Simon Fraser University
Summer 2020
CMPT 459
Professor: Dr. Pei

Using Emotion Intensities in Tweets to Predict Covid-19 Cases in B.C.

August 11, 2020

PAUL DHANDA

#301293694
Bsc. Joint Mathematics and Computer Science

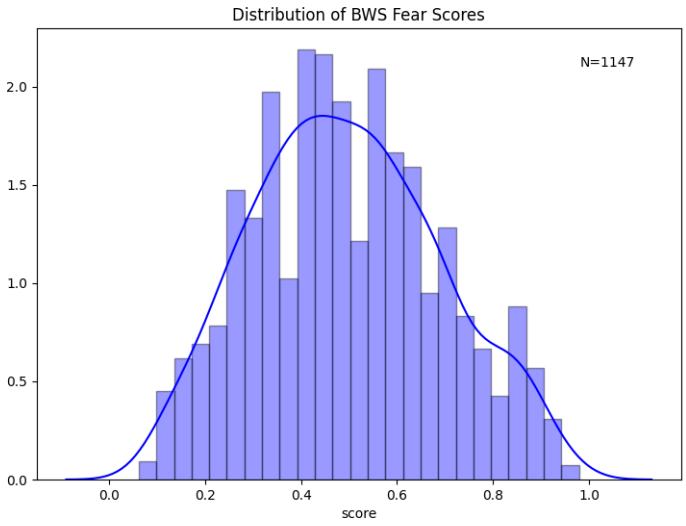
1 Motivation and Background

Covid-19 has disrupted how we live our daily lives and became a nightmare in our globalized world. Here in British Columbia (BC), cases numbers have been relatively stable. However, in recent weeks people have become complacent. The goal of this project was to see if there is a way to detect complacency through tweets and then to see if there is a relationship with the Covid-19 case counts here in BC.

2 Data

For this project, three data sets were used. The first one being *Tweet Emotion Intensity Dataset* (TEI) [MB17] where a Best Worst Scale (BWS) was used to annotate tweets with an emotional intensity score. These are scores that measure the level of a certain emotion on a normalized range from 0 - 1. For this project specifically the fear emotion was used. The assumption here is that fear is an emotion that captions how worry free and individual is. In Figure 1 is the distribution of BWS fear scores for the data set.

Figure 1: Score distribution



The second data set that was used was historical twitter data retrieved from IEEE Data Port [Lam20]. The data had tweets queried of a wide range of topics relating to COVID-19. This data had geographical labels, so we could extract tweets only from the BC region. The data set included tweets from March 20, 2020 to August 3, 2020.

The last data set that was used was the case numbers data set for BC from the BC Centre for Disease Control (BCCDC) [20]. This data contained every reported case in the region.

3 Architecture and Methodology

3.1 Tweet Extraction Pipline

Since the COVID-19 twitter data set had not yet processed, the construction of a robust tweet extraction pipeline was needed. In addition to the code, there is a `sqllite3` database to house our twitter data. All code and raw data for this module can be found in the `tweets/` directory. The raw data set was a collection of tweet ids separated into csv files, one for each day and can be seen in the `tweets/tweet_ids/` directory. One thing to note for the marker here is that the database was too large to put in with the code files, so it was left out. It takes a couple of hours to populate fully.

First, to initialize the database `init_db.py` needs to be run. Then, the entry point of this pipeline is the `recognize_files.py` file which acquires the full path of each csv and inserts them into the database. Next, the `process_files.py` file extracts all the tweet ids from each file and inserts those in database. Finally the `extract_tweets.py` file sends a request for every tweet in the data set and gathers some meta data about the tweet (Geographical location etc..) and inserts the raw tweet into the database. This pipeline has the appropriate error handling to acquire the full data set without interruption. The final result is the table which the following table schema for the tweet with 139,000 tweets in total.

```
Tweets
-----------------------------------
date_time       string
location_name   string
country         string
tweet_text      string
```

3.2 Tweet Pre-processing

Before any tweet was fed into a model or used for training, there were some pre-processing steps that were done. First of all special, characters and stop words were taken out. The retweets were filtered out. All @ references were removed. What was left was a string of the most important emotion capturing words.

3.3 Training the Fear Model

Training a model to detect fear using the TEI dataset was the most challenging part of this project. At first, a regression model was used since we are dealing with scores. However, the results ended-up resulting in too much noise in our output. So, a simpler binary classified model was used.

The model detects whether a high amount of fear is in a tweet or not. To train this model, we converted the fear scores of the TEI data set as follows:

```
def score_to_label(score):
    if(score < .6):
        return 0
    else:
        return 1
```

Dr. Mohammed does investigate a similar case in his paper *Emotion Intensities in Tweets*. He states:

> In some applications, it may be more important for a system to correctly determine emotion intensities in the higher range of the scale than in the lower range of the scale. To assess performance in the moderate-to-high range of the intensity scale,we calculated correlation scores over a subset of the test data formed by taking only those instances with gold emotion intensity scores ≥ 0.5.

This method converts our scores into binary labels. The value .6 here was chosen so tweets with higher levels of fear are classified as tweets containing fear. To this end, multiple classifying methods were compared with the following training scores and train-test split of 40%:

```
kNN classifier:         0.702
Rand forest classifier: 0.697
AdaBoost Classifier:     0.728
LinearSVC Classifier:     0.747
```

The next step was feature selection. To keeps things simple we chose to use unigram features for each tweet. Dr. Hasan decribes the use of uni-grams and lexicons in the paper *Detecting Emotions in Twitter Messages*:

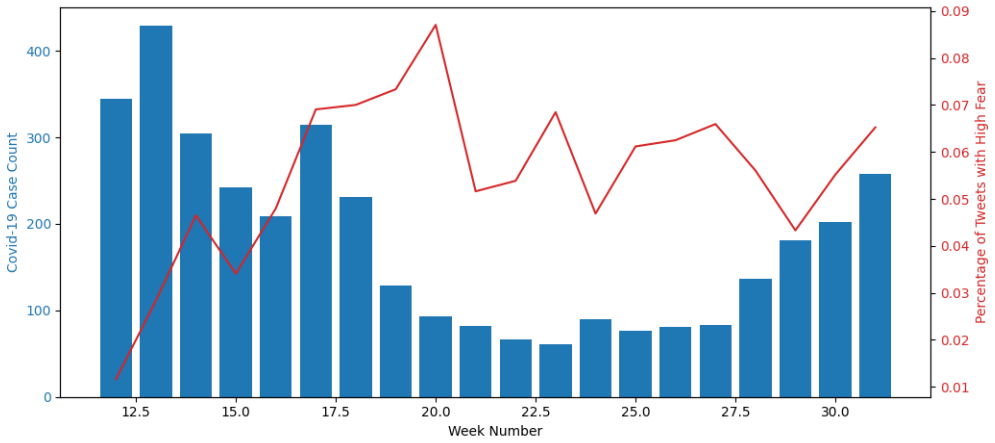
> Unigrams or single word features have been widely used to capture the sentiment or emotion of a tweet. Let (f1, f2, ..., fm) be our predefined set of uni-grams that can appear in a tweet. Each feature in this vector is a word from the dictionary of words in our data-set. Text messages can be classified into emotion categories based on the presence of affect words like ”annoyed”, and ”happy”. Therefore, the problem of high dimensional feature vector can be solved by identifying an appropriate emotion lexicon.We effectively design a domain-specific dictionary by using the lexicon of emotions, instead of all the words in our input data-set.As a result, our feature space does no longer include all the words in our training data-set, but instead it only contains the emotional words from the emotion lexicons. This method reduces the size of feature space dramatically, with-out loosing informative terms. We decide to use the LIWClexicon, which has been well validated and widely used in other studies [22, 23, 15]. LIWC contains a dictionary of several thousand words, wherein we use emotion-indicative categories such as positive emotions, negative emotions,anxiety, anger, sadness, and negation and utilize them effectively as our domain-specific dictionary. [HRA14]

Since, no free source of the LIWClexicon dictionary could be found online, only uni-gram features were used. However, if one would like to improve the model, it can be easily implemented like Dr. Hasan describes.

With the above information, a python model was build using Linear Support Vector Classifier (LSVC) and the use of un-igram features using the SKlearn library. The training was done on the TEI data-set and exported in a pickle format for later use. All code in this section can be found in the `training/` directory.

4 Results

Now that the model has been created and the tweets extracted, we can now use this fear model to come up with an idea of how fear in regards to COVID-19 is changing week to week in B.C. One caveat, however, the volume of tweets in BC was not enough to create a reliable fear related COVID-19 data-set. Thus, the entirety of Canada was used to model BC's fear data-set. Then, with the fear captured we can see how that compares to case counts week to week using the data from the BCCDC. Below is a graph of this comparison:



We can see here the relationship between the Percentage of Tweets from Canada that contain high fear and the case count in BC. While no work has been done to measure the relationship, we can see from week 20 onward, their being a negative relation between red line and the case counts. This verifies part of the assumption that when the general population starts to become more complacent (less fearful) that the covid-19 case counts will be on the rise.

One could use such a model as this to predict if cases are on the rise just depending on how the local/general population is tweeting in regards to COVID-19.

5 Further Improvements

This project is a good base for further work and can be improved in many ways. Feature extraction is one of main areas in which my model could be improved upon. Currently I use uni-grams as the feature for each tweets. However, there are many more features that can be potentially extracted. I mentioned before that a lexicon based approach would be a great added benefit to the model.

In addition to the features, the amount of tweets that were collected could be improved upon if time and cost were not a factor. Due to twitters rate limits, it was very tedious to get tweets from a certain time period and place. If, the reliance on a external data source of tweets was not need, a more rich data-set to test against could have been extracted.

And finally, the way to evaluate a correlation between percentage of tweets with fear ad the case count could also be extended. Currently, it is very trivial, where a the relationship is detect by inspection. However, further statistical tests could be used to measure such a relationship.

6 Final Thoughts

Overall, this was a great project to get acquainted with data mining methodologies. Three main data sources were used. One for model creation, one to test the model the against and the last to be used for comparison.

The project is good start and can be improved on in a multitude of ways. Overall, the methodology was quite simple, however, the results were still quite impactful.

References

[HRA14] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. "E.: Emotex: Detecting emotions in twitter messages". In: *Academy of Science and Engineering (ASE*. 2014.

[MB17] Saif Mohammad and Felipe Bravo-Marquez. "Emotion Intensities in Tweets". In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 65–77. DOI: 10.18653/v1/S17-1007. URL: https://www.aclweb.org/anthology/S17-1007.

[20] *BC COVID-19 Data*. 2020. URL: www.bccdc.ca/Health-Info-Site/Documents/BCCDC_COVID19_Dashboard_Case_Details.csv.

[Lam20] Rabindra Lamsal. *Coronavirus (COVID-19) Geo-tagged Tweets Dataset*. 2020. URL: http://dx.doi.org/0.21227/fpsb-jz61.