.co/facebook/bart-large-mnli

[5] Ghebreyesus, T. [@DrTedros]. (2020, March 5). *In today's media briefing on #COVID19 I stressed that we are continuing to recommend that all countries make containment their* [Tweet]. Twitter. https://twitter.com/drtedros/status/1236011682230607872

[6] Kolluri N, Liu Y, Murthy D. COVID-19 Misinformation Detection: Machine-Learned Solutions to the Infodemic. JMIR Infodemiology. 2022 Aug 25;2(2):e38756. doi: 10.2196/38756. PMID: 37113446; PMCID: PMC9987189.

[7] Memon, S. A., & Carley, K. M. (2020). Characterizing COVID-19 misinformation communities using a novel Twitter dataset. Retrieved from http://arxiv.org/abs/2008.00791

[8] Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and Media, 22(100104), 100104. doi:10.1016/j.osnem.2020.100104

[9] Sorbo, K. [@ksorbs]. (2020, May 13). *If this is true....we have been duped big time. Any fact checkers out there?* [Tweet]. Twitter.

[10] U.S. Mission Uganda. [@usmissionuganda]. (2020, May 26). *In Uganda, we have contributed more than $15 million, or UGX 56 billion, to the COVID-19 response....we are in this together* [Tweet]. Twitter. https://twitter.com/usmissionuganda/status/1265179251185266688

[11] Vasist PN, Sebastian MP. Tackling the infodemic during a pandemic: A comparative study on algorithms to deal with thematically heterogeneous fake news. International Journal of Information Management Data Insights [Internet]. 2022;2(2):100133. Available from: https://www.sciencedirect.com/science/article/pii/S2667096822000763

[12] Williams, K. (2019). Zero shot intent classification using long-short term memory networks. Interspeech 2019.

[13] Wu, T.-W., Su, R., & Juang, B. (2021). A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

[14] Zero-Shot Classification. (n.d.). Huggingface.Co. Retrieved April 28, 2023, from https://huggingface.co/tasks/zero-shot-classification