Decoding US Sentiments on Climate Change

NUS MSBA BT5153 Spring 2020 – Group 3 Project Report

Link to code & data: https://github.com/calvinlian/bt5153_project

Abstract

Climate change is an important and urgent challenge that affects the entire human race. In this project, we aim to measure the sentiment of a population using Twitter posts related to climate change. We first train several machine learning models using a labelled Twitter dataset from Kaggle. Using a scoring table, we compare the models and find the best one to be the CNN model. We then predict the climate change sentiments of different states in the United States using the CNN. Finally, the predicted output is visualized using a choropleth and several insights are drawn. We also discuss some of the shortcomings of our model. We conclude that our solution is a quick and cost-effective product for measuring sentiments with the potential to be used on other social media as well.

1. Introduction

Climate change is mankind's greatest existential challenge in the 21st century. Impacts of rapid climate change (NASA, 2019) are increasingly experienced globally with frequent extreme weather phenomenon, rising sea levels (Sea Level Rise Work Group, 2015), droughts, heatwaves, wildfires, and storms. Climate change not only menaces cities physically but their economies as well.

World leaders, private sectors, and civil societies are coming together to combat climate change, such as the 2019 Climate Submit Action (United Nations, N.D.). Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, 2012). Sentiment trackers for climate change can help government agencies like EPA (U.S. Environmental Protection Agency) and non-government organizations like C2ES (Centre for Climate & Energy Solutions) and CAT (Climate Action Tracker) evaluate public opinions on climate change and identify locations where understanding of climate change is lacking. The sentiment tracking methodology could potentially be transferred to other social subjects.

Twitter is a platform which allows users to post and communicate tweets (short text messages) of up to 280 characters, socially expressing their opinions and experiences on a specific subject. Twitter is thus an ideal platform to capture data for sentiment analysis.

Our hypothesis is that climate change awareness is growing and if EPA, C2ES, or CAT can **identify areas where climate change is disregarded** (indicated by a low and/or decreasing sentiment score), they can either:

- Steer resources accordingly to educate the people,

or

- Tune policies to subtly tackle the economic and environmental impacts of climate change such that it doesn't bring about too much resistance from the public.

This hypothesis is corroborated by economist Mark Jaccard, who opines that "scientists and activists don't have to convince everyone of the seriousness of the threat – they just have to motivate 'climate-sincere' policymakers to instate new regulations" (Jaccard, 2020). To track how successful their policies are, they can also track each area's sentiment change on climate change over time.

2. Proposed Solution

Our project aims to use machine learning models to measure public opinions on climate change in the United States (US). To achieve this, we have built a Climate Change Sentiment Tracker (CCST) based on recent tweets about the subject throughout the US and use it to infer the recent trends of public opinions across different US states. The graph below represents the model process flow.

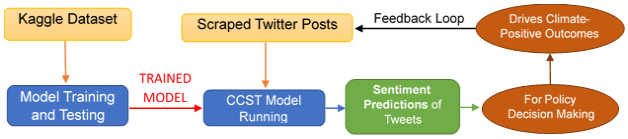


Figure 1: Model Process Flow Chart

We have used labelled Twitter data relating to climate change, obtained from Kaggle, to train and test the machine learning model(s). The model(s) have then been fitted with scraped Twitter posts to predict the average sentiment on climate change per US state. This state-level geographical sentiment would then be passed to key policymakers for decision-making and resource deployment.

3. Data Description

In our study, 2 separate datasets were sourced for model training and sentiment prediction: (1) sentiment-labelled tweet dataset from Kaggle, and (2) scraped Twitter posts.

3.1 Dataset for Model Training

The Kaggle dataset used for model training and tuning, contains 43,943 annotated tweets relating to *climate change* collected between Apr 27, 2015 and Feb 21, 2018. Each tweet is labelled as one of the following classes:

- **2 (News)**: the tweet links factual news about climate change
- **1 (Pro)**: the tweet supports the belief of man-made climate change
- **0 (Neutral)**: the tweet neither supports nor refutes the belief of man-made climate change
- **-1 (Anti)**: the tweet does not believe in man-made climate change

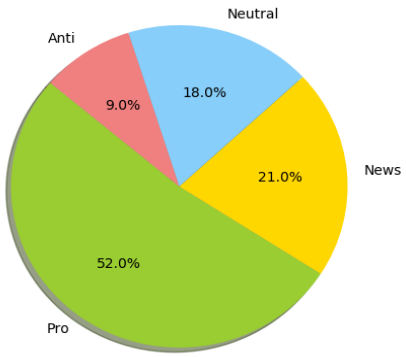


Figure 2: Distribution of Tweets Labels

It was noticed that the dataset is highly imbalanced as 52% of messaged express positive sentiment (class 1) and only 9% of messages express negative sentiment (class -1). This would likely introduce a bias to predict class 1 more than other classes and influence accuracy of the classification model. To minimize this, we have created 2 datasets: the full dataset with all messages, and an under-sampled dataset with equal number of messages for each class. This was achieved by reducing classes 0, 1, and 2 to match the minority class (-1) through random sampling. Both datasets were used in model training. More details on our data resampling strategy will be discussed below in the model training section.

The table below presents the example of messages per class (sentiment).

Table 1: Example of Sentiment-labelled Twitter Data

Senti-ment	Message	tweetid
-1	@tiniebeany climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom	792,927,353,886,371,000
1	Fabulous! Leonardo #DiCaprio's film on #climate change is brilliant!!! Do watch. https://t.co/7rV6BrmxjW via @youtube	793,124,402,388,832,000
2	RT @cnalive: Pranita Biswasi, a Lutheran from Odisha, gives testimony on effects of climate change & natural disasters on the poÃƒÂ¢Ã¢â€šÂ¬Ã‚Â¦	793,125,156,185,137,000
0	@chefBOYERdee1 Al Gore, creator of the internet and global warming.	793,125,429,418,815,000

3.2 Forecast Sentiment of Twitter Data

Real-time Twitter data were scraped using python library GetOldTweets3, which is able to search and parse tweets by specifying the keywords, dates, and number of tweets. However, while looping through the actual tweets, some tweets were unable to be cleaned using the built-in text-cleaning function by GetOldTweets3 and produced an error. To overcome this, the library was edited to bypass the text-cleaning function so that the loops are able to run without interruptions. The messages were then manually cleaned after extraction. Details of cleaning can be found below in Model Creation.

4. Model Creation

4.1 Text Processing

In order for the trained model to also be applicable to real-time scraped twitter, we cleaned the training data and scraped twitter messages so that only English tweets remains. The text processing steps on Kaggle dataset consists of the following steps:

1. Remove duplicate entries (2910 duplicated rows)
2. Removing non-alphanumeric strings
3. HTML tags
4. URL links
5. Applying lower-case for all words
6. Removing excess spaces
7. Lemmatization
8. Removal of non-English tweets

9. The stop-words have been kept in the corpus as they helped better distinguish negative and positive tweets.

We created a **word cloud (Figure 3)** of the corpus to provide an overview and to visualize the recurring keywords for Twitter posts in *Kaggle* dataset. The most frequent words that have appeared in tweets on climate change were: "climate change", "global warming", "scientist", "trump", "people", "change real", and "new".

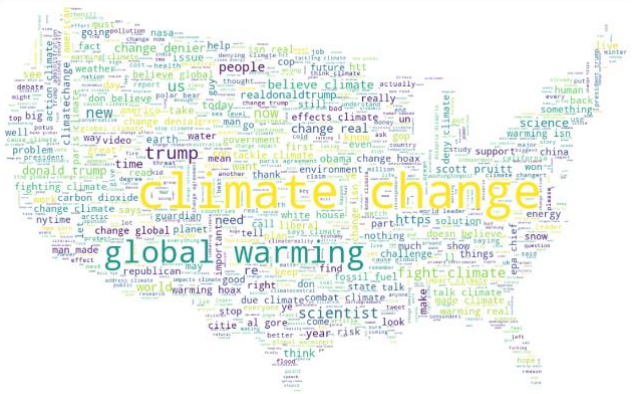


Figure 3: Word cloud visualization of Kaggle training dataset

4.2 Feature Engineering

Then, the **Bag of Words** approach has been applied to transform the text into vectors (features) that can be fitted to different machine learning classification models. The texts were processed into vectors using *sklearn CountVectorizer and TfidfVectorizer*. Both inputs were used in model training for comparison.

Additionally, for the CNN model, fastText word embedding pre-trained from Wikipedia 2017, UMBC webbase corpus and statmt.org news data have been used. The **pre-trained word embedding** was selected over a **randomly-initialized word embedding** to improve model robustness when it encounters words unseen in the training dataset.

We have also explored additional features like *tweet length* and *number of punctuation signs*. However, we found that these new features do not have a significant differentiating factor between classes. Hence, these additional features were not included as model inputs. The outcome of these experiments can be found in Appendix 2.

4.3 Model Validation Strategy

Since the labels are imbalanced, we decided to use a stratified split when selecting our training and validation datasets.

To compare model performance, we selected K-fold cross validation, which permits the use of all training data and measure model capability to generalize. K=5 folds struck a balance between generalizability and training time.

4.4 Model Performance Metric

As the labels are imbalanced, accuracy would give a poor indication of model performance: a model could naïvely predict the majority class to obtain a high accuracy.

Therefore, we selected the **weighted F1-score** as our metric, which calculates the mean of harmonic means of precision and recall for each class, weighted by the number of observations within that class.

4.5 Machine Learning Models

4.5.1 Model Selection

Classification (supervised learning) models are widely used in text analysis, especially in sentiment analysis, where it allows us to identify binary (positive, negative) as well as multiclass (positive, negative, and neutral) sentiments. This project explores **4 classes of labels**: positive, negative, neutral, and factual.

10 models were selected and initially trained. Among these, some classification models we trained were:

- K-Nearest Neighbors
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

However, the above models' predictive powers were found to be significantly poorer than the following models:

- Logistic Regression
- Naïve Bayes
- Linear SVM
- Light GBM
- CNN

Thus, the poorly-performing models were excluded from hyper-parameters optimization. A summary of all models' initial training can be found in Appendix 1.

Logistic Regression, **Naïve Bayes** and **Support Vector Machine (SVM)** gave the best initial results with weighted F1-scores among non-neural network models.

LightGBM, a gradient boosting framework using tree-based learning model was also explored. LightGBM can support parallel and GPU learning, is capable of handling large-scale dataset and was selected over XGBoost as it has faster training speed and lower memory usage.

Finally, the dataset has been trained and tested with **Convolutional Neural Network (CNN)**, known for its capability to reduce model parameters and context capturing.

4.5.2 MODEL HYPER-PARAMETER OPTIMIZATION

Table 2: Summary of ML Models (Before vs after tuning)

Best Final Text Hyper- params	Best Model Final Hyper- params	Sampling Type	Default hyper-params F1-score (accuracy)	Final hyper-params F1-score (accuracy)
CountVectorizer: ngram_range=(1,2)	**Logistic Regression**: *C=10, max_iter=100*	Under	0.65 (0.65)	0.67 (0.67)
	Logistic Regression: *C=10, max_iter=300*	Full	0.67 (0.68)	0.69 (0.69)
CountVectorizer: ngram_range=(1,2)	**NaiveBayes:** *Alpha = 0.2, Class_prior=None Fit_prior=False*	Under	0.62 (0.63)	0.63 (0.65)
	NaiveBayes: *Alpha = 0.1 Class_prior=None Fit_prior=False*	Full	0.68 (0.70)	0.64 (0.72)
CountVectorizer: ngram_range=(1,2)	**Linear SVM:** *Alpha = 0.0001 Loss = modifier_huber Max_iter = 10*	Under	0.64 (0.65)	0.65 (0.66)
	Linear SVM: *Alpha = 0.00001 Loss = modifier_huber Max_iter = 2000*	Full	**0.70** (0.71)	0.68 (0.74)
TfidfVectorizer: *ngram_range: (1, 1) norm: 'l1', use_idf: False*	**LightGBM:** *max_depth: 20 min_data_in_leaf: 10 num_leaves: 30*	Under	0.58 (0.59)	0.62 (0.62)
	LightGBM: *max_depth: 20 min_data_in_leaf: 10 num_leaves: 50*	Full	0.62 (0.65)	0.67 (0.69)
fastText pre-trained wiki-news-300d-1M-subword	**CNN – 2 conv layers** *Embedding weights: trainable Batch size: 256 Num_filters: 100 Kernel_size: 5*	Under	0.61 (0.81)	0.64 (0.83)
	CNN – 2 conv layers *Embedding weights: trainable Batch size: 256 Num_filters: 300 Kernel_size: 5*	Full	0.67 **(0.83)**	**0.70 (0.85)**

The table above summarizes the *F1-score* and *Accuracy* of selected models, computed with *Stratified 5-fold cross validation* method, before and after hyper-parameters tuning.

From this table, we can conclude that under-sampling our dataset hurts our model predictive power across all models. This is likely because reducing the limited no. of training observations (44k to 12k) results in under-fitting.

From this table, we also observe that CNN with the full dataset stands out in terms of both F1-score and accuracy compared to the rest.

4.6 Comparison and Final Model Selection

Aside from predictive accuracy/F1-score, several other factors were considered in our final model selection process.

4.6.1 MODEL ACCURACY/F1-SCORE

Although the best result has been achieved with CNN model (F1 score at 0.7 with full dataset), Linear SVM as well as Logistic Regression have also obtained very good results (F1-score at 0.7 and 0.69 respectively).

4.6.2 ROBUSTNESS

In our context, our models were trained on a Kaggle dataset scraped within a limited period. This being social media, slangs and vocabulary vary wildly across regions and time. *Bag of Words* models are not able to generalize well with unseen words.

The CNN word embedding's ability to manage new vocabulary not from the training vocabulary, but within the pre-trained embedding, is superior to the *Bag of Words* approach used.

4.6.3 TRAINING SPEED

Naïve Bayes was the model that was fastest to train. Followed by Logistic Regression and LightGBM. CNN's high computational cost proved to be its weakness. However, it can be partially mitigated through parallelization with GPU/TPUs.

In our case, our CNN trained on a consumer-grade GPU had comparable training times (in the order of minutes) to Random Forest, XGBoost and LightGBM trained on a CPU.

4.6.4 PREDICTION SPEED

From our scraped tweets, we found that the month with the most tweets was in Feb 2020 at ~18000 number of tweets. This translates to a daily average of 600 tweets. This data velocity can be easily managed by any stream-processing machine learning model. Thus, prediction speed of our models would not be a limiting factor and not considered.

4.6.5 MODEL SCORING TABLE AND FINAL SELECTION

To compare the overall effectiveness of different models, we used a scoring table. Each model was assigned a score of their relative ranking compared to the others in that metric.

For robustness, *Bag of Words* models were ranked based on their relative standard deviations of 5-fold CV F1-scores, as a lower standard deviation means that the model is able to better generalize.

The model with the *highest total score* was our best model, which appeared to be the **CNN.**

Table 3: Scoring Table of Different Models

	Log Reg	Naïve Bayes	Linear SVM	Light GBM	CNN
Accuracy	4	1	3	2	5
Robustness	2	1	3	4	5
Training Speed	4	5	2	3	1
Overall	10	7	8	9	**11**

5. Model Output and Insights

The scraped Twitter posts from 2011 to 2020 was passed through the same text pre-processing pipeline and then passed in the CNN model to generate sentiment predictions. We used the predicted labels of tweets from Mar 2020 to demonstrate our model utility in the following section.

Data visualization can be a powerful tool for presentation of findings. Good visualizations can make our findings easier to interpret and captivating at the same time. Since our target audience are key policy-makers, we chose to use an intuitive visualization in the form of a choropleth map to present a geographical snapshot of sentiment towards climate change.

We have summarized sentiment results from twitter throughout US with geographical information to construct **interactive** choropleth maps using *plotly* library in Python. The maps have allowed us to quickly identify US states with low or high degree of the seriousness of climate change.

To improve our model output interpretability, we explored the visualization of several forms of ***Sentiment Index*** that are intuitive to understand. Several forms of these ***Sentiment Index*** are discussed below.

5.1 Sum of Sentiments

By labelling negative, neutral and positive tweets as -1, 0 and 1 respectively, we can sum them up and use these to calculate a Sentiment Index per state.

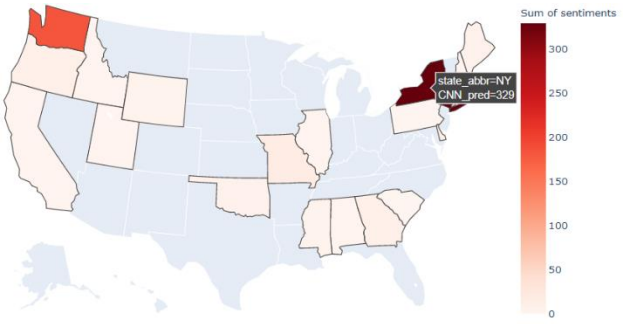


Figure 4: Sum of Sentiments per State

The choropleth map with the US sentiment from the tweets scraped from March 2020. It presents the **sum of climate change sentiments values** per state. It can be noticed that the subject has not been tackled in all states. The positive values indicate that there were more positive tweets in each state than negative ones. Washington and New York State shows a strong positive sentiment towards climate change.

To correlate this with climate-change-related events: In March 2020, there was a global climate strike in New York State, this could explain the abnormally high number of positive tweets from that location.

However, it should be noted that this graph only represents the total sum of sentiment per state and is not normalized to the total number of tweets. Thus, we also considered several other indices to measure sentiment.

5.2 Ratio of Positive vs All Tweets

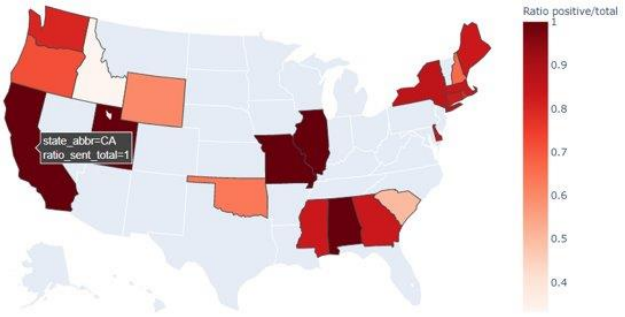


Figure 5: Ratio of Positive vs All

If we look at the ratio between number of positive tweets and total number of tweets, the picture become more complex. Some states show the ratio at 1 meaning that all tweets posted in March 2020 has positive sentiment towards climate change. We can also observe the states with the ratio at 0.3, indicating neutral and negative sentiments being present as well in important proportion.

This form of visualization has greater distinction between states of positive and negative sentiments. From the map, states like California, Utah, Missouri, Illinois and Alabama would likely be more receptive to climate change policies. For more neutral or negative states like Idaho, such policies would likely need to be pitched differently. For example, encouraging adoption of electric vehicles by stating its technological trendiness, or economic savings over fueled vehicles.

5.3 Ratio of Positive vs Negative Tweets

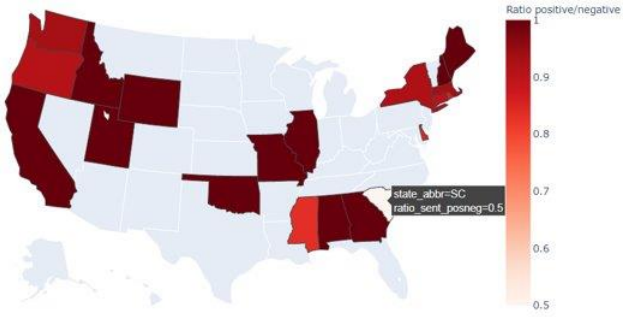


Figure 6: Ratio of Positive vs Negative

Finally, the third choropleth (Figure 6) presents the ratio between number of positive tweets and the sum of positive and negative tweets. Most of the state shows ratio at 1 or very close to 1, indicating that total/ majority of tweets were positive. However, we can see that in South Carolina the ratio is at 0.5, suggesting an important presence of negative sentiment towards climate change. South Carolina is identified among a string of "Deep South" states that will experience the worst effects of climate change. The Climate Change Sentiment tracker could be a good tool for administration to measure potential adverse reactions towards climate change policies in the states like South Carolina.

The Climate Change Sentiment tracker also permit to measure the historical evolution of climate change sentiment. [Appendix3]

6. Potential Pitfalls Related to Twitter Data

The following weaknesses of the approach of using twitter data can be identified:

6.1 Sample may not be representative of population: In US only around 22% of population uses Twitter and 10% of those is responsible for 80% of the tweets. Furthermore, American Twitter users tend to be young and 42% of them have at least bachelor's degree (Pew Research Center, 2019). Furthermore, not everyone shares their opinions online. Those who do tends to be activists who are often extreme in their views. Studies on social media biases have shown the biases can differ geographically (Saez-Trumper et al., 2013).

Addressing the pitfall: An unavoidable degree of sampling bias exists in all surveys and research. In our case, we feel that Twitter data can quickly gauge a population's reception towards new climate policies. In addition, the text medium of Twitter encourages the open sharing of opinions when compared to other forms of social media that have focus on visual media. Despite the listed drawbacks, twitter sentiment analysis remains a powerful tool that can be used by government to gauge public opinion to policy announcements and businesses to see what people think about their product.

6.2 Methodology is susceptible to manipulation: Bad actors might seek to manipulate the sentiment using automated '*bot'* accounts to flood the twitter-sphere.

Addressing the pitfall: Common manipulation can be addressed by duplicate removal. Also, Twitter itself also has safeguards in place to prevent bots manipulation. (Timberg et al., 2018)

7. Conclusion

We have developed a potential solution for measuring and tracking climate change sentiments in the United States based on real-time twitter data. Although there have been surveys conducted to gather such sentiments from the public on climate change, our project offers an innovative approach using free open-source software (Python) that can provide an end-to-end solution -- From data gathering

using twitter-scraping scripts, to training and deploying a sentiment prediction based on tweet messages, to finally visualizing the sentiments on an interactive map.

The use of social media as an informal channel of feedback is a step-up from the traditional surveys. It is a more economical and quicker way of gathering public opinions. Although our project focused on Twitter due to its mainly textual data, the methodology can also be applied to text extracted from Facebook or YouTube.

Other than providing a tool for governments, environmentalists, and businesses to gain insights on the climate change sentiments at each U.S. state, the tracking of such sentiments can also serve as feedback on effectiveness and receptivity of policies or products. Public reactions to natural disasters caused by climate change can also be tracked using our CCST model to serve as early detections on where to deploy aids and resources.

Finally, the interactive map used to visualize the sentiments can be easily exported as HTML scripts to be integrated into existing websites without the need to install additional applications.

8. References

National Aeronautics and Space Administration (NASA) (2019) Climate Change Evidence: How Do We Know? (2019, December 30). Retrieved April 21, 2020, from https://climate.nasa.gov/evidence/

Sea Level Rise Work Group (2015) October 2015. Unified Sea Level Rise Projection for Southeast Florida. A document prepared for the Southeast Florida Regional Climate Change Compact Steering Committee. 35 p.

United Nations (N.D.) Climate Change https://www.un.org/en/sections/issues-depth/climate-change/

Liu, B. (2012) Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Jaccard, M. (2020) The Citizen's Guide to Climate Success. Cambridge University Press. ISBN 9781108783453

Qian, E (2019) Twitter Climate Change Sentiment Dataset Retrieved April 21, 2020 from https://www.kaggle.com/edqian/twitter-climate-change-sentiment-dataset/data

Pew Research Center (2019) Pew: US adult Twitter users tend to be younger, more Democratic; 10% create 80% of tweets https://techcrunch.com/2019/04/24/pew-u-s-adult-twitter-users-tend-to-be-younger-more-democratic-10-create-80-of-tweets/

Saez-Trumper, Diego and Castillo, Carlos and Lalmas, Mounia (2013) *Social Media News Communities: Gatekeeping, Coverage, and Statement Bias*. Association for Computing Machinery. ISBN 9781450322638 https://www.kaggle.com/edqian/twitter-climate-change-sentiment-dataset/data

Timberg, C, Dwoskin, E. (2018) Twitter is sweeping out fake accounts like never before, putting user growth at risk, 7 July 2018. Last retrieved on 22 April 2020 from https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/

AN6U5 (2016) Plotly in Python. Last retrieved 23 Apr 2020 from https://datascience.stackexchange.com/questions/9616/how-to-create-us-state-heatmap

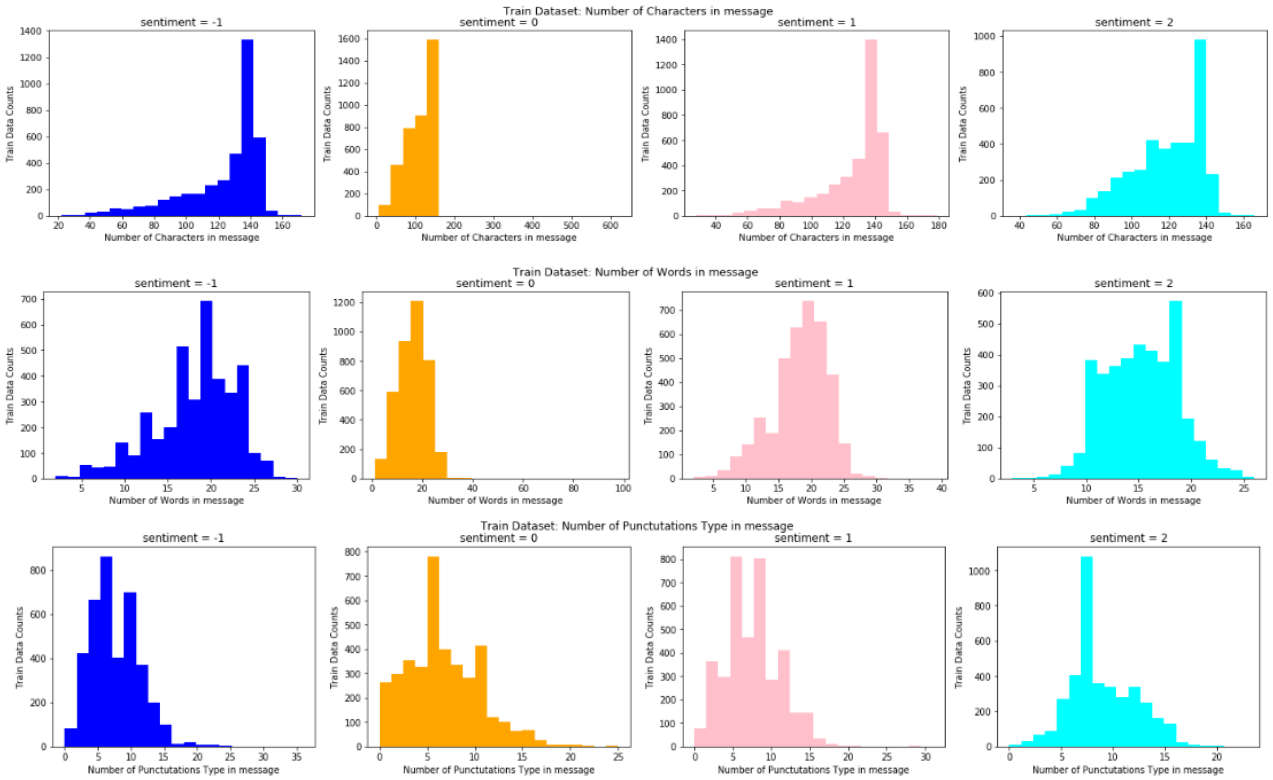
Mohammad, S (2016) Emotion, Sentiment, and Stance Labeled Data, Las retrieved on 23 Apr 2020 from http://saifmohammad.com/WebPages/SentimentEmotionLabeledData.html

9. Appendix

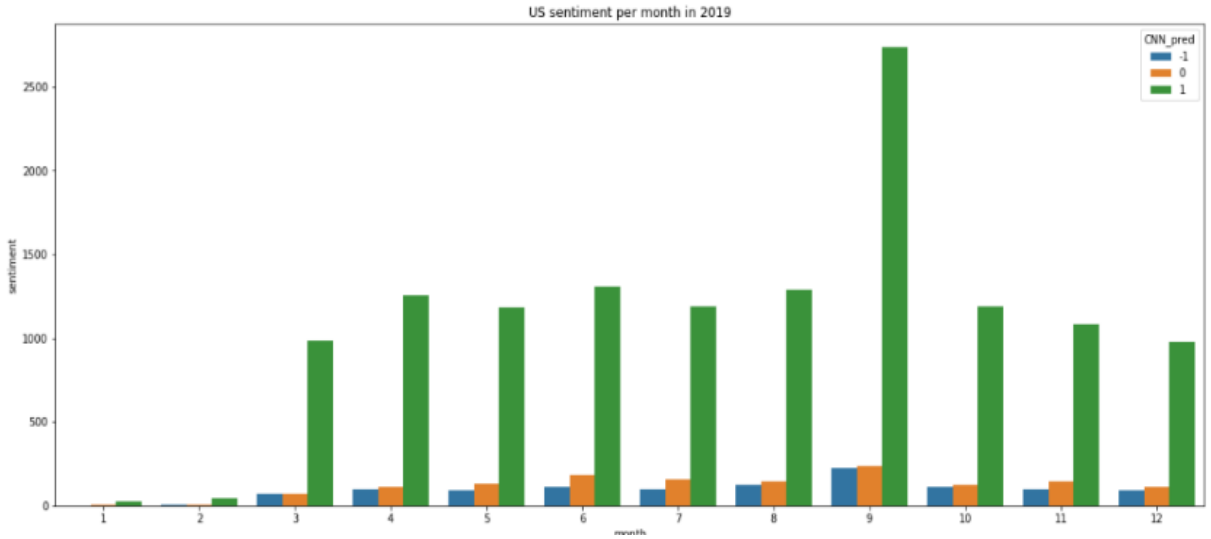
9.1 Appendix 1

Model	Mean_f1 weighted	Std_f1 weighted	Mean accuracy	Std accuracy
LogisticRegression	0.6536	0.0053	0.6542	0.0052
KNN	0.3165	0.0047	0.3492	0.0069
NaiveBayes	0.6279	0.0035	0.6405	0.0029
DecisionTree	0.4909	0.0058	0.4912	0.0059
SVM	0.6240	0.0050	0.6286	0.0052
RandomForest	0.5970	0.0046	0.6007	0.0042
AdaBoost	0.5372	0.0049	0.5408	0.0043
XGB	0.5422	0.0035	0.5479	0.0031
LightXGB	0.5834	0.0067	0.5890	0.0069

9.2 Appendix 2



9.3 Appendix 3



Distribution of tweets categorized based on Positive (1), Neutral (0) and Negative (-1) sentiments from January to December 2019 forecasted by CCST tool.