Validation of Sentiment Markers Extracted by Using Machine Learning: Twitter Mining of Covid-19

ABSTRACT

Aim Although sentiment analysis for Covid-19 tweets is becoming popular, no study has mined a sentiment other than polarity. This study aims to extract 'disaster' sentiment by using various machine learning models, statistically validate sentiment markers extracted by deep learning, and discuss the potentials of 'disaster' as valid sentiment marker.

Methods A total of 7,613 disaster tweets from Kaggle site were utilized to train nine machine learning models. A total of 15,619 tweets in English sent from USA were downloaded using streaming API with keywords of Covid, and Omicron, respectively and were classified into disaster/non-disaster categories using the four best performing models: MNB, deep learning, USE and BERT. Principal component analysis, correlation analysis and regression analysis were performed to determine the psychometric properties.

Results Cronbach Alpha for 13 sentiment markers was 0.71. All 4 machine learning markers were loaded in a factor. A higher level of unfavorable emotions (e.g., fear), a lower level of favorable emotions (e.g., joy), and a higher level of negative polarity were found during surging Omicron variants than early onset of Covid-19. The higher frequency of disaster tweets was found during surging Omicron variants than the early onset of Covid-19.

Conclusion Our study revealed that disaster tweets were characterized to be a higher level of unfavorable emotions and negative polarity, and a lower level of favorable emotions. Since disaster as a sentiment marker was evidently reliable and valid, it should be a part of the sentiment analysis in describing the global health issues.

Keywords: machine learning, sentiment analysis, Covid-19, disaster, natural language processing, Twitter mining

1. INTRODUCTION

With over 192 million daily active users reported [1], Twitter provides a unique data source for public health researchers due to the real-time nature of the content and the ease of accessing and searching publicly available information. Specifically, Twitter data have been used in the areas of surveillance, detection of health problems, health promotion, and professional communication. For example, prior studies have successfully utilized Twitter users' commentaries to investigate health promotion topics such as smoking cessation [2], dietary habits [3], fitness [4], social media as an education resource [5], and cancer screening [6]. Twitter has also been used among health agencies, government organizations, hospitals, and medical journals to disseminate information in a timely and up-

to-date manner [7], resulting in a massive increase in tweets related to COVID-19 in a short time-frame.

Throughout the COVID-19 pandemic, Twitter has been one of the most commonly used social media platforms for people worldwide to express their viewpoints and general feelings concerning the pandemic that has hampered their daily lives [8]. A plethora of studies related to the disease have been published since its outbreak using free, accessible, real-time data on Twitter, with topics including, but not limited to: dissemination of misinformation on COVID-19 [9-11], self-reported symptoms [12-15], mental health and emotional wellbeing [16-19], seeking behavior of global health care [20], and geo-temporal analysis for dispersion of the disease [21].

Sentiment analysis, or opinion mining, for COVID-19-related tweets has been an emergingly popular research area related to Twitter mining. Sentiment analysis is defined as the process of identifying and extracting the subjective information from Tweets, including an opinion, a feeling about a particular topic or subject matter, or judgment [22]. The development of data science technology has allowed researchers to determine the emotional tone behind the messages regarding the COVID-19 pandemic in large volumes. Major topics on Twitter sentiment analyses include public opinion on COVID-19 [23-25], government policies such as social distancing and mask policies [26] [27], lockdowns [28], efficacy of hydroxychloroquine [29], and vaccinations [30-32]. In these studies, either rule-based methods [30] [31] or machine learning techniques [23] [26] [32] were used to extract the sentiments. Only a limited number of studies, however, utilized both methods [25] [33]. Moreover, no study has extracted a sentiment other than polarity (positive, neutral, or negative) from Twitter data, such that machine learning models in previous studies were trained with data labeled as polarity (negative or positive). The COVID-19 pandemic can be regarded as a disaster that combines a biological threat with various vulnerabilities, such as the organizational and response capacity of health systems, overcrowding, informality, social work practices, and public transport [34]. So, it is highly recommended that disaster should be considered as another sentiment to be captured by sentiment analysis on COVID-19.

Although Twitter data demonstrates promising prospects in public health research with the increasing trend of studies using twitter data, the validity of Twitter data is of major concern stemming from computer-generated spam, topically irrelevant information, unorthodox abbreviations, and misspelled words. Despite the development of various self-correcting computer algorithms to capture, clean, analyze, and visualize Twitter data, validation remains an important issue. Bovet and colleagues [35] attempted to validate Twitter data on opinions of supporting presidential candidates by comparing relevant Twitter data with national polling estimates. However, no prior study has determined the psychometric properties of the sentiment data extracted by Twitter mining. This study therefore aimed to 1) extract ‘disastrous’ sentiment from tweets regarding COVID-19 by using various machine learning models; 2) statistically validate sentiment markers extracted by machine learning; and 3) discuss the possibilities of ‘disaster’ sentiment as valid sentiment marker in describing global health issues such as the COVID-19 pandemic. Specifically, with respect to statistical validation the study aims to answer the following questions:

(1) Internal consistency (reliability): How well does the sum score of the selected sentiment markers extracted by deep learning and rule-based methods capture the expected score in the entire domain?
(2) Construct validity: How well do the sentiments generated by the deep learning measure the same construct or idea?
(3) Criterion related validity: How well do the sentiment markers extracted by deep learning correlate with those from rule-based methods? and

(4) Discriminant validity: How well do sentiment markers generated by deep learning detect the change in sentiment of COVID-19?

2. MATERIAL AND METHODS

2.1 Overview of the Processes

Sentiment analysis is a useful way to decipher the mood and emotions of the general public that are expressed in social network platforms (e.g., Twitter for this study) [22]. These sentiments are useful for a better understanding of various events and impact they have caused. The flow chart presents the processes of sentiment analysis for tweets on Covid-19. The public opinion extraction on major health issues was performed through machine learning (left side of the flow chart) and lexicon-based methods (right side of the flow chart). For opinion extraction through machine learning, a total of 7,613 labeled tweets were loaded from the Kaggle site and cleaned using the Natural Language Toolkit and regular expression modules. The cleaned data were used to train and validate the models of four machine learning algorithms, three deep learning models, and two transfer learning models with pre-trained encoders. The performance of the trained models was evaluated by four metrics (accuracy, precision, recall and F1), and we selected the best performing models to predict the test data (i.e., to classify the tweets into disaster or not). Lexicon-based sentiment analysis was conducted only for the test data including 154,719 tweets captured using Twitter's streaming application programming interface (API). The VADER (Valence Aware Dictionary for Sentiment Reasoning) function was used to extract polarity (positive, neutral, and negative). National Research Council Canada (NRC) sentiment lexicons were used to extract the Plutchik's emotions (anger, fear, trust, anticipation, disgust, joy, sadness, and surprise) [36] from the tweets. Results from the machine learning and lexicon-based sentiment analysis were combined to conduct further analyses.

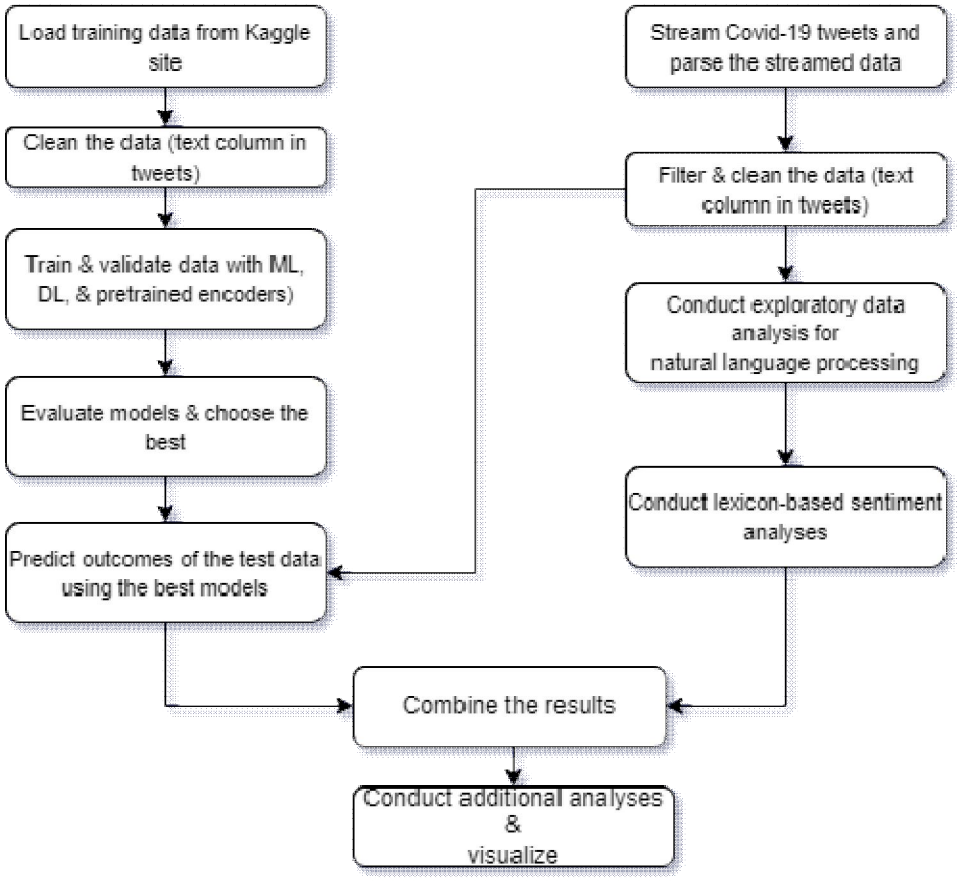

Figure 1. Processes of the Opinion Mining

2.2 Data

2.2.1 Training Data

Unlike prior studies on Twitter sentiment mining where data labeled as positive or negative were used for model training, this study used disaster data for the same purpose. To build models to classify the tweets into disaster and non-disaster categories, we trained the dataset named “Natural Language Processing with Disaster Tweets” [37], where a total of 7,613 tweets were labeled into two groups: 3271 as disaster and 4342 as non-disaster, and included ID, keywords, location, text, and target. Before training, we used the Python module of Natural Language Toolkit (NLTK) and regular expression (RE) for data cleaning to remove emojis, usernames, links, and punctuations in the text of the tweets. Texts excluding “stop” words, which do not convey any meaning (e.g., 'is', 'are', 'and,' etc.), were lemmatized to identify the base form of a word, called Lemma (e.g., 'go' from 'went') based on the dictionary. For this processing, all the tweets were tokenized before applying all the methods for data cleaning. These were then detokenized and converted into NumPy arrays. The data were randomly split into two sets for training (90%) and validation (10%).

2.2.2 Test Data

A total of 66,627 tweets were captured using Twitter's streaming application programming interface (API) method. After filtering to include only tweets written in English sent from the U.S., 15,619 tweets were utilized for the analysis. All test data were cleaned using the same procedure that we used for training data.

COVID Wave I (Early Onset of Covid-19): A total of 1,531 qualified tweets with the keyword "covid-19" were captured between August 20 and August 21, 2020. During this period, a total of 46,022 new cases and 1,100 deaths were reported in a day on average, which were decreasing from the peak on July 17, 2020. Google search returned no salient news related Covid-19 during this period.

COVID Wave II (Surging Omicron Variants): A total of 14,088 qualified tweets with the keywords "covid-19" and "omicron" were captured for 3 days from December 29, 2021. During this period, 344,470 new cases and 1,383 deaths were reported in a day on average. New cases, but not deaths, were rapidly surging due to Omicron variant during this period. Mainstream media released news related to COVID-19 during this time frame, with some example headlines including, " 'The latest wave of COVID-19 cases is "unlike anything we've ever seen,' doctor says" [38]; "U.S. Covid cases rise to pandemic high as delta and omicron circulate at same time" [39]; and "U.S. airlines grapple with Omicron-related disruptions on last day of the year" [40].

2.3 Machine Learning Model Building

For some baseline models which would be utilized as a benchmark for further experiments to build upon, we created a Scikit-Learn Pipeline using the term frequency-inverse document frequency (TF-IDF) formula to convert our words in tweets to numbers, and then model them with the algorithms of Multinomial Naïve Bayes, Ridge Classifier, and Extreme Gradient Boosting and Multi-layer Perceptron classifier. These were chosen by referring to the Scikit-Learn machine learning map [41].

Regarding the model building for deep learning, we started with a single layer dense model. The model took our texts and labels as input, tokenized the texts, created an embedding, found the average of the embedding (using Global Average Pooling), and then passed the average through a fully connected layer with one output unit and a Sigmoid activation function. A long short term-memory (LSTM)-powered recurrent neural network (RNN) was included in our second deep learning model. LSTM is a type of RNN that addresses other RNNs with additional cells, inputs, and outputs. LSTM takes a very similar structure to the dense model but the difference is in adding the LSTM layer between embedding and dense layers. Gated recurrent units (GRUs), which is like a LSTM with a forget gate, was modeled as our last deep model. GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets because of its fewer parameters than LSTM. GRU has a very similar structure to LSTM but the difference is in including GRU layer instead LSTM layer between embedding and dense layers. Our embedding layer yielded a 128-dimensional vector for each word. Each deep learning model used its own trained embeddings because reusing the embedding would involve data leakage between models, leading to an uneven comparison later.

Universal Sentence Encoder (USE) and Bidirectional Encoder Representations from Transformers (BERT) were used to train our data as transfer learning models. There are several substantial benefits to leveraging pre-trained models: simple to incorporate, ability to achieve solid model performance quickly without substantial task-specific architecture modifications, requiring not as much labeled data and versatile use cases from transfer learning, prediction, and feature extraction.

The USE released in 2018 encodes text into high dimensional vectors that can be used for text classification, semantic similarity, clustering and other natural language tasks. The model is trained and optimized for greater-than-word length text, such as sentences, phrases, or short paragraphs. Unlike the embedding we created for the deep learning models which encoded at word level, the USE created a whole sentence-level embedding. Its input is variable length English text and the output is a 512 dimensional vector. The model for the study was built with the USE as our embedding layer which passed through a fully connected layer with a 64-dimensional vector with the Rectified Linear Unit (ReLU) activation function and then output layer with a sigmoid activation function. The trainable parameters were only in our two dense output layers. In other words, the USE weights were kept frozen and used as a feature-extractor.

Our other transfer learning model used BERT, wherein the BERT is pre-trained from unlabeled data extracted from a corpus with 800 million words and English Wikipedia with 2,500 million words, respectively. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Thereby, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. Where each word has a fixed representation under Word2Vec and GloVe regardless of the context within which the word appears, BERT takes into account the context for each occurrence of a given word which allows to produce the context-informed word embeddings capturing other forms of information that result in more accurate feature representations, which in turn results in better model performance. Our model including BERT was pre-trained with a 768-dimension vector and added a single output dense layer. All machine learning models were trained with TensorFlow in Python.

2.4 Model Evaluation and Selection

The four main metrics were used to evaluate the models' performance of classification: accuracy, precision, recall, and F1. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions. Precision is the ratio of correctly predicted observations to total observations. It can answer how many tweets are actually disaster-related among all tweets that were predicted as disaster. Recall is the ratio of correctly predicted positive observations to all observations in actual class. It can answer how many the model classified as disaster from all the tweets that truly were disaster. F1 is the weighted average of precision and recall. A good F1 score (close to 1) means that the model produces low false positives and low false negatives, which allows us to correctly identify real threats and not to be disturbed by false alarms.

Table 1 represents the evaluation metrics by the models that we trained. Overall, the performance of machine learning models was fairly good in classifying the tweets into disaster or non-disaster. While the performance of multi-layer perceptron classifier was relatively lower, the better performance was shown in Multinomial Naïve Bayes, Feed-forward neural network dense model, USE, and BERT. The best performing 4 models were selected to classify the test data: multinomial Naïve Bayes, feed-forward neural network dense model, transfer learning models using USE and BERT.

Table 1. Evaluation of Performance of Machine learning, Deep Learning and Transfer Learning Models

	Machine Learning				Deep Learning			Transfer Learning	
	MNB	XGBoost	RC	MLP	Dense	LSTM	GRU	USE	BERT
Accuracy	0.81	0.78	0.78	0.74	0.81	0.78	0.77	0.81	0.81
Precision	0.82	0.79	0.78	0.74	0.81	0.78	0.77	0.81	0.82
Recall	0.81	0.78	0.78	0.74	0.81	0.78	0.77	0.81	0.81
F1	0.80	0.77	0.78	0.74	0.81	0.77	0.77	0.80	0.81

MNB: Multinomial Naïve Bayes; RC= Ridge Classifier; XGBoost = Extreme Gradient Boosting; MLP= Multi-layer Perceptron classifier; USE=Universal sentence encoder; Dense model= Feed-forward neural network dense model; LSTM= Long short-term memory; GRU= Gated Recurrent Unit; BERT= Bidirectional Encoder Representations from Transformers

2.5 Lexicon-Based Sentiment Analysis

Lexicon-based sentiment analyses were conducted for the entire test dataset including 154,719 tweets captured by using Twitter's streaming application programming interface (API) method. The VADER function was used to extract polarity (positive/negative). VADER is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. The sentiment score of a text can be obtained by summing up the intensity of each word in the text. The vaderSentiment library in python was used to estimate the compound scores, which it is a metric that calculates the sum of all the lexicon ratings that have been normalized between -1 (most extreme negative) and +1 (most extreme positive). The NRC emotion lexicons of the tidytext package in R were used to extract possible emotions such as anger, fear, and trust. The NRC sentiment lexicon categorizes words into mutually non-exclusive sentiments such as anticipation, anger, disgust, joy, fear, surprise, sadness, and trust (the so-called Plutchik wheel of emotions) [36]. The package returns the word counts per emotional category per tweet. Different types of diagrams, such as bar plots, line graphs, and Word Cloud, were implemented to understand patterns between our datasets. For visualization, we used many prebuilt libraries available in Python.

2.6 Statistical Analysis Plan

We conducted additional statistical analyses with the combined results from rule- and machine learning-based opinion mining. All statistical analyses were conducted using SAS 9.4. Cronbach's coefficient alpha was estimated to investigate an internal consistency of the 13 markers extracted from the opinion mining, on a scale from zero to one with a value closer to one being a more reliable measurement instrument and showing higher internal consistency. A principal component analysis was performed to determine if there were the multi-dimensional structures (i.e., possible multi-domains) among markers extracted from opinion mining. Solutions for factors were examined using varimax rotations of the factor loading matrix.

A series of generalized linear models (GLM) were performed to determine any mean difference between health issues in emotions and sentiments extracted from rule-based opinion mining. A series of logistic regression analyses were performed to determine any difference between health issues in opinions (disaster or not) extracted from machine learning-based opinion mining. Pearson correlation coefficients were estimated to determine the correlation of markers between machine learning and rule-based sentiment.

3. RESULTS

3.1 Reliability of Sentiment Markers

The Cronbach's Alpha for the 13 markers that were extracted by rule- and machine learning-based opinion mining was 0.71, which was deemed as having good internal reliability. Principal component analysis was used to identify and compute composite scores for the factors underlying the 13 markers. Initial Eigen values indicated that the first three factors explained 30.8%, 18.5%, and 11.5% of the variance, respectively. The three-factor solution was selected for additional factor analyses because of the 'leveling off' of eigenvalues on the scree plot after three factors. Table 2 presents the factor loadings and communalities based on a principal components analysis with varimax rotation for the 13 markers extracted by rule- and machine learning-based opinion mining. The negative emotions including "anger," "disgust", "fear," "sadness," as well as the VADER score were loaded in the first factor; all four markers generated by machine learning were loaded in the second factor; and positive emotions including "anticipation," "joy," "surprise," and "trust" were loaded in the third factor.

Table 2. Factor Loading of 13 Opinion Markers Extracted by Rule- and Machine Learning-based Methods

	Factor1	Factor2	Factor3
Anger	**0.793**	0.161	0.140
Fear	**0.774**	0.351	0.106
Sadness	**0.774**	0.334	0.048
Disgust	**0.752**	-0.126	0.090
Vader	**-0.578**	-0.258	0.320
DENSE	0.095	**0.823**	0.027
MNB	0.021	**0.792**	-0.050
BERT	0.194	**0.670**	-0.010
USE	0.279	**0.655**	-0.040
Joy	-0.093	-0.125	**0.833**
Anticipation	0.104	-0.062	**0.779**
Trust	0.000	0.075	**0.706**
Surprise	0.418	0.073	**0.594**

3.2 Correlation of Sentiment Markers Between Machine Learning and Rule-based Methods

Table 3 shows the correlation of markers between machine learning rule-based sentiment. The markers generated by machine learning were positively associated with the unfavorable emotions including "anger", "fear", "sadness" and "disgust," and negatively associated with Vader scores and the favorable emotions including "anticipation," "joy," "surprise," and "trust". However, "trust" and "anticipation" had no significant association with USE and BERT.

Table 3. Correlation Coefficients between Machine Learning-based and Rule-based Opinion Markers

	MNB	DENSE	USE	BERT
MNB	1	0.58***	0.21***	0.23***
DENSE	0.58***	1	0.30***	0.28***
USE	0.21***	0.30***	1	0.35***
BERT	0.23***	0.28***	0.35***	1
Anger	-0.04***	0.12***	0.19***	0.05***
Anticipation	-0.02*	0.01	0.09***	0.03**
Disgust	-0.07***	0.01	0.20***	0.14***
Fear	0.06***	0.21***	0.30***	0.17***
Joy	-0.05***	-0.03**	-0.10***	-0.09***
Sadness	0.06***	0.16***	0.22***	0.23***
Surprise	-0.05***	0.11***	0.20***	0.10***
Trust	0.01	0.09***	0.00	-0.06***
Vader	-0.12***	-0.17***	-0.27***	-0.10***

*p-value 0.05-0.01; ** p-value 0.01-0.001; *** p-value <0.001

3.3 Difference in Machine Learning-based Sentiment Between Health Issues

While the dense model classified the highest number of tweets to be disaster (37% for entire dataset, 29% during the early onset of COVID-19, and 38% during surging Omicron variants), only 26% (20% during the early onset of COVID-19, and 27% during surging Omicron variants) were classified as disaster by the BERT model. Table 4 shows the prevalence of tweets predicted to be categorized as disaster by four ML models by each dataset. A higher prevalence of tweets categorized as disaster was found during surging Omicron variants than early onset of COVID-19. This pattern persisted regardless of prediction models with one minor exception: no difference was found when USE was used as the prediction model. There was a non-significant difference in mean Vader scores between the early onset of COVID-19 and surging Omicron variants.

Table 4. Difference in Machine Learning-based Opinion Markers between Early Onset and Surging Omicron Variants

	Early Onset of Covid-19		Surging Omicron Variants		Least Square Mean Difference (95% CI)
	n	%	n	%	
Multinomial Naïve Bayes†	295	19.3	3854	27.4	0.46(0.29~0.62) ***
DENSE model†	446	29.1	5313	37.7	0.39(0.25~0.53) ***
Universal Sentence Encoders †	434	28.4	3996	28.4	0.00(-0.14~0.14)
BERT †	306	20.0	3725	26.4	0.36(0.2~0.52) ***
Vader (mean, SD) ‡	-0.06	0.55	-0.06	0.53	0.01(-0.03~0.04)

† Logistic regression analyses conducted to determine the difference in prevalence of tweets being classified into disaster.
‡ANOVA analysis conducted to determine any mean difference in the level of positivity

4. DISCUSSION AND CONCLUSION

Our study showed that select machine learning models classified 26% - 37% of tweets as a disaster, with more disaster tweets classified by the neural network dense model (37%) than the other three machine learning models (27% by multinomial naïve Bayesian model, 28% by USE, and 26% by BERT). This is comparable with 35% and 40-45% of negative tweets that were classified by support vector machine [26] and CNN-LSTM [42], respectively. Although frequency of disaster tweets varied by the select models, the patterns in different frequencies of disaster tweets among the health issues (i.e., early onset of COVID-19, surging Omicron variants) persisted regardless of the models used for this study. A higher frequency of disaster tweets was found during the surging Omicron variants than during the early onset of COVID-19. The results from our machine learning models may be reliable such that tweets collected during the surging Omicron variants included more unfavorable emotions than those collected during the early stages of COVID-19. However, the frequency of emotions in fear and disgust, and polarity of VADER scores did not differ between the early onset of COVID-19 and Omicron variants, which may be due to people's perceived low severity in Omicron variants juxtaposed with people's perceived surprise in surging cases.

All 13 sentiment markers extracted from rule- and machine learning-based methods were internally consistent and unidimensional: the estimated Cronbach alpha was an acceptable level at 0.71, implying that the markers were only measuring one latent variable or dimension (i.e., sentiment). Further investigation using principal component analysis revealed that all four disaster markers generated by machine learning were loaded in a factor, implying that disaster markers constituted a unique domain other than Plutchik's emotions and polarity. While polarity comprised of a domain with unfavorable emotions like fear, disgust, sadness, and anger, favorable emotions such as surprise, anticipation, joy, and trust made up the other domain. These results suggest that sentiment markers including disaster have a clear construct validity as the markers are measuring their own construct.

Another value of disaster as a valid maker for sentiment analysis is in its responsiveness to change of the COVID-19 pandemic over time. Responsiveness is defined as the ability of a measure to detect changes over time in the construct to be measured [43]. Since the early onset of COVID-19 when a total of 46,022 new cases were reported in a day on average, new cases of COVID-19 rapidly increased by 344,470 in a day during the period of the surging Omicron variant. It is, therefore, expected that the sentiment also changed as new cases did. Our data showed that a higher frequency of disaster tweets was found during the surging Omicron variants than during the early onset of COVID-19 while the level of polarity (VADER scores) did not differ between two periods. That is, disaster markers as well as emotion markers successfully detected the change in sentiment between the early onset of COVID-19 and the surging Omicron variant, implying that the disaster marker, unlike polarity, is well-responsive to the changes in sentiment. Further studies are needed to determine the responsiveness of polarity sentiment extracted from machine learning in comparison to that of rule-based polarity.

Our findings should be interpreted in the context of specific limitations. The streaming API used for the study did not identify social bot accounts where automated accounts are created by industry groups and private companies that aim to influence discussions and promote specific ideas or products [44]. Further research is needed to determine the inter-issue difference after excluding tweets created by social bot accounts.

The second limitation of our study is the exclusion of emojis from the analyses. Global emoji use has reached an all-time high, with more than one in five tweets containing an emoji in comparison to one-in-ten tweets in 2014 [45]. Emojis are defined as "visual representations

of an emotion, idea or symbolism" and may enhance the exchange of emotional information by providing additional social cues beyond those found in a text message used to augment the meaning of a message as a whole [46]. It is not new to use emojis to convey ideas about health or disease. During the pandemic, Emojis were used in quite expressive ways using various symbols of, generally, smiley faces to express their emotions ranging from confusion to sadness, crying, alarm, frustration, and anxiety [47]. Emojis as self-reported emotions should thus be considered an important part of sentiment analysis. It is, therefore, important for future studies to understand users' emotions when they use Emojis.

Another limitation is that our study may be subject to issues related to representativeness of study samples. We included only tweets in English sent from the U.S. While language is clearly indicated in tweets, location is not. Only 0.85% of all users [48] and 3.1% of all original tweets [49] were estimated to have an exact location (e.g., country code). We filtered out the tweets based on information in the location field of tweets, which was not clean. Of a total of 66,627 tweets written in English, 34% included 'Not Applicable' or a missing value in this field in our data. Some tweets indicated only a city name, which made it difficult to identify the country where the city belonged to (e.g., Birmingham in UK or USA?). Our filtering procedure identified only 23% tweets sent from the U.S. Thus, tweets sent from the U.S. may be under-filtered because of massive 'Not Applicable' location statuses, duplication of city name across the country, etc. It is also important to note that the keyword method for retrieving data that the study adopted may have excluded tweets from users who tweeted about the issues without using our targeted keywords, such that the inclusion of other search terms may have modified the results. Another issue related to representativeness is the unbalanced time span for data collection across the two health issues. We collected early onset COVID-19 Twitter data for 2 days but surging Omicron Twitter data for 5 days. It is unknown that the time spent for data collection is appropriate to capture the representative samples per health issue.

Despite these limitations, our study revealed that a higher frequency of disaster tweets was found during the surging Omicron variants than during the early onset of COVID-19. This pattern persisted regardless of the machine learning models used for this study, although the models classified a varied frequency of disaster tweets. Disaster tweets were characterized to be a high level of unfavorable emotions and negativity and a lower level of favorable emotions. To our best knowledge, this is the first study to include disaster in sentiment analysis of COVID-19. It was reliable and valid in describing the global health issues. Not only was it internally consistent with polarity and emotions (internal reliability), but it also measured a unique sentiment other than emotions and polarity (construct and criterion validity). It also achieved better responsiveness to the changes of sentiment in COVID-19 than polarity (discriminant validity). Therefore, it is highly recommended that disaster should be a part of sentiment analyses in describing global health issues.

REFERENCES

1. Ahmed A. Twitter's Daily Active Users Number Reached to 192 Million in the Fourth Quarter of 2020. Digital Information World. 2021. Accessed 7 August 2021. Available: https://www.digitalinformationworld.com/2021/02/twitters-daily-active-users-number.html

2. Prochaska JJ, Pechmann C, Kim R, Leonhardt JM. Twitter=quitter? An analysis of Twitter quit smoking social networks. Tob Control. 2012;21(4):447-449. doi:10.1136/tc.2010.042507

3. Hingle M, Yoon D, Fowler J, Kobourov S, Schneider ML, Falk D, Burd R. Collection and visualization of dietary behavior and reasons for eating using Twitter. J Med Internet Res. 2013 Jun 24;15(6):e125. doi: 10.2196/jmir.2613. PMID: 23796439; PMCID: PMC3713881.

4. Vickey T, Breslin JG. Online Influence and Sentiment of Fitness Tweets: Analysis of Two Million Fitness Tweets. JMIR Public Health Surveill. 2017;3(4):e82.. doi:10.2196/publichealth.8507

5. Pizzuti AG, Patel KH, McCreary EK, et al. Healthcare practitioners' views of social media as an educational resource. PLoS One. 2020;15(2):e0228372. doi:10.1371/journal.pone.0228372

6. Lyles CR, Godbehere A, Le G, El Ghaoui L, Sarkar U. Applying Sparse Machine Learning Methods to Twitter: Analysis of the 2012 Change in Pap Smear Guidelines. A Sequential Mixed-Methods Study. JMIR Public Health Surveill. 2016;2(1):e21. doi:10.2196/publichealth.5308

7. Rosenberg H, Syed S, Rezaie S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. CJEM. 2020;22(4):418-421. doi:10.1017/cem.2020.361

8. Ilyas SZ, Hassan A, Hussain SM, et al. COVID-19 persuaded lockdown impact on local environmental restoration in Pakistan. Environ Monit Assess. 2022;194(4):272. doi:10.1007/s10661-022-09916-7

9. Rosenberg H, Syed S, Rezaie S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. CJEM 2020 Jul;22(4):418-421. doi: 10.1017/cem.2020.361.

10. Krittanawong C, Narasimhan B, Virk HUH, Narasimhan H, Hahn J, Wang Z, Tang WHW. Misinformation Dissemination in Twitter in the COVID-19 Era. Am J Med. 2020 Aug 14:S0002-9343(20)30686-0. doi: 10.1016/j.amjmed.2020.07.012. Online ahead of print.

11. Kawchuk G, Hartvigsen J, Harsted S, Nim CG, Nyirö L. Misinformation about spinal manipulation and boosting immunity: an analysis of Twitter activity during the COVID-19 crisis. Chiropr Man Therap. 2020 Jun 9;28(1):34. doi: 10.1186/s12998-020-00319-4.

12. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, Liang B, Cai M, Cuomo R. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health Surveill. 2020 Jun 8;6(2):e19509. doi: 10.2196/19509.

13. Guo JW, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. Public Health Nurs. 2020 Nov;37(6):934-940. doi: 10.1111/phn.12809.

14. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang YC. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. Am Med Inform Assoc. 2020 Aug 1;27(8):1310-1315. doi: 10.1093/jamia/ocaa116.

15. Panuganti BA, Jafari A, MacDonald B, DeConde AS. Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google

and Twitter. Otolaryngol Head Neck Surg. 2020 Sep;163(3):491-497. doi: 10.1177/0194599820932128. Epub 2020 Jun 2.

16. Guntuku SC, Sherman G, Stokes DC, Agarwal AK, Seltzer E, Merchant RM, Ungar LH. Tracking Mental Health and Symptom Mentions on Twitter During COVID-19. J Gen Intern Med. 2020 Sep;35(9):2798-2800. doi: 10.1007/s11606-020-05988-8.

17. Doogan C, Buntine W, Linger H, Brunt S. Public Perceptions and Attitudes Toward COVID-19 Nonpharmaceutical Interventions Across Six Countries: A Topic Modeling Analysis of Twitter Data. J Med Internet Res. 2020 Sep 3;22(9):e21419. doi: 10.2196/21419.

18. Chehal D, Gupta P, Gulati P. COVID-19 pandemic lockdown: An emotional health perspective of Indians on Twitter. Int J Soc Psychiatry. 2020 Jul 7:20764020940741. doi: 10.1177/0020764020940741.

19. Singh P, Singh S, Sohal M, Dwivedi YK, Kahlon KS, Sawhney RS. Psychological fear and anxiety caused by COVID-19: Insights from Twitter analytics. Asian J Psychiatr. 2020 Jul 11;54:102280. doi: 10.1016/j.ajp.2020.102280.

20. Amat-Santos IJ, Baladrón C, San Román JA. Twitter and the pursuit of global health-care during COVID-19 pandemic. Med Clin (Engl Ed). 2020 Sep 25;155(6):268-269. doi: 10.1016/j.medcle.2020.06.006.

21. Klein A, Magge A, O'Connor K, Cai H, Weissenbacher D, Gonzalez-Hernandez G. A Chronological and Geographical Analysis of Personal Reports of COVID-19 on Twitter. medRxiv. 2020 Apr 22:2020.04.19.20069948. doi: 10.1101/2020.04.19.20069948

22. Carchiolo V, Longheu A, Malgeri M. Using twitter data and sentiment analysis to study diseases dynamics. In: Inter-national conference on information technology in bio-and medical informatics 2015 Sep 3 (pp 16–24). Springer, Cham

23. Garcia K, Berton L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Appl Soft Comput. 2021;101:107057. doi:10.1016/j.asoc.2020.107057

24. Valle-Cruz D, Fernandez-Cortez V, López-Chau A, Sandoval-Almazán R. Does Twitter Affect Stock Market Decisions? Fi-nancial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods. Cognit Comput. 2022;14(1):372-387. doi: 10.1007/s12559-021-09819-8. Epub 2021 Jan 23. PMID: 33520006; PMCID: PMC7825382

25. Crocamo C, Viviani M, Famiglini L, Bartoli F, Pasi G, Carrà G. Surveilling COVID-19 Emotional Contagion on Twitter by Sentiment Analysis. Eur Psychiatry. 2021 Feb 3;64(1):e17. doi: 10.1192/j.eurpsy.2021.3. PMID: 33531097; PMCID: PMC7943954.

26. Shofiya C, Abidi S. Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. Int J Envi-ron Res Public Health. 2021;18(11):5993. doi:10.3390/ijerph18115993

27. Sanders AC, White RC, Severson LS, Ma R, McQueen R, Alcântara Paulo HC, Zhang Y, Erickson JS, Bennett KP. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. AMIA Jt Summits Transl Sci Proc. 2021 May 17;2021:555-564. PMID: 34457171; PMCID: PMC8378598.

28. Su Y, Xue J, Liu X, Wu P, Chen J, Chen C, Liu T, Gong W, Zhu T. Examining the Impact of COVID-19 Lockdown in Wu-han and Lombardy: A Psycholinguistic Analysis on Weibo and Twitter. Int J Environ Res Public Health. 2020 Jun 24;17(12):4552. doi: 10.3390/ijerph17124552.

29. Mutlu EC, Oghaz T, Jasser J, Tutunculer E, Rajabi A, Tayebi A, Ozmen O, Garibay I. A stance data set on polarized con-versations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. Data Brief. 2020 Dec;33:106401. doi: 10.1016/j.dib.2020.106401.

30. Lyu JC, Han EL, Luli GK. COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. J Med Internet Res. 2021;23(6):e24435. doi:10.2196/24435

31. Marcec R, Likic R. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgrad Med J. 2021 Aug 8. doi:10.1136/postgradmedj-2021-140685

32. Aygun I, Kaya B, Kaya M. Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic with Deep Learning. IEEE J Biomed Health Inform. 2021 Dec 7; Epub ahead of print. doi: 10.1109/JBHI.2021.3133103. PMID: 34874877.

33. Alam KN, Khan MS, Dhruba AR, Khan MM, Al-Amri JF, Masud M, Rawashdeh M. Deep Learning-Based Sentiment Anal-ysis of COVID-19 Vaccination Responses from Twitter Data. Comput Math Methods Med. 2021 Dec 2;2021:4321131. doi: 10.1155/2021/4321131. PMID: 34899965; PMCID: PMC8660217.

34. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. Botornot: A system to evaluate social bots. The 25th International Conference Companion on World Wide Web; 2016; Montreal, Canada. pp. 273–274.

35. Bovet A, Morone F, Makse HA. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. Sci Rep. 2018;8(1):8673. doi:10.1038/s41598-018-26951-y

36. Plutchik R. A general psychoevolutionary theory of emotion. In: Robert P, Henry K, editors. Theories of Emotion. Cam-bridge, MA: Academic Press; 1980. pp. 3–33.

37. Anonymous. Natural language processing with disaster tweets. Kaggle. 2021. Accessed 8 October 2021. Available: https://www.kaggle.com/c/nlp-getting-started

38. Mogul R, Renton A, John T, Upright E. December 29 coronavirus pandemic and Omicron variant news. CNN. 2021. Accessed 8 January 2022. Available: https://www.cnn.com/world/live-news/omicron-variant-coronavirus-news-12-29-21/index.html

39. Rattner N. U.S. Covid cases rise to pandemic high as delta and omicron circulate at same time. CNBC Health and Science. 2021. Accessed 8 January 2022. Available: https://www.cnbc.com/2021/12/29/us-covid-cases-rise-to-pandemic-high-as-delta-and-omicron-circulate.html

40. Singh K. U.S. airlines grapple with omicron-related disruptions on last day of the year. Reuters. 2021. Accessed 8 January 2022. Available:

https://www.reuters.com/markets/commodities/covid-driven-flight-delays-cancellations-persist-2021s-final-day-2021-12-31/

41. Anonymous. Choosing the right estimator. Scikit Learn. 2021. Accessed 8 January 2022. Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

42. Bokaee Nezhad Z, Deihimi MA. Twitter sentiment analysis from Iran about COVID 19 vaccine. Diabetes Metab Syndr. 2021 Dec 13;16(1):102367. doi: 10.1016/j.dsx.2021.102367. PMID: 34933273; PMCID: PMC8667351.

43. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul; 63(7):737-45.

44. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. Botornot: A system to evaluate social bots. The 25th International Conference Companion on World Wide Web; 2016; Montreal, Canada. pp. 273–274.

45. Broni K. Emojipedia analysis of over 6.5 billion global tweets dated between 2011-09 and 2021-07. Emojipedia. 2021. Accessed 8 January 2021. https://blog.emojipedia.org/emoji-use-at-all-time-high/

46. Al-Rawi A, Siddiqi M, Morgan R, Vandan N, Smith J, Wenham C. COVID-19 and the Gendered Use of Emojis on Twitter: Infodemiology Study. J Med Internet Res. 2020;22(11):e21646. doi:10.2196/21646

47. Miyake E, Martin S. Long Covid: Online patient narratives, public health communication and vaccine hesitancy. Digit Health. 2021;7:20552076211059649. doi:10.1177/20552076211059649

48. Sloan L, Morgan J, Housley W, Williams M, Edwards A, Burnap P, et al. (2013) Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online. 2013 Aug; 18(3):74-84.

49. Sloan L, Morgan J. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. PLoS One. 2015;10(11):e0142209. doi:10.1371/journal.pone.0142209