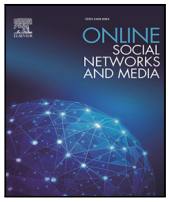
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Online Social Networks and Media 23 (2021) 100135

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem



EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets

Md. Yasin Kabir, Sanjay Madria *

Department of Computer Science, Missouri University of Science and Technology, USA

ARTICLE INFO

Keywords:
COVID-19 data
Coronavirus
Twitter Data
Data analytics
Topics tracker
Emotion analysis
Machine learning

ABSTRACT

The adversarial impact of the Covid-19 pandemic has created a health crisis globally all over the world. This unprecedented crisis forced people to lockdown and changed almost every aspect of the regular activities of the people. Thus, the pandemic is also impacting everyone physically, mentally, and economically, and it, therefore, is paramount to analyze and understand emotional responses during the crisis affecting mental health. Negative emotional responses at fine-grained labels like anger and fear during the crisis might also lead to irreversible socio-economic damages. In this work, we develop a neural network model and train it using manually labeled data to detect various emotions at fine-grained labels in the Covid-19 tweets automatically. We present a manually labeled tweets dataset on COVID-19 emotional responses along with regular tweets data. We created a custom Q&A roBERTa model to extract phrases from the tweets that are primarily responsible for the corresponding emotions. None of the existing datasets and work currently provide the selected words or phrases denoting the reason for the corresponding emotions. Our classification model outperforms other systems and achieves a Jaccard score of 0.6475 with an accuracy of 0.8951. The custom RoBERTa Q&A model outperforms other models by achieving a Jaccard score of 0.7865. Further, we present a historical emotion analysis using COVID-19 tweets over the USA including each state level analysis.

1. Introduction

Every country is taking preventive measurements to fight against the COVID-19 pandemic. By the end of 2020, there were more than 83 million confirmed cases of novel coronavirus globally, and about 20 million people are infected[1] in the USA alone. The number of total fatal cases exceeded 1.8 million globally in 2020. The number of infected people and fatality keeps rising every day. Social distancing or stay-at-home became the most widely used directive all over the world. Social distancing is impacting public events, business activities, the educational domain, and almost every other activity associated with human life. People are losing their jobs and earning sources and thus, the stress level is rising at both the personal and community levels. The emotional responses became overwhelming and inconsistent as people are facing an unprecedented challenge. The studies of behavioral economics show that emotions can deeply affect individual behavior and decision-making.

Social networks have the hidden potential to reveal valuable insights on human emotions at the personal and community level. The monitoring of emotions at fine-grained labels could be valuable during and after the COVID-19 pandemic as the reactions of the people are changing every moment during this unpredictable time. The exploration of tweets to track emotions might play a significant role to understand people's behaviors and responses during the COVID-19 pandemic. The recent works [1–4] show that Twitter data and human emotions analysis can be highly useful and it is not limited to only predict crimes, stock market, election polarity, and managing disasters. Therefore, it is paramount to analyze the social media data to understand the human behavior and reaction in the ongoing pandemic. To find out useful insights from the public reactions and shared posts in social media, and to model the public emotions, we have started collecting tweets from 5th March 2020. We have collected and processed over 600 million tweets related to Coronavirus (focused on the USA only) which is more than 4.5 terabytes in raw data. We developed a web application that processes the collected data in real-time and produces interactive graphs and charts. The website is accessible publicly and enables anyone to observe the sentiments, topic trends, and user mobility with interactive visualizations including maps, time charts, and word clouds. Detailed information about the website and visualizations is available in [5].

* Corresponding author.
E-mail addresses: mkabir@mst.edu (M.Y. Kabir), madrias@mst.edu (S. Madria).

[1] https://mykabir.github.io/coronavis/index.html.

https://doi.org/10.1016/j.osnem.2021.100135
Received 21 September 2020; Received in revised form 26 February 2021; Accepted 3 April 2021
Available online 16 May 2021

There is a wide range of research works available where sentiments are explored using different techniques. Sentiments analysis [6–9] became a popular field of natural language processing. In most of the sentiment analysis work, sentiments are explored considering high-level emotion categories such as positive, neutral, and negative. Several works also considered sentiment as a form of feeling using numerical scores such as 1 to 5 defining very bad to very good or something like that. However, to understand the emotional response of the people and correlate that with the socio-economic situation, we need fine-grained labels of emotion. For example, labeling the emotions like sadness, worry, or angry as negative sentiment only might not enable us to understand the proper reaction of a person as all three of those emotions may lead to different behaviors and decisions. Furthermore, while detecting and labeling the emotions into different categories is highly useful, it is also necessary to understand the reasoning behind an emotion. People might be angry or sad for different causes, and treating all of those causes similarly might not be ideal. To understand the reasoning behind an emotion, it is necessary to label a few words or a phrase from a text which will enable us to understand the emotions better and use them appropriately.

However, there is a lack of available labeled emotion data. In our research, we were able to find two available tweet emotion datasets. One of those datasets [10] has a total of 14,827 annotated tweets in 11 emotion categories (e.g. anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust) comprises with English, Arabic, and Spanish tweets. However, this dataset does not contain any COVID-19 related tweets. The other dataset [11] which is annotated using COVID-19 tweets contains 10K English tweets and 10K of Arabic tweets in 10 different categories (e.g. optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official report, and joking). We found that this dataset is useful for developing machine learning models to automatically detect and classify the tweet's emotions. However, the 10K labeled English tweets for 10 different categories are fairly low for creating an effective machine learning model. Moreover, none of those datasets provide the selected words or a phrase denoting the reason for the corresponding emotions.

Due to the lack of available datasets, most of the research works on COVID-19 sentiments such as [5,12,13] are mostly limited to the positive, neutral, and negative sentiments or researcher rely on some available API or lexicon-based tools that provides emotion categories without understanding the proper context which essentially is not appropriate for fine-grained emotion analysis tasks. There is also a lack of available machine learning models to automatically classify the emotion in the tweets using the context. To resolve those problems, we started annotated COVID-19 English tweets manually to 10 different emotion categories (e.g. neutral, optimistic, happy, sad, surprise, fear, anger, denial, joking, pessimistic) as well as also select the words or phrases that are mostly responsible for the selected emotion label. The phrase selection makes our dataset unique as there is no other such data available on COVID-19 tweet emotion to the best of our knowledge. Our annotated dataset can be used with the conjunction of the available dataset by [11] for the similar emotion labels to classify the emotion of the tweets and train a better machine learning model. In this work, we not only presented our dataset but also develop and train machine learning models to detect the emotion of the tweets and extract phrase which is mostly responsible for the detected emotion. We explore and created custom pipelines for the classification and phrase extraction tasks and perform a comparative study of the model performance. The primary contributions of this work are:

- A multi headed binary classifier using deep learning to automatically classify the COVID-19 tweets into above specified 10 emotions. The classifier determine the high-level relationships among the labels, and extract a contextual representation of the tweets to detect different emotions. The developed classification model achieves a Jaccard score of 0.6475 with an accuracy of 0.8951 outperforming other systems.
- A custom Q&A roBERTa model to extract the phrase that is mostly responsible for a particular emotion on a tweet. The model predicts the positions of the start and end tokens from a given text that represent the specified emotion. The proposed model achieves a score of 0.7865 in Jaccard metrics.
- Manually labeled (by three annotators) 10,000 tweets into 10 different emotions (e.g. neutral, optimistic, happy, sad, surprise, fear, anger, denial, joking, pessimistic). Along with the labels, we also selected the phrase that might be responsible for the respective emotion.
- An experimental historical emotion analysis on COVID-19 tweets using the developed classification model.

2. Related works

Throughout the recent years social media has seen a tremendous increase in its use during times of crisis. Many researchers from all around the globe are creating COVID-19 datasets using Twitter APIs [14,15] Putting together millions of tweets composed of largely English tweets related to keywords like: covid, corona, pandemic, and quarantine similar to those used in our research for the initial collection of tweets. Researchers are investigating methods of promoting healthy social media use during times of pandemic similar to the COVID-19 outbreak. With the growing amount of open data available relating to the public opinion from platforms like Twitter, Facebook, Instagram, Snapchat, Tumblr, LinkedIn, Youtube, Twitch, and Reddit [16,17] as grown exponentially. Researchers are looking into different methods with the hope of developing an effective method of utilizing all the public data available through Twitter. Some suggest that the possibility of reaching an accuracy of sentiment classification is between 60–80 percent [9,18].

Twitter being especially popular for anomaly detection, response and communication monitoring during crisis (disease outbreaks [19, 20], hurricanes [21–23], floods [24], terrorist bombing [25], misinformation propagation [26,27] and others [28,29]). Lisa et al. [30] and Ramez et al. [31] presented their works on misinformation propagation and quantification during COVID-19 using twitter. The authors in [31] conclude, there is an alarming rate of medical misinformation and non-credible content sharing on Twitter throughout the pandemic. It is very crucial to quantify the misinformation on social media and take the necessary action to prevent unnecessary anxiety and medically harmful methods to fight against COVID-19.

Catherine et al. [32] are exploring the possibility of illustrating topics such as spreading of corona case, healthcare workers and personal protection equipment (PPE) and seventeen others using a pattern matching and topic modeling system with Latent Dirichlet Allocation (LDA). The authors are investigating the use of five methods of analysis on features like key terms and features, information dissemination and propagation and network behavior during COVID-19 pandemic. These produced a model that could detect high level topic trends in news briefings over time. Alaa et al. [33] also performed topic modeling using word frequencies and Latent Dirichlet Allocation (LDA) with the aim to identify the primary topics shared in the tweets related to the COVID-19. Choudhury et al. [34] developed a dataset of classified tweets for a more refined set of emotions. Using a hashtag word classification system the authors were able to classify millions of tweets quickly. An example of this would be the translation of the word smile into the class of joviality.

Although there are many works available on tweet classification and phrase extraction we found only a few attempts to classify the tweets emotion during the COVID-19 pandemic using context-based machine learning models as there is a lack of available datasets. Most of the traditional tweet emotional classification works [10,35,36] treat the problem as a text classification problem and rely on a large amount of labeled data and focus mostly on effective feature engineering. Baziotis et al. [37] and Meisheri et al. [38] who hold the first and

second place of the multi-label emotion classification task of SemEval-2018 Task1, developed classifiers using a bidirectional LSTM with an attention mechanism. Using two different trained models: regularized linear regression and logistic regression classifier chain, Park et al. [39] try to classify the emotions for the same problem discussed above. The authors captured the correlation of emotion labels using logistic regression classifiers. However, none of those works perform emotion classification on a crisis datasets which might represents a wide verify of emotions with unbalance labeled data. Yang et al. [11] introduce a COVID-19 dataset and implemented XLNet, AraBert, and ERNIE for classifying the emotion in English, Arabic, and Chinese language text respectively which is the only available emotion classification work on the COVID-19 tweets or text. For phrase extraction there are several transformation based models [40,41] available from different research works. However, to our knowledge there is no available phrase extraction work on tweets emotion.

While there are ongoing research works for emotion detection and classification using the tweets there is a lack of publicly available datasets. Moreover, in most of those works, researchers are trying to label and detect emotion categories only for the tweets. However, the phrase that is responsible for a particular emotion in a tweet might help us understand the tweets better and can allow us to dive deep into data mining on emotional response. There is also a lack of available machine learning model that is developed particularly for automatic emotion detection of the COVID-19 tweets. In this work, we present the EMOCOV dataset that provides emotion category labels along with the phrase responsible for that emotion. We also propose two different machine learning models: one is for emotion classification using deep learning approach with attention mechanism and auxiliary features input, and another one for extracting the responsible phrase for that emotion using a custom Q&A roBERTa head.

3. Data collection, annotation and description

At the early stage of our research, we have performed data analysis to observe and understand the differences between the available Twitter datasets for sentiment analysis and COVID-19 tweets. We observed that due to the ever-evolving nature of the tweets' linguistics and the newly allowed length of the tweet text (280 vs previous 140) there are noticeable contrasts between the available datasets and recent tweets. Moreover, during the ongoing pandemic, there is a frequent change in the events, guidelines, restrictions, news which creates a roller-coaster ride of emotions. Fig. 1 represents the word clouds created using the tweet texts from the ongoing COVID-19 dataset and using a combined dataset created from the Crowdflower sentiment dataset and SemEval-2018 dataset. We randomly select 5000 data points from each category for generating word clouds. Fig. 1(a) depicts the word cloud for COVID tweets, and Fig. 1(b) represents the word cloud for the combined dataset of non-COVID tweets. From the figures, we can observe a good variation among the frequent words in the datasets. While general tweets contain usual words (e.g., love, going, today, thank) in the texts, COVID tweets are dominated by the words specifically related to the ongoing pandemic (e.g., death, patient, lockdown, death). We can also observe that only a few words in the non-COVID dataset are very frequent while the frequency of the top words in the COVID-tweets is much closer which is represents by the size of the words. We have also noticed emotional variations among the people for the same news or events. For example, while many people considered lockdown as positive, there were another group of people who were against it. Therefore, the same words with a little variation changed the emotion of the tweets. Machine learning models are highly dependable on the quality of the data. Most of the models rely on good data annotation and embedding techniques. This encouraged us to create our own for emotion analysis on the ongoing pandemic. Further, to make a robust model that can adapt to the change of the emoticon and punctuation uses in the tweets, we have developed a deep learning model pipeline. In the following subsection, we briefly describe the process of data collection and data annotation along with an overview of our dataset.

3.1. Data collection

We are collecting tweets since 5th March 2020 using Twitter Streaming API and the python Tweepy package. We have collected more than 500M tweets in 2020. We run the queries using COVID-19 related keywords (e.g. COVID, corona, coronavirus) for the tweets collection. The module listens to the stream of the tweets and tries to check if a tweet text contains any of the desired words. While checking the module it converts all the text to lower case and tries to find out sub-strings within the text. By doing this, the module identifies a qualified tweet and saves it in the JSON format. Further, the collected data is processed in real-time for the CoronaVis[2] application. We will keep collecting the data and update the collected tweets ids in the data repository[3] periodically. The repository contained those tweet ids for which we were able to estimate a state-level geo-location.

3.2. Data annotation

We randomly selected 10K English tweets generated from the USA for the emotion annotation from the collected COVID-19 tweets in our first phase of data annotation. The tweets are annotated manually by 3 different people to reduce any bias. Among three annotators, one is a Ph.D. student working on social media data mining since 2017. The other two annotators are undergraduate students from the computer science department and are native English speakers. We have selected 10 dominant emotions based on the study in [34] to label the tweets. Those 10 labels are neutral, optimistic, happy, sad, surprise, fear, anger, denial, joking, and pessimistic. Each tweet has annotated with primary and secondary emotion based on the tweet text. The primary label is selected from the majority agreement of all the annotators considering both primary and secondary labels. For example, if an annotator selected "Optimistic" as the primary label and another annotator selected "Optimistic" as a secondary label for a tweet, we have considered the primary label for that tweet as "Optimistic". The secondary emotion is selected based on the majority agreement. If a majority agreement is unavailable, then that the tweet was discarded. By this process, the agreement for primary emotion between two annotators is 87% and the agreement from three annotators is 68%. For the secondary emotion, the inter-annotators agreements are 54% and 41% respectively by two annotators and three annotators. Further, the annotators marked a phrase associated with the primary emotion for each tweet. The whole tweet text has been selected for the tweets with neutral emotions. We will share our annotated emotion data publicly for further research and analysis.

3.3. Data description

3.3.1. COVID-19 tweets data

Table 1 represents a high-level summary of the tweets ids that is available in the git repository. However, we are continuously collecting the data and thus the data statistics can be changed in the repository with future updates. In the repository, we have included processed tweet ids that have geolocation information. However, we will also include the list of all tweets ids with or without geolocation information.

The processed tweets ids are saved and updated in the git repository within the folder named as data. The data folder contains several csv files. Every file contains tweets ids fetched in the respective date that is specified as the name of that file. For example, 2020-03-05.csv contains the tweets that was fetch on 5th March, 2020. The name was formatted as Year-Month-Date.

[2] https://mykabir.github.io/coronavis/.
[3] https://github.com/mykabir/COVID19.

Fig. 1. Word clouds from: (a) COVID-19 tweets, (b) Non-COVID tweets.

Table 1
COVID-19 tweets data summary.

Attribute	Summary
Collection period	March 5, 2020 to December 31, 2020.
Number of unique tweets	5,60,14,158.
Location	USA (state label).
Number of unique users	Total: 54,27,831; Verified: 56,387;

Table 2
The label distributions in COVID-19 annotated emotion dataset (%).

Type	Neu.	Opt.	Hap.	Sad	Sur.	Fea.	Ang.	Den.	Jok.	Pes.
Primary	23.47	8.43	8.29	7.82	16.64	8.79	16.83	1.16	4.64	3.93
Secondary	38.54	7.90	3.99	12.61	7.17	9.28	4.99	1.43	2.21	11.86

3.3.2. Annotated EMOCOV data

Table 2 provides the label distributions of different types of emotions in the annotated datasets. We can see that there is a good variation in the label distribution. We can also see that a large number of tweets were annotated in the Surprised, Anger, and Neutral categories where there are only a few tweets in Denial and Joking categories.

Table 3 presents a few examples of annotated tweets. The first emotion is the primary emotion and selected text represents the reasoning behind that emotion. Combining the labels from different annotators we decide the primary and secondary emotions. In Section 5.2.2, Table 12 contains few more examples of phrase selection by annotators where we discuss the performance of our model.

4. Emotion detection and extraction

4.1. Neural network for emotion classification

We develop a Deep Neural Network to classify the tweet text into a specific emotion category. To create the network, we modify the deep neural network that we have proposed in our previous research work [4]. Fig. 2 illustrates the architecture of the model starting from input sequence generation. The modified deep learning model comprises 6 primary components.

1. Input layer: Processed tweets are used as input in this layer as vectors.
2. Embedding layer: Using lookup tables, this layer encodes the input into real-valued vectors. We used a pretrained word vectors named GloVe [42] which generates a feature word vectors using co-occurrences based statistical model. This layer map all tokenized words in every tweet to their respective word vector. Padding is used at the end of the vector list for the tweets with shorter length.
3. BLSTM layer: The Long–Short Term Memory (LSTM) is a specialized version of Recurrent Neural Network (RNN) that is capable of learning long term dependencies. While LSTM can only see and learn from past input data, Bidirectional LSTM runs input in both forward and backward direction. This bidirectional feature of BiLSTM is critical to understand of complex language context [43].
   The implemented LSTM version in this work can be defined by the Eqs. (1)–(5) where the input gate i_t, forget gate f_t, output gate o_t, and cell state activation c_t. In the equations σ represents the logistic sigmoid function, h represents the respective hidden vectors, and W is the weight matrix. A detailed explanation of each equation and more about LSTM is available in [44].

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (2)$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \quad (4)$$

$$h_t = o_t \tanh\left(c_t\right) \quad (5)$$

4. Attention layer: We use a word-level deterministic, differentiable attention mechanism to identify the words with the closer semantic relationship in a tweet. Eq. (6) represents the attention score $e_{i,j}$ of each word t in a sentence i and g is an activation function. More information on the attention mechanism is available in [45].

$$e_{i,j} = g\left(W h_t c\right) \quad (6)$$

5. Auxiliary features input: A tweet can only contain 280 characters which forces a user to express emotions in a different way compared to a traditional English sentence. People use extra punctuations and emoticons to intensify the meaning of a

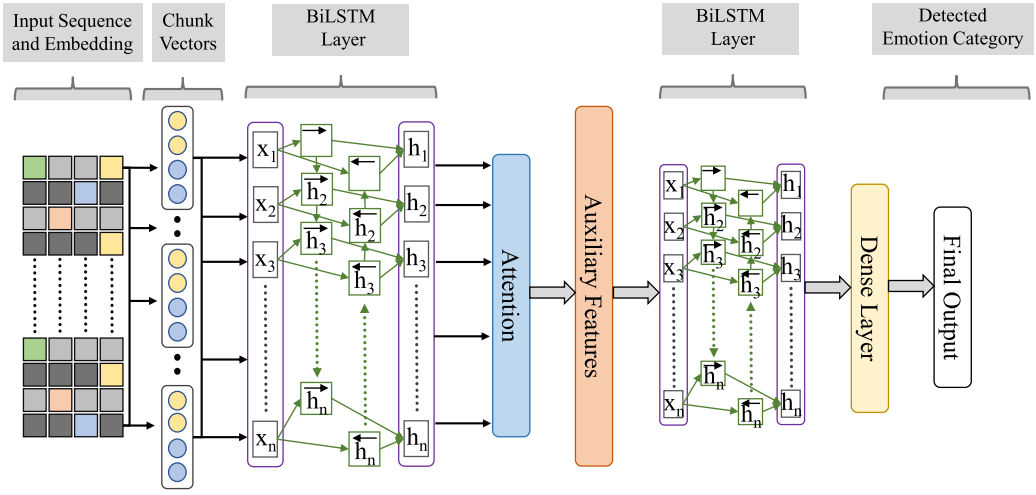


Fig. 2. The illustration of the emotional classification Deep Neural Model.

Table 3
Example of annotated tweets.

Example tweet and selected text	Emotion category
Tweet: Relief provided to the poor needy during lockdown and to facilitate medical reserves to combat COVID Selected text: Relief provided to the poor	Happy Optimistic
Tweet: In the Covid era mathematical models are deciding matters of life and death. @mathbabedotorg explains how they wor… Selected text: mathematical models are deciding matters of life and death	Surprise Fear
Tweet: We pay an obscene amount of taxes in NY. We are not broke bc of COVID. We are broke because #GovernorDeath puts illegals… Selected text: We are broke	Anger Pessimistic

Table 4
Auxiliary features.

Polarity, subjectivity, wordsVsLength, digitVsLength, punctuationVsLength, nounsVsWwords, SadVsWords, capitalsVsWords, uniqueWords, TagNumbers.

tweet. We perform feature engineering to obtain a set of specific auxiliary features that can assist the classification model. A list of extracted auxiliary features that shows noticeable influence during the model evaluation is given in Table 4. The well-known NLTK package is used to extract those features.

6. Output layer: The output layer is created using dense layers which use *sigmoid* as activation function and predict the output class. The layer produces binary values for all the label categories.

4.1.1. Classification model parameters and training

A set of optimal parameters is crucial for achieving the desired performance results. We performed rigorous parameter tuning and selected an optimal set that is used in all the experiments. We used the same set of parameters as presented in Table 5 for performance evaluation and model reproducibility. To build a robust model, we used 5-fold cross-validation with an 80/20 split ratio for training and testing. Initially, we have trained and tested our model starting from 20 epochs to 100 epochs. To get the optimal learning rate, we employed an LR-scheduler with an initial learning rate of 0.001. We observe that the learning rate drops to 0.00001 by the time the model reaches the best validation score. We noticed that each model performed best around 40 epochs and after that start overfitting. Therefore we use 50 epochs for the final training and testing.

Table 5
Hyperparameter values.

Hyperparameter	Value/Description
Text embedding	Dimension: 250
BLSTM layer	2 layers; 250 hidden units in each (Forward and backward)
Dense layer	3 layers; First 2 layers have 150 and 75 units respectively and the last one is output (Dense)
Drop-out rate	Word embedding: 0.3; Dense layer: 0.2 each;
Activation function	Conv1D, BLSTM, Dense: ReLU; Output dense layer: Sigmoid;
Adam optimizer	Learning rate 0.001–0.00001; $beta_1 = 0.8$;
Validation	Training and validation split = 80/20;
Epochs and batch	Epochs = 50; batch size = 68;

4.2. Custom RoBERTa for phrase extraction

RoBERTa, a Robustly Optimized BERT Pretraining Approach [40] is developed using the Google's BERT language masking strategy [46]. The accuracy of RoBERTa is 2%–20% higher compared to BERT. Both of these approaches provide transformer to learn a language representation. However, BERT is more suitable for Question and Answer problem solution as BERT tries to predict the Next Sequence Probability of a token. As RoBERTa does not use NSP, we have to develop a custom Q&A head to predict the probability of the start and the end sequence of the tokens.

The developed model uses two Q&A heads that is illustrated in Fig. 3 for the position of start sequence of a phrase and the end position of the sequence. Practically, the model provides a probability for each character position for being a start or end sequence. Further, using the maximum probability value, the model selects the final start and the end positions. Primarily, the developed models have the following three components:

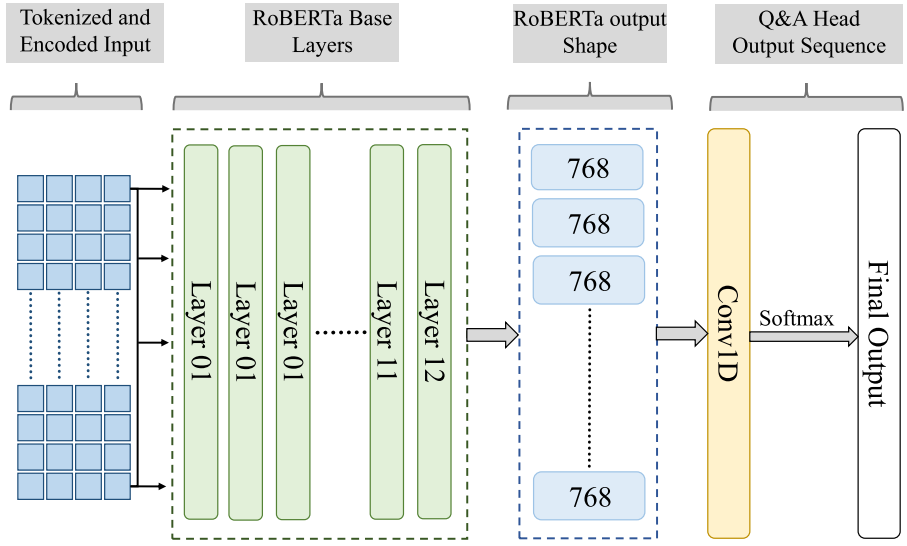


Fig. 3. RoBERTa model illustration with custom Q&A head.

1. Tokenizer: The tokenizer takes the input text and split it into words using the black space between the characters. It performs the similar split for both the input tweet text and the selected text. Further, it translates those words into the respective numerical values using RoBERTa base vocabulary files. After that it creates two masked lists where any other values apart from the start and end positions of the sequence is set as 0. The formal list puts 1 for the start position and the second list puts 1 for the end position of the selected text sequence. The tokenizer also creates a same size attention mask for all the input tweet texts where available words position presented as ones with the padding zeros.
2. RoBERTa Base: We used pretrained RoBERTa base model for further training with our input data. RoBERTa base uses the BERT-base architecture with 125M parameters. For the implementation, we used Huggingface transformer library. Detailed information about RoBERTa base is available in Liu et al. [40].
3. Custom Q&A head: We created a custom Q&A head for the start and the end position prediction of the sequence. RoBERTa is developed primarily for question and answering task. In our model, we treated it for the similar purpose where the emotion label is the question, and the selected phrase is the respective answer. To achieve that, we use a convolution layer that transform the base output of RoBERTa to a pre-determined vector size. Further, applying the softmax function, it produces two one hot encoded lists for the starting and the ending index position of the given text.

4.2.1. Model parameters

We used several parameters to tune our model. Table 6 demonstrates the final parameter for our model with the best performance.

5. Experimental result and analysis

We present the experimental result and historical emotion analysis in this section. We use two different machines to perform data collection, model training, and analysis. We use a machine with Intel Xeon E5-2650 v4 @ 2.20 GHz CPU (12 cores, 24 threads) with 64GB RAM and an Nvidia RTX-2070 super GPU. Another machine comprises of Intel® Core™ i9-9900K CPU @ 3.60 GHz (8 cores, 16 threads) with 64GB RAM and an Nvidia RTX-2080Ti GPU. In the following subsections, we describe the evaluation metrics, experimental results, and emotion analysis.

Table 6
Hyperparameter values.

Hyperparameter	Value/Description
MAX input length	196
Pre-trained network	RoBERTa base
Dense layer	2 layers; One for start position and one for end position of the sequence.
Dropout	0.1 before each output dense layer
Activation function	Output dense layer: Softmax;
Cross validation	Folds = 5; Training and validation split = 80/20;
Epochs and batch	Epochs = 10 (each fold); batch size = 68;

5.1. Evaluation metrics

To evaluate the classification model, we have used Micro F1, Macro F1, Jaccard, and Accuracy. Let L denotes the number of label categories, TP denotes True Positive, FP denotes False Positive, and FN denotes False Negative. We can define F1 micro average score using Eqs. (7)–(9).

$$Precision_{micro} = \frac{\sum_{k=1}^{L} TP_k}{\sum_{k=1}^{L}(TP_k + FP_k)} \tag{7}$$

$$Recall_{micro} = \frac{\sum_{k=1}^{L} TP_k}{\sum_{k=1}^{L}(TP_k + FN_k)} \tag{8}$$

$$F1_{micro} = \frac{2 * Precision_{micro} * Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{9}$$

Eqs. (10)–(13) denote the macro average F1 score calculation which is a simple averaging of F1 scores for different labels.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{12}$$

$$F1_{macro} = \frac{1}{|L|} \sum_{k=1}^{L} F1_k \tag{13}$$

Jaccard score is a popular metrics for multi-label binary classifier accuracy as this metric consider every label category similarly. The jaccard score is calculated using Eq. (14). In the equation, T denotes

Table 7
The label distributions in combined emotion dataset (%)

		Neu.	Opt.	Hap.	Sad	Sur.	Fea.	Ang.	Jok.	Pes.
# data		10K	25K	15K	25K	25K	15K	15K	20K	25K
% label	**Neu.**	23.47	7.18	11.12	7.18	7.18	11.12	11.12	10.77	7.18
% label	**Opt.**	8.43	15.35	13.03	15.35	15.35	13.03	13.03	14.33	15.35
% label	**Hap.**	8.29	9.83	15.23	9.83	9.83	15.23	15.23	3.89	9.83
% label	**Sad**	7.82	14.66	13.05	14.66	14.66	13.05	13.05	13.19	14.66
% label	**Sur.**	16.64	11.75	9.96	11.75	11.75	9.96	9.96	16.04	11.75
% label	**fear**	8.79	6.40	9.91	6.40	6.40	9.91	9.91	4.16	6.40
% label	**Ang.**	16.83	12.72	19.71	12.72	12.72	19.71	19.71	7.93	12.72
% label	**Jok.**	4.64	14.67	2.47	14.67	14.67	2.47	2.47	22.02	14.67
% label	**Pes.**	3.93	7.43	5.52	7.43	7.43	5.52	5.52	7.67	7.43

the number of test data, Y_k denotes the truth label of data k, and P_k denotes the predicted label.

$$Jaccard = \frac{1}{|T|}\sum_{k=1}^{T}\frac{Y_k \cap P_k}{Y_k \cup P_k} \tag{14}$$

$$Accuracy = \frac{1}{T}\sum_{k=1}^{T}\sigma(Y_k == P_k) \tag{15}$$

We used accuracy as another metrics as it provides a better observation of model performance while the data has imbalanced categories. Eq. (15) defines the Accuracy score where $\sigma(Y_k == P_k)$ returns 1 if the prediction for a data is correct, otherwise it returns 0.

We evaluated phrase extraction model using word-level jaccard similarity score. It calculates the performance using the similarity between the predicted words respective ground truth. In Eq. (16), Y_k denotes the ground truth string, and P_k refers to the predicted string.

$$Jaccard_{similarity} = \frac{1}{|T|}\sum_{k=1}^{T}\frac{len(Y_k \cap P_k)}{len(Y_k) + len(P_k) - len(Y_k \cap P_k)} \tag{16}$$

5.2. Experimental results

5.2.1. Classifier evaluation

We evaluate our proposed classification model using two different data sets. First, we use our own labeled emotion data that we described in Section 3.3.2. The primary and secondary emotion label was processed as distinctive data points for the classification purpose. For example, if a tweet has a label of Angry and Pessimistic, we use that tweet for both of the label categories individually. Further, we have created an aggregated data combining our data, the emotion dataset by Yang et al. [11] and the emotion classification dataset of SemEval-2018 Task1: Affect in Tweets [10]. For the aggregated data, we evaluate the models only on the selected labels that are similar across all the three datasets. We have converted some of the labels to reduce the imbalance in the label categories. The aggregation module produces a combined dataset across 9 different emotion categories that are presented in Table 7. The *#data* row in the table presents the number of tweets that are used for training, validation, and testing of the classification models for each category. Other rows indicated as *%label* represents the distribution of labeled data in each dataset. To elaborate, to train and test the models to identify the neutral tweets we used a dataset containing 10K tweets where 23.47% tweets were neutral and the rest of the tweets were labeled as other emotion categories. To train and test, those 23.47% tweets were assigned binary label 1 - indicating neutral emotion, while the rest of the 77.53% tweets was assigned label 0 - indicating non-neutral tweets. Similarly, for optimistic we used a dataset containing 25K tweets, where 15.35% tweets were optimistic and rest of the tweets were labeled as other categories.

Tables 8 and 9 represent the performance of 3 different classification models. To compare our model performance, we compare our model with SVM-Unigrams [10] and NTUA-SLP [37]. NTUA-SLP is the submitted system that became the winner of the SemEval-2018 Task1: E-challenge. For our annotated dataset which has highly imbalanced

Table 8
Classifier evaluation and comparison.

Model	F1-Micro	F1-Macro	Jaccard	Accuracy
SVM-Unigrams	0.5294	0.4076	0.4138	0.7383
NTUA-SLP	**0.5981**	0.4887	**0.5472**	0.8492
BiLSTM$_{AAf}$(Our)	0.5514	**0.5392**	0.5366	**0.8647**

Table 9
Classifier evaluation and comparison using combined emotion data.

Model	F1-Micro	F1-Macro	Jaccard	Accuracy
SVM-Unigrams	0.5532	0.4849	0.5185	0.8227
NTUA-SLP	**0.7058**	0.5829	0.6293	0.8746
BiLSTM$_{AAf}$(Our)	0.6893	**0.6342**	**0.6475**	**0.8951**

categories, NTUA-SLP performed better in F1-Micro and Jaccard score. However, our model performed better in the other two metrics. Our model BiLSTM$_{AAf}$ outperforms both SVM-Unigrams and NTUA-SLP in terms of F1-Macro, Jaccard, and Accuracy while we train and test those models using the combined dataset described in Table 7.

Table 10 represents some sample tweet texts and respective emotion labels predicted by our proposed model (BiLSTM) and NTUA-SLP. We omit SVM-Unigrams from this comparison as the performance of this model is considerably lower. Although both models predicted labels for all of the emotion classes for a given text, here we only present the emotion labels for which the models have different predictions. Column 'GT' in the table denotes the ground truth (annotated) labels. We observe that NTUA-SLP is struggling with sarcastic and contrasting emotions. For example, in the first tweet, the tone of the text seems happy until we see the word #ignorant. Due to this hashtag, we can infer that this tweet is sarcastic. In the second and third tweets, we observe contrasting emotions or meanings. While the struggle of the families during covid is sad, stimulus check brings optimism. In the third tweet, the literal meaning of the word 'Losing' is not something positive. However, losing weight could be a positive thing. We find it fascinating that our proposed model is doing well do identify these contexts compared to the NTUA-SLP. To find out the probable reason behind this we perform several evaluations. In the evaluation, we observe the impact of auxiliary feature input that we describe in Section 4.1. Using auxiliary features input we explicitly provide a set of features that helps the model to detect the contrast in the tweet. For example, in Table 10, we observe a significant number of capital words or letters in the tweets with contrasting meanings. The auxiliary features input helps the model to catch this information which is otherwise might have less impact due to the attention on the words and word-embedding. By the architecture, NTUA-SLP uses a self-attention mechanism that identifies the dominant words related to the emotion. However, this leads to misclassification in some cases. To confirm this hypothesis we further train and evaluate our model without using the auxiliary features input. Without the auxiliary features, the performance of the model drops by 5%–10% for different emotion classes. Furthermore, we assess the weakness of our model. Our proposed model underperforms for the emotion classes with small training data such as pessimistic and

Table 10
Sample output comparison between the proposed model and NTUA-SLP model.

	Tweet text	Emotion	BiLSTM$_{AAf}$	NTUA-SLP	GT
1	Those who are following trump regarding MASK, have a happy get together. #covid #ignorant	Happy	0	1	0
		Anger	1	0	1
2	With all of the sad news during COVID, the only hopeful thing is Stimulus check for the struggling family.	Sad	1	1	1
		Optimistic	1	0	1
3	If this #lockdown does not end now it will not be just the covid that is flattened but the economy FLATLINED.	Sad	1	1	1
		Anger	1	0	1
4	Plans to alter own clothes after los ing 17 Pounds in COVID-19 #Lockdown	Happy	1	0	1
		Optimistic	1	0	1
5	This is nothing more than targeting the old to get increased numbers of deaths with COVID. OBVIOUS!	Pessimistic	0	1	1
		Anger	1	0	1

Table 11
Performance evaluation of the phrase extractors.

Model	Jaccard	Jaccard (EXT.)
BERT base	0.6852	0.7349
ALBERT	0.6879	0.7529
Custom RoBERTA (Our model)	**0.7196**	**0.7865**

fear. The 5th tweet in the table represents such an example. NTUA-SLP is outperforming our model for such a situation. In the future, we are planning to develop and train multiple models architecture with and without auxiliary features and ensemble those models to address the weakness of our model.

5.2.2. Phrase extraction evaluation

Similar to classifier evaluation we evaluate models for phrase extraction using our annotated dataset and a combined dataset that is available in Kaggle Tweet Sentiment Extraction competition.[4] However, due to some automated data processing, there were some issues in the text in the available dataset. We processed that dataset using the original tweet text that is available in crowdflower dataset.[5] Combining our data with the external dataset, we were able to make the models robust and it increased the performance of the models. Table 11 represents the performance evaluation of the models. In the table Jaccard (EXT.) denotes the performance of the model when we also used the external data for model training and testing. The developed RoBERTa model with a custom Q&A head outperforms both BERT and ALBERT models for both datasets. For BERT and ALBERT implementation, we have used the publicly available top kernels used and available in the Kaggle Tweet Sentiment Extraction competition.

Few examples of phrase extractions are presented in Table 12 to demonstrate the effectiveness of the model in different contexts. The table also includes the output from BERT and ALBERT models along with our proposed Custom RoBERTa model. We observe that in most cases, all of the models selected smaller phrases or fewer words compared to the annotators' selection. However, our proposed model selected longer phrases in many cases compared to other models. All three models follow the similar concept of question and answer modeling. In the context of this work, the provided emotion acts like a question and the answer is the selected phrase by the models related to the given emotion. Both BERT and ALBERT encode each word in a tweet and selected text. However, we created a custom head in our model which encodes each letter in the text instead of the word. Hence, while BERT and ALBERT try to predict the starting and ending word positions, our model tries to predict the starting and ending letter positions. We believe this behavior is the primary reason for the better performance of our model as it helps to mimic the longer phrase. In Table 12, we observe both BERT and ALBERT are omitting the preposition, adverbs, or adjectives in the predicted text in many cases. For example, both models omitted "should, biggest, have, and some" for tweets 1–4. While our proposed model included those words. In the 5th tweet, our model predicted 'What kind of' compared to the 'What kind' predicted by ALBERT. The research on the phrase extraction models is still in the primary stage for emotion context. Also due to the subjectivity of the annotators, the selected text varies a lot. The models perform miserably with fear, surprise, and sarcastic tweets. In future, we need to conduct more experiments and analysis to have more concrete reasoning on why the models performing differently. Also, we need more data for generalizing the models better.

5.3. Historical tweets emotion analysis

In this section, we present the historical emotion analysis on the COVID-19 tweets. We present the analysis of the six dominant emotions (e.g. Happy, Sad, Optimistic, Pessimistic, Fear, and Anger) all over the USA. Further, we analyze the emotions of six individual states (NY, CA, CO, TX, MO, and FL) to perform a comparative study of the emotions among the states from the east coast, midwest, and west coast. To infer the state from the tweet we have used geo-tag and user profile information. If a tweet is not geo-tagged, we fetched the user profile to lookup the location info. We discarded the tweets if we were unable to infer a location. We have also discarded tweets from any user profile which has more than 5 tweets on a day. This is to ensure the filtering of the spamming and also reducing the bias of having tweets from the same person. We have also removed the duplicates or retweets. Using our location detection strategy and filtering module, we get more than 56M tweets originated from the USA from 5th March 2020 to 31st December 2020. On average there are 188765 tweets per day. For the above specified six states that we have used for the analysis have the following numbers tweet per day on average: NY-11419, CA-24230, CO-3681, TX-19328, MO-2297, FL-13014. For the analysis, we use our proposed machine learning model to classify the tweet emotions. In this section, we include analysis on weekly and monthly emotion distribution. However, we primarily focus on the monthly analysis at which enables us to correlate the important events during the pandemic in limited space. To calculate the emotion scores in the figures, we use the weekly and monthly mean of the classified tweets emotions.

Fig. 4 presents the weekly ratio of emotion categories. We can see that happy, sad, and fear are the identified emotions for most of the tweets. We observe that while 70%–80% of the tweets are showing those 6 emotions, there are still 20%–30% tweets that are either neutral or can be categorized in other emotion categories. In the figure, Y-axis represents the distributions of emotions on a scale of 0 to 1. The distribution is calculated using the total number of tweets identified for an emotion divided by the total number of tweets in that periods. For example, in the first week the distribution of the

[4] https://www.kaggle.com/c/tweet-sentiment-extraction/.
[5] https://data.world/crowdflower/sentiment-analysis-in-text.

Table 12
Example of phrase extractions by proposed model.

	Example of phrase extraction
1	Tweet: Almost 70% of PA's Covid-19 deaths 2611 of 3806 have occurred in nursing homes or long-term care — PA should never have... Emotion: Anger, Selected text: pa should never have **BERT Base:** never have, **ALBERT:** never have **Custom RoBERTa:** should never have
2	Tweet: Rare Thai Turtle Nests Make Biggest Comeback In 20 Years Thanks to COVID-19 Emotion: Happy, Selected text: biggest comeback in 20 years **BERT Base:** comeback, **ALBERT:** thanks **Custom RoBERTa:** biggest comeback
3	Tweet: If hygiene JUST became a priority for you ... you have bigger issues than Corona. Emotion: Pessimistic, Selected text: bigger issues than Corona **BERT Base:** bigger issues, **ALBERT:** bigger issues **Custom RoBERTa:** have bigger issues
4	Tweet: Fight Corona by staying indoors. Spend some quality time with your family that is otherwise difficult in our busy schedules. Emotion: Optimistic, Selected text: Spend some quality time **BERT Base:** quality, **ALBERT:** quality time **Custom RoBERTa:** some quality time
5	Tweet: He believes the Democrats want people to die of COVID-19 so they can win the election? What kind of hatred is in his heart!! Emotion: Surprise, Selected text: What kind of hatred is in his heart!! **BERT Base:** election? What, **ALBERT:** What kind **Custom RoBERTa:** What kind of

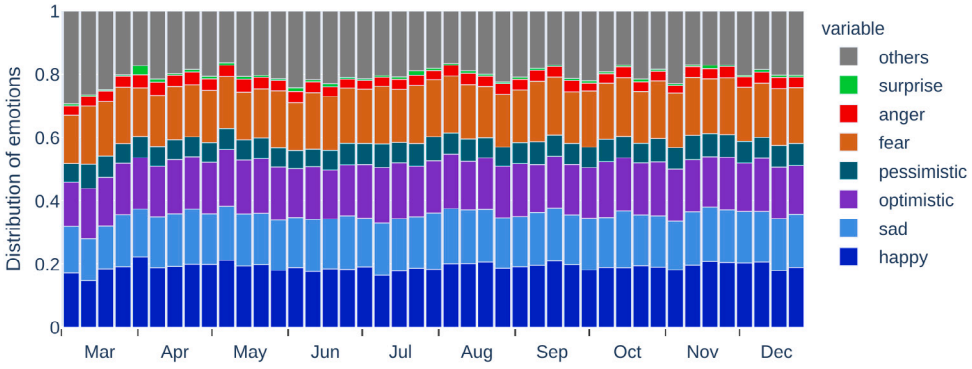


Fig. 4. Weekly emotion distribution in the USA.

emotions are as follows: Happy = 0.1714, sad = 0.1477, optimistic = 0.1394, pessimistic = 0.0589, fear = 0.1537, anger = 0.0298, surprise = 0.0092, and others = 0.2899. We conduct further emotion analysis on the six dominant emotions that we have stated earlier. Fig. 5 provides a better idea of weekly emotion distribution. It shows the variation in the emotions in each week. We use the exact emotion range in the Y-axis without scaling. This allows us to recognize the dominant emotions in the tweets. For example, the Y-axis values of pessimistic and anger charts denote that the number of tweets with those emotions is lower than other emotions.

While Fig. 5 represents the emotional roller-coaster in the USA, Fig. 6 depicts a better picture of emotion evolution during the pandemic using monthly emotion distribution. We can observe a similar emotion range in monthly and weekly charts. We present some of the critical events during the pandemic in Fig. 7 to correlate the emotions. This also allows us to observe the accuracy of the models with respect to historical events. In Fig. 7, the events are ordered in a way such that, closer events to the timeline occurred earlier in the respective month. From the figure, we can see that in mid-February US stock market crashed from the fear of COVID-19. By the end of February US reported the first COVID related death.

From the emotion chart in Fig. 6 we observe the high range of fear and pessimism at the beginning of March as people became aware of the situation. In March WHO declares COVID-19 as a pandemic and a national emergency also announced in the USA. By the end of March, the death count became 2000 in the US and the total number of cases surpassed 102K+. However, stimulus bills were also signed in March and people started to receive their first stimulus check in April. There was also a lack of proper guidance regarding the pandemic and many people thought COVID-19 is only harming the adult people severely. Because of this, the fear is reduced and people became optimistic in April. However, people were still sad and disappointed by the pandemic and economic situation. We can see a sharp rise in anger in April. In April, the death count increased rapidly and president Trump suggested disinfectants can be helpful for COVID treatment which surges the anger among the people. Until May, most of the COVID cases in the USA were came from New York. However, by the end of May, COVID cases and hospitalization started to spike in other states which triggers negative emotions all over the USA. This reflects in Fig. 6 as we can see fear and pessimism rise sharply from June. In August, the daily reported new cases declined and because of that, we see a drop in the fear. People were scared again after August as the second wave of COVID infection started and the daily new cases started to break the previous record regularly. By September 200K people died in the USA and a total of 1M people died worldwide because of COVID. In October several reports were published about positive vaccine trials which gave optimism to the people. In November, US reported 100K+ news cases in a single day. People started to lost hope and both anger and pessimism started to rise. In December people started to gain confidence because of the vaccine roll-out. However, the USA experienced a record single-day death. Furthermore, several reports stated the 20M vaccination goal of the USA might not be fulfilled in 2020. All those events trigger mixed reactions but we can observe an increasing amount of anger.

Fig. 5. Weekly emotion variation in the USA (March 2020–December 2020). Y axis represents the weekly emotion range on a scale of 0–1 combining all emotions.

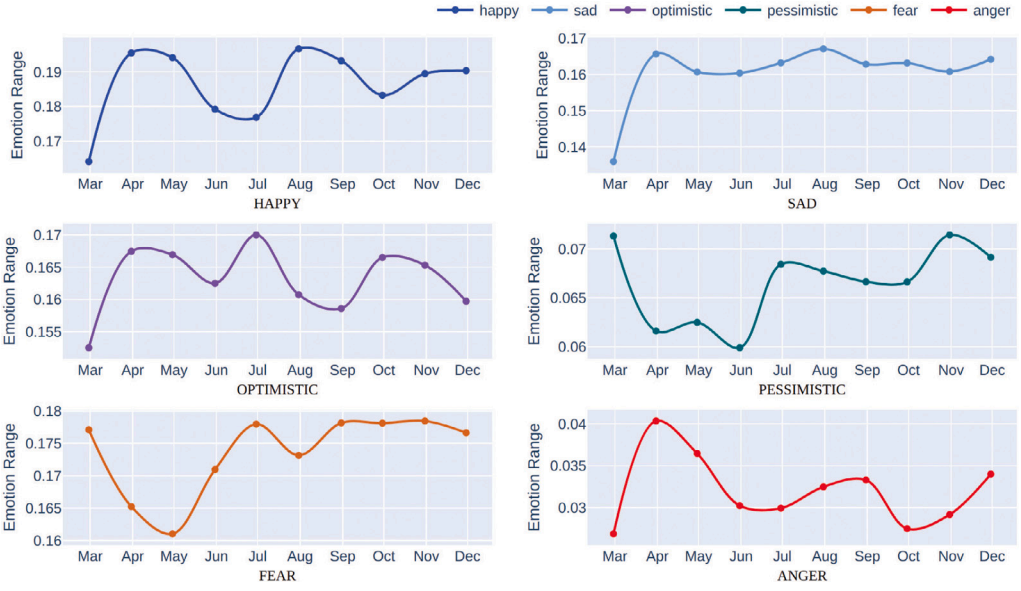


Fig. 6. Monthly emotion variation in the USA (March 2020–December 2020).

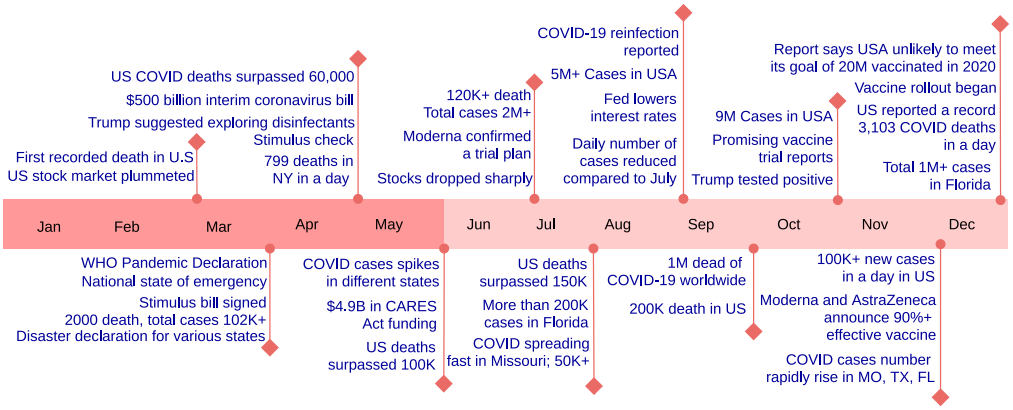


Fig. 7. Timeline of Events Related to the COVID-19 Pandemic in the USA.

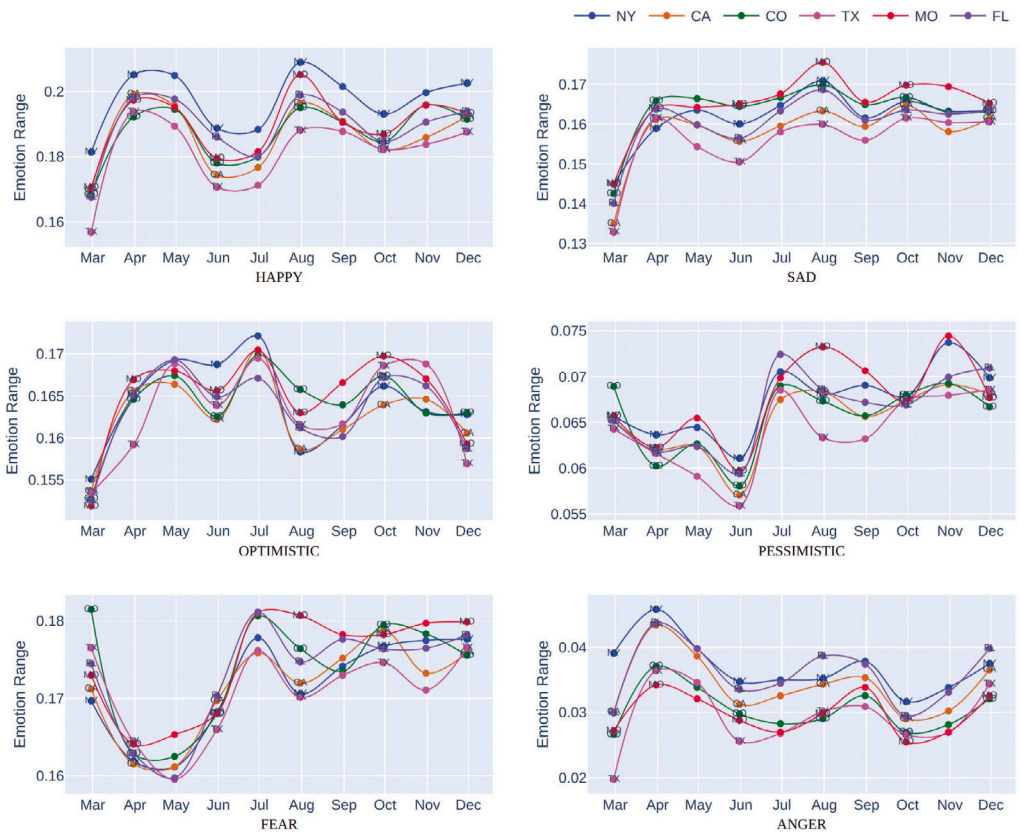


Fig. 8. Monthly emotion variations in 6 states of USA (March–December 2020).

The monthly emotions variations for six different states (NY, CA, CO, TX, MO, FL) are depicted in Fig. 8. We can observe a similarity in emotion timeline across the USA and the states. While most of the states have similar emotional trends, we can observe some significant variations at some points. For instance, we can observe a higher amount of negative emotions such as fear, pessimism, and sadness in Missouri (MO) and Florida (FL) during July, August, and September. MO exhibits a higher amount of fear, pessimism, and sadness in August compared to other states. If we look back at the timeline of the events in Fig. 7, we see that in July COVID-19 cases spiked in MO and FL and it was spreading fast. This correlates with the higher negative emotion as we can see in the chart. In November and December, the new cases again started to rise rapidly in MO and FL which make people scared and sad. As a result, we can see those states showing high fear and pessimism. We can see during NOV–DEC, MO is showing the highest fear among the six states and FL is showing maximum anger. From the charts and COVID-19 events timeline, we can state that the classification model performed satisfactorily to identify the emotions during the pandemic.

6. Conclusion and future work

In this work, we proposed two machine learning models for multi-label binary classification and phrase extraction applied on a unique emotion dataset on COVID-19 tweets for classifying 10 different emotion labels, and to select a phrase that represents each emotion the most. This paper also presents a comparative performance evaluation and analysis of the proposed models. Our developed models outperformed other systems under different performance metrics. We use a set of auxiliary features that improve the performance of the classifier. For phrase extraction, we use RoBERTa pre-trained model with a custom Q&A head which takes the emotion label as a question and tries to find a phrase that can best be suited for that emotion. The output analysis of the model shows the robustness to understand the context of a given tweet. Further, we perform a historical emotion analysis over some of the states in the USA using the COVID-19 tweets. The analysis shows how the negative emotions increased during the pandemic. It also shows how people were adapting to the pandemic over time, and being more optimistic. In the future, we will integrate our models in our live application to continue the emotion analysis during the pandemic over the entire USA. We will also analyze phrase extraction model output over the historical COVID-19 tweets and incorporate those in the live application. We will keep exploring the different ideas on phrase extraction for emotion context in the future to improve our results further. We plan to use data augmentation and transfer learning to train our model so that it can perform robustly with effectiveness. We will share our data publicly for the different research communities on Github.

CRediT authorship contribution statement

Md. Yasin Kabir: Methodology, Formal analysis, Software, Writing - review & editing, Investigation. **Sanjay Madria:** Conceptualization, Data curation, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research has been partially supported by a grant from NSF, USA, CNS-1461914.

References

[1] X. Chen, Y. Cho, S.Y. Jang, Crime prediction using twitter sentiment and weather, in: 2015 Systems and Information Engineering Design Symposium, IEEE, 2015, pp. 63–68.

[2] M.S. Gerber, Predicting crime using Twitter and kernel density estimation, Decis. Support Syst. 61 (2014) 115–125.

[3] P. Grover, A.K. Kar, Y.K. Dwivedi, M. Janssen, Polarization and acculturation in US Election 2016 outcomes–Can twitter analytics predict changes in voting preferences, Technol. Forecast. Soc. Change 145 (2019) 438–460.

[4] M.Y. Kabir, S. Madria, A deep learning approach for tweet classification and rescue scheduling for effective disaster management, in: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2019, pp. 269–278.

[5] M. Kabir, S. Madria, et al., Coronavis: A real-time COVID-19 tweets analyzer, 2020, arXiv preprint arXiv:2004.13932.

[6] A.F. Anees, A. Shaikh, A. Shaikh, S. Shaikh, Survey paper on sentiment analysis: Techniques and challenges, 2020, EasyChair2516-2314.

[7] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 8 (4) (2018) e1253.

[8] X. Wang, W. Jiang, Z. Luo, Combination of convolutional and recurrent neural network for sentiment analysis of short texts, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2428–2437.

[9] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: LREc, vol. 10, 2010, pp. 1320–1326.

[10] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation, 2018, pp. 1–17.

[11] Q. Yang, H. Alamro, S. Albaradei, A. Salhi, X. Lv, C. Ma, M. Alshehri, I. Jaber, F. Tifratene, W. Wang, et al., SenWave: Monitoring the global sentiments under the COVID-19 pandemic, 2020, arXiv preprint arXiv:2006.10842.

[12] J. Xue, J. Chen, C. Chen, C. Zheng, T. Zhu, Machine learning on big data from Twitter to understand public reactions to COVID-19, 2020, arXiv preprint arXiv:2005.08817.

[13] C. Ziems, B. He, S. Soni, S. Kumar, Racism is a virus: Anti-Asian hate and counterhate in social media during the COVID-19 crisis, 2020, arXiv preprint arXiv:2005.12423.

[14] E. Chen, K. Lerman, E. Ferrara, Covid-19: The first public coronavirus twitter dataset, 2020, arXiv preprint arXiv:2003.07372.

[15] J.M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, G. Chowell, A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration, 2020, arXiv preprint arXiv:2004.03688.

[16] P. Martí, L. Serrano-Estrada, A. Nolasco-Cirugeda, Social media data: Challenges, opportunities and limitations in urban studies, Comput. Environ. Urban Syst. 74 (2019) 161–174.

[17] P.R. Spence, K.A. Lachlan, A.M. Rainear, Social media and crisis research: Data collection and directions, Comput. Hum. Behav. 54 (2016) 667–672.

[18] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, in: Coling 2010: Posters, 2010, pp. 36–44.

[19] R. Nagar, Q. Yuan, C.C. Freifeld, M. Santillana, A. Nojima, R. Chunara, J.S. Brownstein, A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives, J. Med. Internet Res. 16 (10) (2014) e236.

[20] M. Dredze, D.A. Broniatowski, K.M. Hilyard, Zika vaccine misconceptions: A social media analysis, Vaccine 34 (30) (2016) 3441.

[21] M.Y. Kabir, S. Gruzdev, S. Madria, STIMULATE: A system for real-time information acquisition and learning for disaster management, in: 2020 21st IEEE International Conference on Mobile Data Management (MDM), IEEE, 2020, pp. 186–193.

[22] L. Zou, N.S. Lam, S. Shams, H. Cai, M.A. Meyer, S. Yang, K. Lee, S.-J. Park, M.A. Reams, Social and geographical disparities in Twitter use during hurricane harvey, Int. J. Digit. Earth 12 (11) (2019) 1300–1318.

[23] Z. Yang, L.H. Nguyen, J. Stuve, G. Cao, F. Jin, Harvey flooding rescue in social media, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 2177–2185.

[24] E. Hirata, M. Giannotti, A. Larocca, J. Quintanilha, Flooding and inundation collaborative mapping–use of the Crowdmap/Ushahidi platform in the city of Sao Paulo, Brazil, J. Flood Risk Manag. 11 (2018) S98–S109.

[25] C. Buntain, J. Golbeck, B. Liu, G. LaFree, Evaluating public response to the Boston Marathon bombing and other acts of terrorism through Twitter, in: Tenth International AAAI Conference on Web and Social Media, 2016.

[26] B.G. Southwell, J. Niederdeppe, J.N. Cappella, A. Gaysynsky, D.E. Kelley, A. Oh, E.B. Peterson, W.-Y.S. Chou, Misinformation as a misunderstood challenge to public health, Am. J. Prevent. Med. 57 (2) (2019) 282–285.

[27] S.O. Oyeyemi, E. Gabarron, R. Wynn, Ebola, Twitter, and misinformation: a dangerous combination?, Bmj 349 (2014) g6178.

[28] Z. Wang, N.S. Lam, N. Obradovich, X. Ye, Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data, Appl. Geogr. 108 (2019) 1–8.

[29] D. Wladdimiro, P. Gonzalez-Cantergiani, N. Hidalgo, E. Rosas, Disaster management platform to support real-time analytics, in: 2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), IEEE, 2016, pp. 1–8.

[30] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, Y. Wang, A first look at COVID-19 information and misinformation sharing on Twitter, 2020, arXiv preprint arXiv:2003.13907.

[31] R. Kouzy, J. Abi Jaoude, A. Kraitem, M.B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E.W. Akl, K. Baddour, Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter, Cureus 12 (3) (2020).

[32] C. Ordun, S. Purushotham, E. Raff, Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs, 2020, arXiv preprint arXiv:2005.03082.

[33] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, Z. Shah, Top concerns of tweeters during the COVID-19 pandemic: infoveillance study, J. Med. Internet Res. 22 (4) (2020) e19016.

[34] M. De Choudhury, M. Gamon, S. Counts, Happy, nervous or surprised? classification of human affective states in social media, in: Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[35] M. Jabreel, A.M. Ribas, SiTAKA at SemEval-2017 Task 4: Sentiment analysis in Twitter based on a rich set of features, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 694–699.

[36] M. Jabreel, A. Moreno, SentiRich: Sentiment analysis of tweets based on a rich set of features, in: CCIA, 2016, pp. 137–146.

[37] C. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, A. Potamianos, Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning, 2018, arXiv preprint arXiv:1804.06658.

[38] H. Meisheri, L. Dey, TCS research at SemEval-2018 Task 1: Learning robust representations using multi-attention architecture, in: Proceedings of the 12th International Workshop on Semantic Evaluation, 2018, pp. 291–299.

[39] J.H. Park, P. Xu, P. Fung, Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and# hashtags, 2018, arXiv preprint arXiv:1804.08280.

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[41] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[42] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[43] S. Wang, J. Jiang, Learning natural language inference with LSTM, 2015, arXiv preprint arXiv:1512.08849.

[44] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6645–6649.

[45] A. Kumar, S.R. Sangwan, A. Arora, A. Nayyar, M. Abdel-Basset, et al., Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network, IEEE Access 7 (2019) 23319–23328.

[46] C. Alberti, K. Lee, M. Collins, A bert baseline for the natural questions, 2019, arXiv preprint arXiv:1901.08634.