Warnken, Julia, and Gokhale, Swapna S. (2022) Classifying the evolving mask debate: A transferable machine learning framework. Journal of Computer-Assisted Linguistic Research 6: 1-18. ISSN: 2530-9455.

Classifying the Evolving Mask Debate: A Transferable Machine Learning Framework

Julia Warnken, Swapna S. Gokhale*
University of Connecticut, Storrs, CT 06269, USA
*Corresponding author: swapna.gokhale@uconn.edu

Received: 13 April 2022 / Accepted: 4 July 2022 / Published: 25 November 2022

Abstract
Anti-maskers represent a community of people that opposes the use of face masks on grounds that they infringe upon personal freedoms. This community has thoroughly exploited the convenience and reach of social media to spread discordant information about the ineffectiveness of and harm caused by masks in order to persuade people to shun their use. Automatic detection and demoting of anti-mask tweets is thus necessary to limit their damage. This is challenging because volumes of anti-mask misinformation continuously evolve with creative arguments that embed emerging knowledge about the virus, changing socio-political landscape, and present policies of public health officers and organizations. Therefore, this paper builds a transferable machine learning framework that can be applied to identify anti-mask tweets from different time periods. What makes this framework transferable is that it identifies and extracts features from the tweet data that remain unchanged and faithfully capture the linguistic and interaction patterns of the anti-mask dialogue. These features are used to train machine learning classifiers that can identify anti-mask tweets with approximately 80% accuracy and F1-score from data collected at four different time periods. Thus, by the virtue of relying upon features that remain invariant, the framework built on data from one time period can be transferred to detect anti-mask tweets from other time periods. This framework may also form the basis of automated classifiers that can efficiently separate falsehoods and misinformation on other prominent issues such as vaccines and climate change from huge volumes of social media data.

Keywords: Anti-mask, Classification, Twitter, Longitudinal Data, Machine Learning

1. Introduction and Motivation

The coronavirus pandemic has caused significant health and economic damage to our society. Simple, commonsense public health measures recommended by the World Health Organization (WHO) and the Centers for Disease Control (CDC) to control the pandemic have ignited fierce

DOI: HTTPS://DOI.ORG/10.4995/JCLR.2022.17493
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

debates and exposed bitter rifts in U.S. society. Of the many interventions, the most basic practice, which causes only minor inconvenience, but holds the promise of saving many lives, is the use of face masks. This seemingly trivial measure, however, has been the subject of intense scrutiny. People who advocate the use of face masks cite several scientific sources that show their value in stemming the propagation of coronavirus (Godoy 2020). Those who oppose the use of face masks, however, question the validity of these studies. They argue that these studies have evaluated the effectiveness of face masks in medical settings, and not during community interactions. They also claim that these inferences are drawn through observations, and not through experimental design, and hence, they are inadequate to justify cause-effect relationships.

This polarizing debate on masks is always resurrected when there is a noticeable rise in the numbers of infections, hospitalizations, and deaths. Anti-maskers, or people who oppose face masks, seek these opportunities to employ a variety of strategies to shun and discourage their use. They organize protests and rallies in public spaces, and openly defy masking orders in stores, airports, and airplanes. Their goal is to draw attention through these actions. Social media platforms also offer them convenient channels to rapidly spread misinformation about how masks are ineffective and can in fact be harmful. Anecdotal evidence shows that social media users believe the health-related information that is shared by family and friends (even though it is misleading and false) more than the official guidance put forth by public health organizations (Perry 2020). This faith in social media poses a real danger that the discordant information that circulates on these platforms may in fact convince many that the measures recommended by public health officials are not effective. It is thus imperative to separate "fact from fiction", that is, to identify and demote such misinformation in order to warn the public.

Given the massive volumes of content shared on social media platforms each day, it is impractical that the content that carries falsehoods and misinformation can be separated manually. Although machine learning may be employed for such classification from a specific data set (Cerbin 2021), it remains a challenge to transfer over or to apply these classifiers to tweets shared outside of that specific period. This is because the percentage of anti-mask content may vary with time and the anti-mask rhetoric incorporates new information about the virus and also presents social and political circumstances. The extent to which the classifiers include this context into their learning, and then rely upon it during classification, is unclear.

In this paper, we first analyze the mask dialogue collected from Twitter at four crucial moments during the pandemic. This analysis reveals that the mask debate has indeed evolved significantly by creatively incorporating new and emerging information. This confirms the need for a transferable machine learning framework, one that can be applied to detect anti-mask tweets from any time period with minimal effort, without the need to build it from ground up every time. The paper then presents such a transferable framework. The heart of the framework lies in identifying and extracting those features from tweet data that are likely to remain unchanged over time, regardless of the background information. These features are used to train ensemble and neural network classifiers. These classifiers are applied to detect anti-mask tweets from data collected at four different time periods, and also from the set that combines all

of them. Our results show that the transferable framework is robust and can separate anti-mask tweets from those leaning pro-mask with an accuracy and a F1-score of at least 0.80 from each individual data set.

These results are promising because they suggest that many linguistic patterns and parameters that determine how users interact and engage with anti-mask and pro-mask tweets remain invariant despite the passage of time, and that the framework has successfully managed to extract these features. These features can thus look beyond the background chatter to train classifiers with a small amount of annotated data in order to automatically detect anti-mask tweets from different time periods. Given that misinformation is a serious threat to public health as warned by the surgeon general (Reuters 2021), these classifiers may also hold the promise of detecting misinformation about other topics that lead to cultural conflicts such as vaccinations, vaccine passports, and climate change.

The rest of the paper is organized as follows: Section 2 presents steps in the preparation of data. Section 3 studies how the mask debate has evolved. Section 4 describes the classification framework. Section 5 discusses the results. Section 6 compares and contrasts related research. Section 7 offers concluding remarks and future directions.

2. Data Collection and Labeling

In this section we discuss the steps in data preparation, namely, data collection and data labeling.

2.1. Data Collection

The mask debate has waxed and waned during the course of the pandemic, coinciding with the rise and fall of the numbers of cases and deaths (Lang, Erickson, and Jing-Schmidt 2021). We identified four such time periods over the course of the first year from July 2020 through August 2021 when the mask debate was resurrected because of a significant turn of events. At these four time periods, tweets were collected using two hashtags *#wearadamnmask* and *#nomask* using the rtweet library (Kearney 2020). The former represented a pro-mask hashtag, and the latter represented an anti-mask hashtag.

2.2. Data Labeling

Anecdotally, it is now known that the use of pro- or anti-mask hashtags in a tweet is not a reliable indicator of whether a tweet as a whole leans pro- or anti-mask. Many anti-mask hashtags are creatively embedded in pro-mask tweets and vice versa (Cerbin 2021), and some tweets contain both pro-mask and anti-mask hashtags in an attempt to reach a wider audience and increase engagement with the tweet. Therefore, we manually annotated each tweet into 'A' for anti-mask, and 'P' for pro-mask.

The corpus collected for each time period was labeled by two annotators. Prior to annotation, duplicate tweets were eliminated. The final corpus for each time period consisted of only those tweets for which the labels from the two annotators matched. The statistical distributions of the tweets into two groups are shown in Table 1.

Date	Pro-Mask	Anti-Mask	Total
July 2020	1019 (50.30%)	1007 (49.70%)	2026
August 2020	1323 (65.63%)	693 (34.37%)	2016
March 2021	1573 (77.83%)	448 (22.17%)	2021
August 2021	1449 (71.20%)	586 (28.79%)	2035
Total	5364 (66.24%)	2734 (33.76%)	8098

TABLE 1. STATISTICAL DISTRIBUTION OF TWEETS

3. EMERGENCE OF MASK DEBATE

In this section, we seek to understand how the mask debate has evolved during the course of the pandemic by analyzing the news and the other literature:

- **July 2020:** Masks were first considered around July 2020, when the country began to emerge from the lockdowns. At this point, masks were promoted as a way to restore a sense of normalcy. This was the earliest time when the mask debate began to circulate on social media platforms. What empowered people to oppose masks is the early stance taken by the U.S. officials that face masks are not effective in preventing the spread of the coronavirus (Cheng, Lam, and Leung 2020; Taylor 2020).

- **August 2020:** In the next one-month period, the spotlight on masks was brighter because of the conflict over reopening of schools and colleges. Masks also became a divisive political issue at the highest level when the then Democratic presidential candidate Joe Biden promised a national mask mandate upon being elected (Morrison 2020). Public health experts also promoted the use of masks as "life-saving", highlighting their importance, by requesting everyone to commit to wearing masks to save a significant number of lives (CDC 2020).

- **March 2021:** The debate entered a new phase during the rollout of vaccines. People continually questioned the effectiveness and the necessity of vaccines especially if fully vaccinated people still needed to wear masks. The CDC countered that masks, especially in indoor spaces with poor ventilation, would still be necessary to prevent the spread of the virus from asymptomatic carriers. Many governors explicitly rescinded the mask mandates or allowed them to expire (Morrison 2020). This dialogue was featured prominently in many school committees and boards. Clashes between Dr. Fauci and Congresspeople erupted, the latter demanding a timeline on when we would be able to drop the masking (Hilder 2021).

- **August 2021:** In August 2021, the Delta variant took hold, straining hospitals and health care systems, especially in the South. The CDC reversed course from its May 2021 guidelines and recommended that even individuals fully vaccinated against Covid-19 wear masks indoors in public if they are in an area of substantial or high transmission. Conservatives cried foul, questioning the effectiveness of vaccines and claiming that the CDC is relying on flimsy data and not science in revising its guidelines (Ernst 2021; Kessler 2021).

Examples of pro-mask and anti-mask tweets from each data set, summarized in Table 2, suggest the significance of these four time periods.

Date	Text	P/A
Jul. 2020	*#NoMasks - as people begin to get really annoyed with them. https://t.co/mfDPyK1XUe*	A
	Difficult to get tested here without severe symptoms, and by then it's too late. Masks are critical to fend off asymptomatic carriers. #WearADamnMask #StayAtHomeSaveLives https://t.co/JbzEUQ7zEB	P
Aug. 2020	*Hurts every time I see a CHILD with a MASK on! IT IS SO WRONG! #NoMasks https://t.co/QQxLAO35S8*	A
	@SenatorLoeffler @CDCgov The surgeon general also says we should all #MaskUp , but our @GOP leaders and @realDonaldTrump refuse to #mandatemasks . Have you been to North Georgia? People are not wearing masks and #COVID19 is spreading, so no school for kid	P
Mar. 2021	*With all the anti-vaxers like me refusing to take a vaccine and wearing a mask as little as possible shouldn't we be dead by now if this were as deadly as the MSM would have you believe?*	A
	@pietepiet Yeah, Canada's been a total mess with the vaccine rollout. Ontario is in shambles over here too. We have all these vaccines stored that we're just...not using? We have government folks espousing anti-mask bullshit.	P
Aug. 2021	*@Garner4Senate Anyone who believes masks aren't part of a comprehensive solution, time to show your work and explain it. For about 18 months I've challenged folks to show the scientific evidence that masks are detrimental to that end. Articula*	A
	I'm living a surreal experience in Florida. Either all have given up and assume they will get and/or die from #COVID19 or they live in an alternate reality. NO MASKS, no social distancing, people living their lives normally. This will never end.	P

TABLE 2. EXAMPLES OF PRO-MASK AND ANTI-MASK TWEETS

We also analyzed the commonly used hashtags from the four data sets to further confirm the above trends. For each hashtag in a given corpus, the ratio of the number of times it appears to the total number of occurrences of all hashtags was computed. We consolidated *#covid19*, *#coronavirus*, *#covid* into a single hashtag as *#covid19*. Similarly, the counts of singular and plural uses of a specific hashtag were merged. These include: *#nomask* and *#nomasks*, *#mask* and *#masks*, *#nomaskmandate* and *#nomaskmandates*, etc. However, *#wearadamnmask* and *#wearamask* are treated separately because of the use of an expletive in the former. The top 15 hashtags, ranked in the order of decreasing percentages, are listed in Table 3. It should be noted that the percentages in each column do not add up to 100%, because the table lists only the most frequently occurring hashtags. Each data set contains many more hashtags that occur just once and these are not included in the table.

Beyond the usual pro-mask and anti-mask hashtags that respectively endorse or reject masks, Table 3 shows the masking dialogue to be embedded in the background chatter of unfolding events. Hashtags in July 2020 are related to re-opening and ending the lockdown, with some expressing skepticism whether Covid is legit as evidenced by the use of the hashtags *#covidhoax* and *#scamdemic*. References to the election in the form of *#maga* and *#trump2020* are also seen in July 2020. As the majority of the states re-opened in August 2020, at least in some limited form, those hashtags disappeared from the dialogue. However, skepticism about Covid continued, albeit in a muted form compared to July 2020. In July 2020 and August 2020, the anti-mask hashtags were somewhat counter balanced by the pro-mask hashtags. However, the greatest percentage of anti-mask hashtags appear in March 2021. For example, in March 2021, the topmost hashtag is *#takeoffyourmask*, perhaps because people expected mask restrictions to ease once the vaccines began to roll out. Overall, the number of anti-mask hashtags, perhaps indicating a greater resistance to masks, is the highest in March 2021. A new symbol of culture wars, vaccine passports, also make their first appearance in the March 2021 corpus. A majority of the hashtags in August 2021, such as *#deltavariant* and *#covidisairborne* are related to the Delta variant, which was the latest variant of the virus. Many such as *#getvaccinated* and *#getvaccinatednow* encourage people to get vaccinated, with the latter emphasizing the urgency to do so. Florida and its governor *(#desantis*, *#deathsantis*) and *#texas* also appear in August 2021 data, and may correspond to the hardline stance of these politicians and states against masks.

In summary, the mask dialogue has evolved during the course of the year by incorporating new knowledge about the virus itself (how it spreads, treatment options, etc.), present social and political landscape (including opening of schools, holidays, sporting events such as the Olympics, Super Bowl, Sturgis rally, election, etc.), and actions and guidance from the public health officials (such as reinstating and relaxing the mask mandate, rolling out of vaccines, etc.). This evolving landscape is reflected in the anti-mask tweets as well. For example, in Table 2, the anti-mask tweet from July 2020 refers to masks as a way of government control, the August 2020 tweet talks about the risk of forcing children to wear masks, in March 2021 interaction between masks and vaccines is mentioned, and in August 2021 the tweet questions the effectiveness of masks based on evidence collected over the 18-month course of the pandemic.

It would be infeasible to build a new classification approach from scratch each time. Thus, our objective is to identify and extract a suite of features that capture patterns that are relevant to anti-mask dialogue, but ignore the background chatter. We apply the framework to identify anti-mask tweets from four different data sets to demonstrate that it can indeed transfer over to different time periods with minimal effort.

4. Classification Framework

Anti-mask tweets can spread myths, lies, and conspiracy theories. These fake theories can provoke people into letting go of a simple public health measure. These tweets must thus be detected and demoted in a timely manner to limit their damage. However, because of the excessive volume, automated detection of such deviant content becomes necessary. Furthermore, the emergence of the mask debate and socio-political details, which penetrates into anti-mask misinformation as seen in Table 2, also calls for transferable classifiers that can

bypass these details. This section presents the steps in such a classification approach to distinguish between pro-mask and anti-mask tweets that can be applied across time periods.

4.1 Text Pre-Processing

The labeled data was converted to UTF-8 encoding, and transformed to lower case. Then, numbers, punctuation and stop words were removed. We observed that only a small percentage of tweets in our data sets contained emojis. Therefore, we eliminated emojis from the tweets, because we found that mapping emojis to their meaning in words increased the size of the feature vectors without any increase in the performance of the classifiers. We eliminated the hashtag (#) symbols, and treated the remainder of the hashtag as a single word. We did not break hashtags such as *takeoffyourmask* into separate words, because doing so automatically was infeasible. We stemmed the remaining words and stripped white space. After this, we identified domain-specific words heuristically, and if these words occurred with similar frequencies in both pro-mask and anti-mask tweets, they were removed as they were unlikely to be informative and contribute meaningfully towards the classification.

4.2 Feature Extraction

The next step was to map the properties of the tweets into features that can be fed as input to machine learning models. We derived the following groups of features.

4.2.1. Content Features

We considered both statistical and semantic features to reflect the content of the tweets. Statistical features included the TF-IDF score, computed for every token in the pre-processed text [36]. TF-IDF is a weighted score that considers the frequency of the term in a single document, and its frequency of occurrence across the entire corpus of documents.

TF-IDF scores cannot capture contextual information, and hence, they cannot tell us how the words are related to each other. Therefore, we used word embeddings to represent semantically related words as closely related vectors. We used two types of word embeddings, Word2Vec and BERT. Word2Vec maps words to features using a distributed numerical representation computed with a two-layer neural network with back propagation (Mikolov et al. 2013). Using the Gensim library, we implemented the CBOW model (Rehurek and Sojka 2010). Each token in the list was represented by a 10-dimensional vector, where min count is 1, with 8 partitions. Word2Vec features were generated for all the words using 25 epochs. BERT, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al. 2018), was used to enhance the contextual information for words in order to further improve classification. BERT features were computed using the base uncased model, which uses masked language modeling and next sentence prediction.

4.2.2. Emotion Features

Emotion features capture the intensity of the hidden feelings that are beyond the meaning

conveyed by words. In face-to-face communication, these feelings can be gathered through facial expressions and body language, and they significantly enhance the perception of the meaning of the tweets. Unfortunately, these clues are not available in written texts such as social media feeds. To substitute for these clues, users employ a variety of techniques such as using emoticons, upper case letters, specific arrangements of punctuations, and quotes. Therefore, we included counts of punctuations, words in uppercase letters, question marks, periods, quotes, and exclamation marks in our framework as surrogates for hidden emotions.

July 2020	August 2020	March 2021	August 2021
#nomask (18.81)	*#maskup* (12.88)	*#takeoffyourmask* (15.97)	*#covid* (13.77)
#wearadammask (18.45)	*#nomask* (9.60)	*#nomask* (8.97)	*#masks* (3.93)
#covid (5.68)	*#wearadamnmask* (8.88)	*#covid* (5.55)	*#deltavariant* (3.93)
#wearamask (2.01)	*#covid* (5.93)	*#antimasker* (4.78)	*#deathsantis* (3.28)
#nomaskmandates (1.03)	*#masks* (1.97)	*#masksdontwork* (4.70)	*#wearamask* (2.95)
#covidhoax (0.92)	*#masksoffamerica* (1.65)	*#covidiot* (1.62)	*#covidisairborne* (1.64)
#scamdemic (0.85)	*#wearamask* (1.64)	*#masks* (1.20)	*#getvaccinated* (1.64)
#dobetter (0.81)	*#nomaskmandates* (0.74)	*#wearamask* (1.1)	*#getvaccinatednow* (1.31)
#endthelockdown (0.65)	*#covidiots* (0.54)	*#antimask* (1.02)	*#texas* (0.98)
#maga (0.63)	*#masksoff* (0.52)	*#justsayno* (0.60)	*#maskmandate* (0.98)
#reopenamerica (0.62)	*#socialdistancing* (0.44)	*#ableg* (0.60)	*#maskup* (0.98)
#maskup (0.60)	*#staysafe* (0.44)	*#novaccinepassports* (0.51)	*#antivaxxers* (0.98)
#trump2020 (0.28)	*#antimaskers* (0.39)	*#florida* (0.51)	*#wearadammask* (0.98)
#mndonothinggov (0.58)	*#scamedemic* (0.36)	*#nomaskinclass* (0.51)	*#covidiots* (0.98)
#masks (0.56)	*#pandemic* (0.36)	*#abpoli* (0.51)	*#vaccinated* (0.66)

TABLE 3. FREQUENT HASHTAGS IN LONGITUDINAL MASK DIALOGUE

Structural organization of the words is another technique which involves using a greater number of certain parts of speech over others. For example, adjectives, adverbs and verbs may be used to convey heightened emotions (Benamara et. al. 2007). POS (part-of-speech) tagging (Loper and Bird 2002) can capture such structural organization. Therefore, each word was classified into one of 35 POS; processing for POS tags was done on raw tweets.

Finally, we also included sentiment and polarity scores computed using VADER (Hutto and Gilbert 2014) and TextBlob (Loria 2018) libraries. These scores were computed for raw tweets just like POS tagging. TextBlob provides the sentiment polarity for each tweet over the range -1 to +1. The compound score computed by Vader ranges from -1 to +1 depending on the net sentiment, computed from the sentiment of each word.

4.2.3. Engagement Features

These features captured how the tweets spread and are received across the platform. They were grouped into five categories:

- The first group included users' explicit actions to boost the visibility, spread and engagement of their tweets. Users may add hashtags, mention other users, add links, and media such as images and videos, and quote other users. We included the counts of hashtags and mentions. We also included binary features indicating whether or not a tweet quotes another tweet, includes media, and links. A tweet from a verified account can also improve its authenticity. Thus, a binary indicator of whether the account is verified is also included.
- The second group included the numbers of likes and retweets, parameters that directly measure the engagement with a tweet.
- The third group captured the network characteristics of the authors, including the numbers of followers and friends.
- The fourth group captured how active the authors are. Numbers of tweets posted and liked by the authors in the lifetime of their accounts, and the number of public lists in which the authors claim membership all belong in this group.
- The fifth group included the properties of the quoted tweets and their authors. These included the numbers of retweets and likes received by quoted tweets, and the numbers of followers and status updates from the authors of quoted tweets. If a tweet does not quote another tweet, then these four features are set to zero.

Of these, the parameters in the third through the fifth groups indirectly influence the spread of the tweets. If an author has a larger network of friends and followers, then their tweet is just that much more likely to be seen by people. Similarly, an author who is active in posting status updates, liking other users' tweets, and is a member of many groups, is more likely to be noticed. Finally, the visibility of the quoted tweets and their authors will further enhance the visibility of the tweets that quote them.

Because the values of these features differed by an order of magnitude, we transformed each

feature using the MinMaxScaler in sklearn (Pedregosa et al. 2011). This function scales and translates each feature to map it to the range of 0 and 1. This transformation is used many times instead of zero mean, unit variance scaling (Pedregosa et al. 2011).

4.3 ML Models

We considered the following two types of machine learning models: ensemble learners and neural networks. Scikit (Buitinck et al. 2013) and Keras implementations of these models were used (Chollet et al. 2015), and the following parameters were chosen:

- **Ensemble Learners:** Random Forests (RF) and Gradient Boosting (GB) are ensemble learners based on the underlying weak learner, which is a decision tree. In Random Forests, the parameters include: number of trees in the forest (100), and the number of features tried at each split in a tree (square-root of the number of total features). We let each decision tree grow fully up to its leaves (Liaw and Matthew 2002). In Gradient Boosting, the parameters include: the number of trees (1600), the fraction of observations selected for each tree (0.55), the maximum depth of each tree (5), the minimum samples in each leaf (1), and the learning rate (0.05) (Freidman 2001).

- **Neural Networks:** Multi-Layer Perceptron (MLP) is a feed-forward Artificial Neural Network (ANN) with input, output, and hidden layers. The number of neurons in the input layer is set to 10 and in the output layer is set to 2. The numbers of neurons in the hidden layers are set to 8 and 5. The activation function used was ReLu (Delashmit and Manry 2005). Long Short-Term Memory (LSTM) is an artificial recurrent neural network architecture used in deep learning (Hochreiter and Schmidhuber 1997). The parameters are: input sequences truncated and padded to 60, vectors of length 100 in the first layer, 100 memory units in the LSTM layer, and a dense output layer with a single neuron and a sigmoid activation function. Binary cross entropy is the loss function, and the efficient ADAM optimization with batch sizes of 64 and 100 epochs is used.

4.4 Performance Metrics

To define performance metrics, we designated the anti-mask and pro-mask classes as positive and negative respectively. Labels predicted by a classifier are compared with those assigned by manual annotators to classify the tweets into four groups. A true positive (TP) occurs when an anti-mask tweet is predicted to be anti-mask, a true negative (TN) occurs when a pro-mask tweet is predicted to be pro-mask, a false positive (FP) occurs when a pro-mask tweet is predicted to be anti-mask, and a false negative (FN) occurs when an anti-mask tweet is predicted to be pro-mask. These four groups can be aggregated into the following four metrics:

- **Accuracy:** Percentage of tweets whose predicted labels match the ground truth or the manually assigned label. It is given by:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

- **Precision:** Considering the universe of tweets predicted to be anti-mask, precision is the

percentage of tweets whose ground truth label is anti-mask (Zafarani, Abbasi, and Liu 2014). The universe of tweets predicted to be anti-mask is the sum of true and false positives. Thus, precision is the ratio of true positives to the sum of true and false positives. It is given by:

$$Precision = TP /(TP + FP)$$

- **Recall:** Considering the universe of tweets that were annotated as anti-mask by the manual coders, recall measures how many of those anti-mask tweets were actually identified correctly by a classifier (Zafarani, Abbasi, and Liu 2014). The universe of tweets annotated as anti-mask is the sum of true positives and false negatives. Thus, recall is the ratio of true positives to the sum of true positives and false negatives. It is given by:

$$Recall = TP /(TP + FN)$$

- **F1-Score:** F1-score is the harmonic mean of precision and recall, and it trades off one metric against the other. It is given by:

$$F1\text{-}score = 2 \; x \; (Precision * Recall)/(Precision + Recall)$$

Precision is optimized when the consequences of false positives are unacceptable, while recall is optimized when the consequences of false negatives are unacceptable. When a pro-mask tweet is labeled as anti-mask (false positive), it may be tagged and demoted based on the policies of the platform. On the other hand, when an anti-mask tweet is labeled as pro-mask (false negative), it will escape the tagging and circulate over the platform. While neither of these have seriously damaging consequences, recall may be a slightly more relevant metric to optimize because it advances the goal of the classification which is to find as many discordant tweets as possible. A tradeoff between precision and recall may also be sought in this case.

5. Results and Discussion

We applied the classification framework to each data set separately, as well as for the combined data set. Each time, the corpus under consideration was split into train/test partitions respectively containing 75%/25% of the tweets. All the models, except for LSTM, were trained and tested on all the features. LSTM was fed pre-processed text directly along with emotion and engagement features. Performance metrics for the individual data sets, as well as for the combined data set are displayed in Table 4.

The Gradient Boosting classifier offers the best accuracy for the individual and the combined data sets. The highest accuracy for individual data sets is around 90%, and for the combined data is around 80%. The lowest accuracy, around 77% is for the August 2021 data. No model is a clear winner with respect to precision and F1-score. Gradient Boosting offers the highest F1-score for the July 2020, August 2020 and the combined data sets. However, for the March 2021 and August 2021 data sets, it is the LSTM classifier that offers the best F1-score. In all the experiments, the F1-score is over 0.80, with the largest value seen in July 2020, and the lowest value in August 2021. Gradient Boosting offers the best precision for July 2020, August 2020 and the LSTM outperforms the others in March 2021 and August 2021. Thus, precision follows the same trend as that of the F1 score.

Overall, the classifiers show the worst performance on the August 2021 data set for all the performance metrics, and the best performance with respect to F1-score, precision and accuracy for the July 2020 data set. Considering the surrounding chatter in these two data sets, this is perhaps not unexpected. The July 2020 data set is nearly balanced between the pro-mask and anti-mask tweets, while the August 2021 data set is mildly imbalanced. Moreover, the commonly used hashtags in August 2021 indicate a significantly greater degree of other concurrent related topics such as politics, vaccination, and new knowledge of the virus. By contrast, most of the common hashtags in July 2020 are focused exclusively on the pandemic, and are probably the most homogeneous. Because of the diversity of the topics, the August 2021 data may be the most challenging, and yet our framework can achieve a F1-score of over 0.80. The F1-score of the Random Forest classifier varies, sometimes it is very close to the Gradient Boosting classifier, and for some data sets it lags. The MLP classifier shows consistently worst performance across all the data sets. Generalizing, it appears that the ensemble classifiers (Random Forest and Gradient Boosting) show consistent and better performance on these data sets compared to neural network architectures.

These performance results suggest that although the anti-mask tweets from different time periods incorporate current knowledge of the virus and other socio-political information, there are sufficient commonalities in their linguistic characteristics, and also in how tweeters engage with them across these time periods. These commonalities can be mapped into features that can allow automated classifiers built for one period to transfer over to other periods with minimal effort and training. These classifiers can thus be employed to detect anti-mask tweets in particular, and misinformation in general from large volumes of social media data.

6. Related Research

During the Covid-19 pandemic, people increasingly turned to social media platforms to voice their opinion about safety and health-related measures. Naturally, these conversations have been mined extensively for public outlook on topics such as vaccines, masks, vaccine passports, physical distancing, etc. In this section, we compare and contrast the work that analyzes the mask dialogue.

Ahmed et. al. (2020) found influential users by an analysis of network centrality measures by building networks of users from mask-related tweets. Patterns in the geographical activity of anti-mask tweets is analyzed to find states with the greatest number of pro- and anti-maskers (Staff 2020). Both Lang et al. (2021) and Al-Ramahi et al. (2020) inferred positive correlation between the number of cases and volume of masking hashtags. Lang et al. (2021) undertook a hierarchical classification scheme, where they manually split pro-mask hashtags into those that encouraged use of masks, asserted their effectiveness, and were a collective benefit to society. They classified anti-mask hashtags into those that rejected masks and those that carried wrong information (Lang, Erickson, and Jing-Schmidt 2021). Al-Ramahi et al. (2020) also undertook hierarchical classification of anti-mask tweets into those mentioning constitutional rights, conspiracy theories, and fake news, pandemic, and data. Pascual-Ferra (2021) found that tweets with anti-mask hashtags were more likely than tweets with pro-mask hashtags to contain toxic

language. He et. al. (2021) confirmed Al-Ramahi's findings by understanding the common attitudes and reasons for resistance towards the wearing of masks. Our analysis of hashtags is similar to Al-Ramahi et al. (2020). However, rather than analyzing the elements related to the mask debate, our motive is to understand how the debate is playing out against the backdrop of the present social and political landscape.

July 2020				
Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
RF	85.40%	0.86	0.85	0.85
GB	89.35%	0.89	0.89	0.89
LSTM	71.99%	0.64	0.92	0.75
MLP	71.20%	0.72	0.71	0.71
August 2020				
Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
RF	75.99%	0.76	0.76	0.74
GB	82.34%	0.82	0.82	0.82
LSTM	71.03%	0.73	0.86	0.78
MLP	58.93%	0.63	0.59	0.60
March 2021				
Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
RF	83.20%	0.84	0.83	0.80
GB	83.40%	0.83	0.83	0.82
LSTM	80.04%	0.84	0.90	0.87
MLP	71.2%	0.71	0.71	0.71
August 2021				
Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
RF	72.01%	0.70	0.72	0.65
GB	77.01%	0.74	0.77	0.74
LSTM	70.92%	0.77	0.86	0.81
MLP	65.82%	0.63	0.66	0.64
Combined				
Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
RF	76.69%	0.79	0.77	0.73
GB	80.44%	0.80	0.80	0.80
LSTM	70.57%	0.69	0.71	0.69
MLP	70.77%	0.69	0.71	0.69

TABLE 4. PERFORMANCE OF ML CLASSIFIERS

Prior work has applied supervised machine learning on tweets related to masks. This includes distinguishing between tweets that are relevant to the wearing of masks, with further filtering

of those that do not express any personal opinions (Ahmed et al. 2020). Cotfas et. al. (2021) classified tweets that expressed personal opinion into for, against, and neutral in terms of the use of masks. In this paper, we build a robust, transferable machine learning framework that can identify anti-mask tweets from data collected over the course of a year, at four critical moments during the pandemic. Our framework achieves better performance than contemporary efforts, despite the socio-political context in which the mask dialogue is framed, by augmenting the linguistic features with additional properties that represent interactions and emotions.

7. Conclusions and Future Research

This paper presents a transferable machine learning framework that can distinguish between anti-mask and pro-mask tweets from data collected at four crucial points in time during the Covid-19 pandemic. The framework successfully extracts invariant features related to content, emotion, and engagement from the tweets that allows it to identify anti-mask tweets with a F1-score of over 0.80. These invariant features can look beyond the background social and political chatter and zero in on those tweets that carry misinformation, and thus can form the basis of automated classifiers that detect misinformation from social media feeds despite the topic or circumstantial information that it may be embedded in.

While this work was focused on detecting anti-mask tweets, our future research will consist of understanding the mask debate through network analysis by identifying the key players at each time period. Studying the mask debate to understand how citizens' attitudes have evolved using topic modeling is also a concern of the future. Applying the framework to detect other types of misinformation on vaccines and climate change is also currently underway.

References

Ahmed, Wasim, Vidal-Alaball, Josep, Segui, Francesc, and Moreno, Pedro. 2020. "A Social Network Analysis of Tweets Related to Masks during the COVID-19 Pandemic." *International Journal of Environmental Research and Public Health*, 17: 8. 10.3390/ijerph17218235.

Al-Ramahi, Mohammad, Noshokaty, Ahmed, El-Gayar, Omar, Nasralah, Tareq, and Wahbeh, Abudllah. 2020. "Public Discourse Against Masks in the COVID-19 Era: Infodemiology Study of Twitter Data" (Preprint). *JMIR Public Health and Surveillance*, 7. 10.2196/26780.

Benamara, Farah, Cesarano, Carmine, Picariello, Antonio, Reforgiato Recupero, Diego, and Subrahmanian, Vs. 2007. "Sentiment analysis: Adjectives and adverbs are better than adjectives alone." *In International Workshop on Web and Social Media*.

Breen, Kerry. 2021. "How Long Do We Need to Wear Masks? Here's What Experts Predict". https://www.today.com/health/how-long-will-we-need-wear-masks-here-s-what-t200771. Accessed: 2022-01-31.

Buitinck, Lars, Louppe, Gilles, Blondel, Mathieu, Pedregosa, Fabian, Mueller, Andreas, Grisel, Olivier, Niculae, Vlad, Prettenhofer, Peter, Gramfort, Alexandere, Grobler, Jacques, Layton, Robert, Vanderplas, Jake, Joly, Arnaud, Holt, Brian, and Varoquaux, Gael. 2013. "API design for machine learning software: Experiences from the scikit-learn project." Proc. of ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 108–122.

CDC. 2020. "CDC Calls on Americans to Wear Masks to Prevent Covid-19 Spread." https://www.cdc.gov/media/releases/2020/p0714-americans-to-wear-masks.html. Accessed: 2021-01-21.

Cerbin, Luca, DeJesus, Jason, Warnken, Julia, and Gokhale, Swapna. 2021. "Unmasking the Mask Debate on Social Media." In 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 677–682. 10.1109/COMPSAC51774.2021.00098.

Cheng, K. K., Lam, Tai, and Leung, Chi. 2020. "Wearing face masks in the community during the COVID-19 pandemic: altruism and solidarity." *The Lancet*. 10.1016/S0140-6736(20)30918-1.

Chollet, Francois et al. 2015. Keras.

Cotfas, Liviu-Adrian, Delcea, Camelia, Gherai, Rare, and Roxin, Ioan. 2021. "Unmasking People's Opinions behind Mask-Wearing during COVID-19 Pandemic: A Twitter Stance Analysis." *Symmetry*, 13(11): 1995. 10.3390/sym13111995.

Delashmit, Walter H., and Manry, Michael T. 2005. "Recent Developments in Multilayer Perceptron Neural Networks." In the 7th Annual Memphis Area Engineering and Science Conference.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *CoRR*, abs/1810.04805.

Ernst, Diana. 2021. "CDC Reverses Course on Masks for Fully Vaccinated in Light of SARS-CoV-2 Variant." https://www.hematologyadvisor.com/general-medicine/cdc-reverse-course-on-masks-for-fullyvaccinated-in-light-of-sars-cov-2-variants/. Accessed: 2022-01-05.

Freidman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of Statistics*, 1189-1232.

Gabe, Nicole, and Hill, Drew. 2021. "New Version of Mask Debate Asks if Fully Vaccinated People Still Need to Wear Them." https://www.winknews.com/2021/03/09/a-new-version-of-the-mask-debate-asks-if-fully-vaccinated-people-still-need-to-wear-them/. Accessed: 2022-01-31.

Godoy, Maria. 2020. "Yes, Wearing Masks Helps, Here's Why." https://www.npr.org/sections/health-shots/2020/06/ 21/880832213/. Accessed: 2021-01-21.

He, Lu, He, Changyang, Reynolds, Tara L., Bai, Qiushi, Huang, Yicong, Li, Chen, Zheng, Kai, and Chen, Yunan. 2021. "Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic." *Journal of the American Medical Informatics Association*, 28: 1564 – 1573.

Hilder, Alex. 2021. "Fauci: Sen. Rand Paul is Dead Wrong in Assuming Masks Aren't Needed After Vaccination". https://www.thedenverchannel.com/news/national/coronavirus/fauci-sen-rand-paul-is-dead-wrong-in-assuming-masks-arent-needed-after-vaccination. Accessed: 2022-01-31.

Hochreiter, Sepp, and Schmidhuber Jurgen. 1997. "Long Short-Term Memory." *Neural Computation*, 9(8): 1735-1780. 10.1162/neco.1997.9.8.1735.

Hutto, C., and Gilbert, Eric. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." Proceedings of the International AAAI Conference on Web and Social Media, 8(1): 216–225.

Kearney, Michael W. 2020. "R: Collecting and Analyzing Twitter Data." https://cran.r-project.org/web/packages/rtweet/rtweet.pdf.

Kessler, Glen. 2021. "The GOP's Attack on the CDC's Mask Reversal and a Study from India." https: //www.washingtonpost.com/politics/2021/08/12/gops-attack-cdcs-mask-reversal-study-india/. Accessed: 2022- 01-05.

Kiely, Eugene. 2021. "Misinformation About Face Masks." https://www.factcheck.org/2021/08/scicheck-misinformation-about-face-masks/. Accessed: 2022-01-05.

Lang, Jun, Erickson, Wesley, and Jing-Schmidt, Zuo. 2021. "#MaskOn! #MaskOff! Digital polarization of mask-wearing in the United States during COVID-19." *PLoS ONE*, 16: e0250817. 10.1371/journal.pone.0250817.

Liaw, Andrew, and Wiener, Matthew. 2002. "Classification and Regression by randomForest." *R News*. 3:18-22.

Loper, Edward, and Bird, Steven. 2002. "NLTK: The Natural Language Toolkit." *CoRR*, cs.CL/0205028.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. 2013. "Distributed Representations of Words and Phrases and their Compositionality." *CoRR*, abs/1310.4546.

Morrison, Sara. 2020. "Biden Wants A National Mask Mandate. Can He Do That?" https://www.vox.com/2020/8/21/21395570/biden-mask-mandate-for-all-national-states. Accessed: 2021-01-21.

Nicholson, Chris. 2019. "A Beginner's Guide to Word2Vec and Neural Word Embeddings." https://pathmind.com/wiki/ word2vec. Accessed: 2022-04-02.

Novitsky, Mikala. 2021. "Governor Ducey's Decision Refuels Mask Debate in Schools." https: //www.kold.com/2021/04/20/governor-duceys-decision-refuels-mask-debate-schools/. Accessed: 2022-01-31.

Pascual-Ferra, Paola, Alperstein, Neil, Barnett, Daniel, and Rimal, Rajeev. 2021. "Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic." *Big Data & Society*, 8.

Paul, Nijhum, and Gokhale, Swapna. 2020. "Analysis and Classification of Vaccine Dialogue in the Coronavirus Era." In 2020 IEEE International Conference on Big Data (Big Data), 3220–3227. 10.1109/BigData50022.2020.9377888.

Pedregosa, Fabian, Varoquaux, Gael, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent et al. 2011. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 12(Oct): 2825–2830.

Perry, Susan. 2020. "Social Media Users are More Likely to Believe False Information About Covid-19 and to Ignore Public Health Advice, Study Suggests". https://www.minnpost.com/second-opinion/2020/07/social-media-users-are-more-likely-to-believe-false-information-about-covid-19-and-to-ignore-public-health-advice-study-suggests/. Accessed: 2021-01-21.

Rehurek, Radim, and Sojka, Petr. 2010. "Software Framework for Topic Modelling with Large Corpora." 45–50. 10.13140/2.1.2393.1847.

Reuters. 2021. "Misinformation is Serious Threat to Public Health Surgeon General Warns". https://www.nbcnews.com/tech/tech-news/misinformation-serious-threat-public-health-surgeon-general-warns-rcna1428. Accessed: 2022-01-05.

Shen, Yanqing. 2020. "Covid-19 Outbreak: Tweet Analysis on Face Masks." https://towardsdatascience.com/covid-19-outbreak-tweet-analysis-on-face-masks-27ef5db199dd. Accessed: 2021-01-21.

Staff, Knau. 2020. "Twitter Analysis Shows Arizona is #1 in Anti-Face Mask Activity." https://www.knau.org/post/ twitter-analysis-shows-arizona-1-anti-face-mask-activity. Accessed: 2021-01-21.

Taylor, Adam. 2020. “How the Split Over Masks Sums Up America’s Chaotic Coronavirus Response.” https://www.washingtonpost.com/world/2020/06/25/face-masks-america-divided/. Accessed: 2021-01-21.

Zafarani, Reza, Abbasi, Mohammad Ali, and Liu, Huan. 2014. “Social media mining: an introduction.” *Cambridge University Press*.