COMSW 6998 Practical Deep Learning

Correlation Between Election and COVID Using Sentiment Analysis

Authors: Xin Gu (xg2327), Yvette Wang (yw3396)

Introduction

Covid cases surged in the US since March in 2020 and there were shutdowns and pandemic policies that Americans have never experienced before. Meanwhile, election 2020 was happening. The US 2020 election has the highest percentage of voter turnout since 1908[1]. Many people cared about the election especially this year partly because the new president will determine how COVID will be handled in 2021. Their attitudes towards COVID have been highly affiliated with their political beliefs.

Analysis of how COVID impacted the election results based on election statistics have been done by media and institutes. We are inspired by this phenomenon and interested in knowing how both of them are correlated in a data standpoint, specifically sentiment scores.

Background

COVID and Election are the two elements that influenced sentiment of twitter users, which is our objective. Thus, we need two models: COVID model and Election model. Both of them should be trained on data that should not overlap with the target dataset. We then designed to use Election 2016 data to finetune a Election model while using COVID Canada data to finetune a COVID model.

Data

Twitter has a premium API: full archive API[2], which provides tweets since 2006 based on query. This is the ideal way to query data but this API sandbox has limits of 50 requests per month and 100 results per request. Due to time restriction of this project, this API can only be used when necessary.

Due to Twitter Developer policy[3], only Tweet IDs can be disclosed publicly. We then made use of tweet IDs in three public datasets: Election 2016 Tweet IDs by Justin Littman, et al[4], Election 2020 Tweet IDs by Emily Chen[5], et al, and COVID19 Tweet IDs by Emily Chen, et al[12]. With tweet IDs, twitter tweet lookups endpoint gives the content, which we later used as training data and election 2020 data.

COVID Canada has the two requirements that tweets have COVID19 related keywords and are posted from Canada. There is no existing dataset for this at least to our knowledge, so the data of COVID Canada was queried using Twitter Full Archive API.

Sentiment Analysis

Sentiment Analysis is a task in Natural Language Processing. The goal of this task is to detect the latent sentiment in a given text content. Usually, people care about whether a text shows a positive view or a negative view. Sentiment Analysis is a popular approach analyzing posts in social media to understand the trending views of the society. Approaches in this task can be divided into two main methods, lexicon-based approaches and machine-learning-based approaches.

Lexicon-based approaches score sentiment words in a sentence and combine the scores to get a final sentence level score. The word-level sentiment scores are obtained by creating a lexicon manually or automatically using selected seed words to expand a list of words [7]. The choice of sentiment words are mainly adjective words. When analyzing sentiment under different content, a lexicon may need some modification. In our task of analyzing election text, a lot of sarcasm is used. We expect that we may need a big modification from a general English sentiment lexicon, which could be difficult.

Machine learning approaches have various algorithms. The common idea in this approach is to treat sentiment analysis tasks as supervised classification tasks. Hence, classifiers such as SVM and Naïve Bayes are usually used in machine-learning-based sentiment analysis. Recently, neural network approaches are explored and have achieved the state-of-the-art performance [6]. Common deep learning approaches are LSTM [8] and BERT [10]. LSTM, Long Short-Term Memory, is based on RNN. LSTM works well in natural language processing tasks, as sentences are simply structured (i.e., usually from left to right.) and it is good at learning long distance contextual dependency among words [9]. However, due to the structure, even bidirectional LSTM can still miss some inter-word dependency information. The other method is BERT, short for Bidirectional Encoder Representation from Transformers, which is an encoder which learns the features of a data set and can be used for multiple tasks such as machine translation (with a decoder), sentiment analysis, and question-answering, etc [10]. The use of Masked LM and Next Sentence Prediction in BERT improves robustness and learning of sentence-level relationships. The encoder transfers the input sentence into token embeddings, segment embeddings, and position embeddings [10]. When applying to sentiment analysis tasks, the result embeddings are fed into a classifier and train on sentiment labels [11].

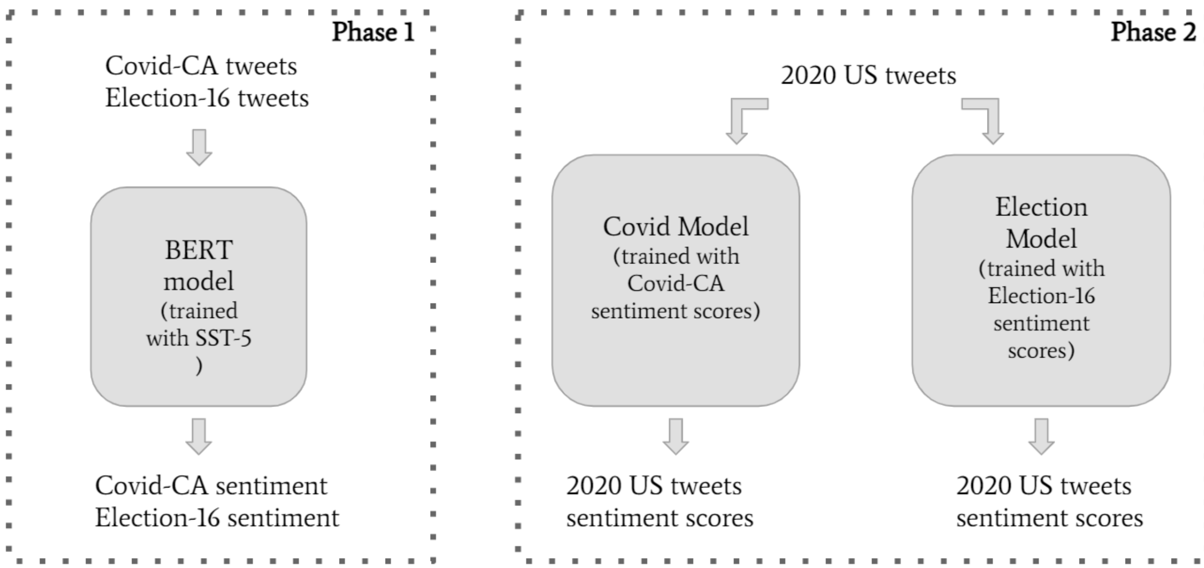
Implementation

Our approach can be divided into three steps. First, we “faked” the training dataset. Election 2016 data and COVID19 data do not have labels (sentiment score), so we used a BERT model to approximate them, which gives scores from 0 to 4. Then, we trained a Election model on 2016 dataset and a Covid model on COVID Canada dataset. Next, we predicted scores of election 2020 and US Covid data using these two models. Final step was to compare and analyze the prediction differences.

We collect data using Twitter Developer API, which allows us to search the full-archive of the tweets with keywords and to lookup specific tweets with their tweet IDs. We use the search API to find tweets related to Covid in Canada, because the data sources we find are either tweets in the US or globally. For the other datasets, we retrieve the tweets content by the tweet IDs from datasets mentioned in the previous sections. After the data retrieval, we process the tweets with twitter preprocessor library to remove all the mentions (username with a prefix of “@”), links, and special characters.

In Figure 1, we show the training process in two phases. All the three BERT models and corresponding tokenizers we use are based on BERTweet [13]. Besides general-purpose

English tweets, BERTweet pre-trained models also provide weights trained on Covid-19 twitter datasets. We attach a classifier on the base BERTweet model. The first phase is to train the model using the Stanford Sentiment Treebank dataset which consists of 0.3 million training data labelled as five sentiment classes. With this model, we can get both sentiment scores (in the scale of 0 to 4) of Canada's Covid tweets and Election 2016's tweets. These sentiment scores predicted from phase 1 are put into phase 2 to train the two models respectively. After training, the two models each predict sentiment scores of tweets collected in 2020. We compare the two sets of predictions with respect to time.



Two phases of the training process (Figure 1)

Result

RT @mmpadellan: I'm old enough to remember when #RacistTrump was disrespectful of the AMAZING reporter @Yamiche.	0	election
RT @AntillanaSoy_: nobody recovers from COVID-19 in 3 days ... specially not a 74 year old man	1	covid
RT @HassanPeppe: Karasuwa local government inspection and monitoring on observance of covid-19 protocols today 17th August 2020 https://t.c?€?	2	covid
RT @BillKristol: This is moving and powerful. America is better than Trump. https://t.co/a123CCG8UR	3	election

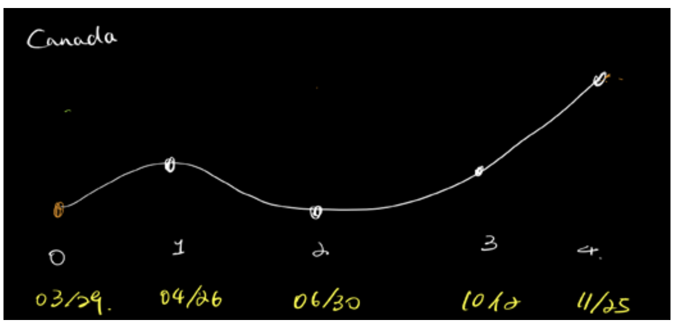
RT @TeamTrump: Congratulations to President @realDonaldTrump and Vice President @Mike_Pence!	4	election

Example of Sentiment Scores predicted by Election and Covid Models

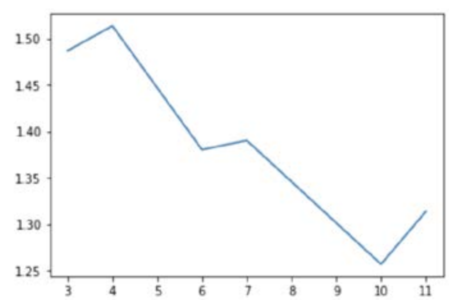
The table above are tweets sentiment predictions by two models on the target data. score 0 is the most negative and 4 is the most positive. By humans reading the tweets, we agree with the predictions. We found that in most cases, the predictions are fairly accurate.

We chose Canada to collect representative COVID data for training purposes. Due to the API rate limit, we picked five turning points in the Canada COVID trend and collected tweets around these dates (Figure 2). The plots of sentiment scores approximated by the BERTweet model (Figure 3) along the time makes a lot of sense as the longer the pandemic is, the more negative feeling people would have.

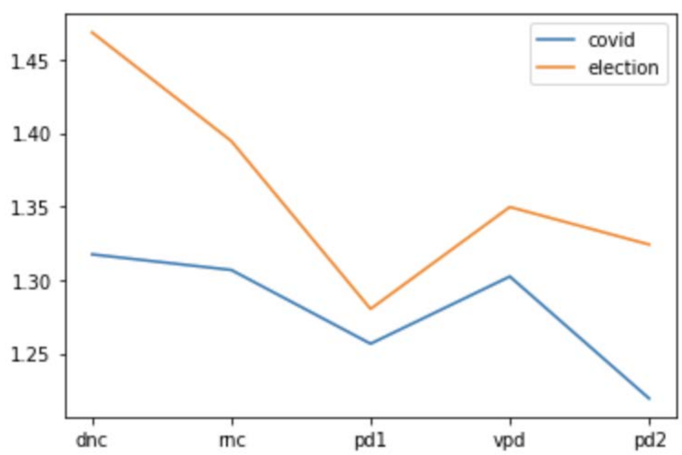
The sentiment scored on the target dataset by two models is shown in Figure 4. The two score trends are very similar to each other while covid sentiments are always lower than Election sentiment predictions. This meets our assumptions that people in general would feel more negative as the pandemic stays longer and COVID has more negative side of feeling than election.



COVID Confirmed Cases Trend in Canada (Figure 2)



COVID-CA Sentiment Score Trend (Figure 3)

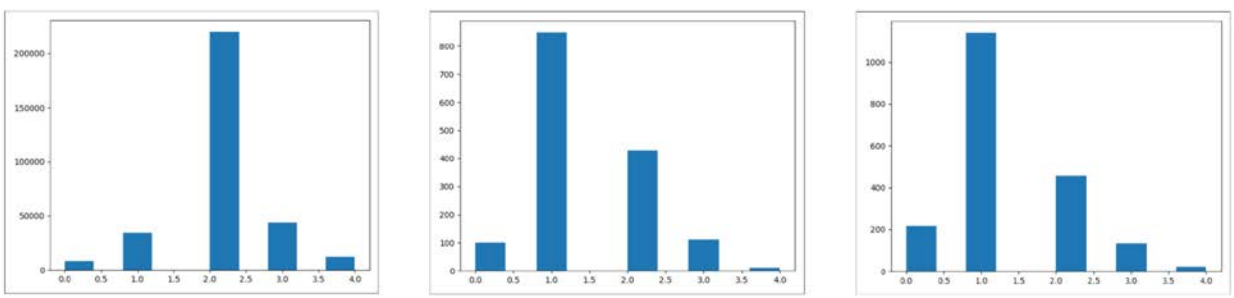


Sentiment Score differences on target dataset by Election and Covid models (Figure 4)

Discussion

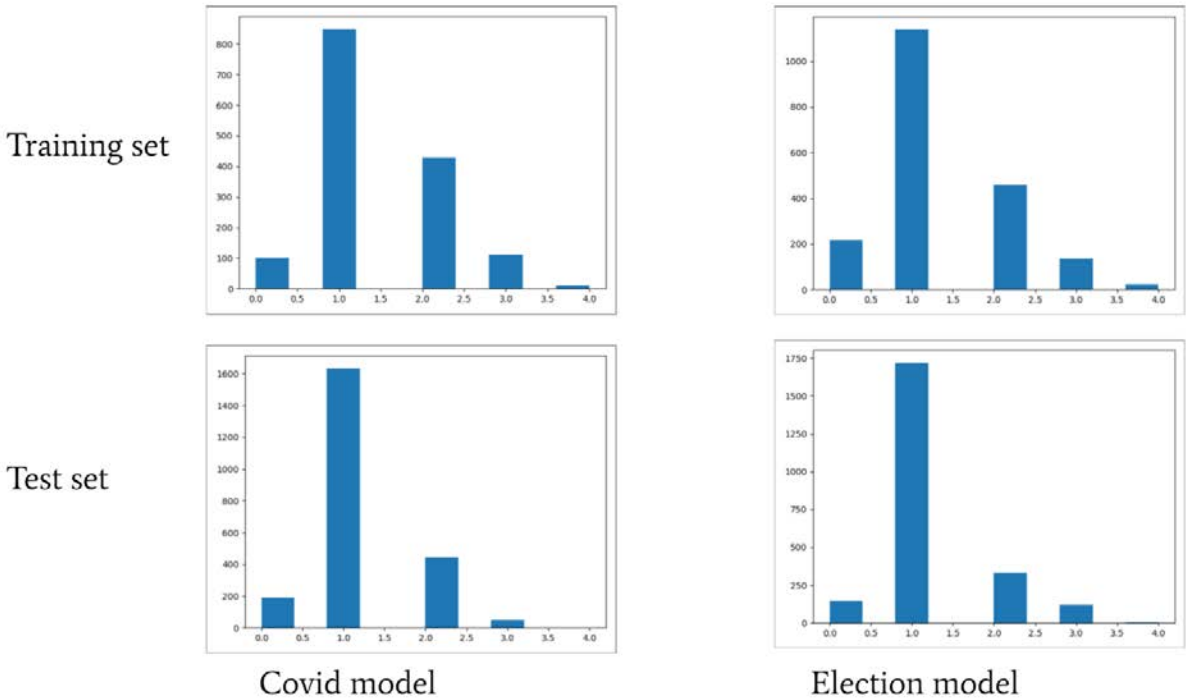
There were challenges at each step of our project. Full Archive API rate limits restrict the size of COVID Canada data. At the same time, we do not have any data with labels. The labels here we use in training data were approximated by a BERT model. This means there are errors in training data for which we cannot know the percentage. The two models we obtained after Phase I have low accuracies, which combining with the previous challenge leads to much lower accuracy. Except these, the BERTweet model used in Phase I was trained on twitter data with COVID topic, which can have overlap with our target dataset.

In both training phases, we use a VM with 4 vCPUs, 15 GB RAM, and one NVIDIA P100 GPU. Training the model of phase 1 takes 6 hours on a dataset with 0.3 million records and gets a validation accuracy of 0.82. Training the models of phase 2 takes 3 minutes each on a 2000-tweet dataset and gets a validation accuracy of 0.73 on the Covid-CA dataset and 0.70 on the Election-16 dataset. Our assumption before training is the sentiment class distribution of the Covid datasets and the Election datasets should be different from that of the general English training set. In phase 1, we observe (Figure 5) a left-skewed distribution from the predictions on Covid-CA and Election-16 tweets by the model. The phase-1 model predicts about half of the tweets in the two datasets are slightly negative.



SST Covid-CA Election-16

Sentiment class distributions of phase 1 (Figure 5)



Sentiment class distributions of phase 2 (Figure 6)

In phase 2, we make a similar analysis as phase 1 (Figure 6). We assume that the training set has a similar sentiment class distribution as the testing tweets in the US 2020. Compared with the training sets (Covid-CA and Election-16), the prediction results from Covid model and Election model are slightly more left-skewed. It may be because the phase-1 model is trained on a dataset with a normal distribution on number of classes which makes its predictions more neutral. However, in phase 2, we train the model with the negative skewed datasets, which makes the predicted classes more negative.

From Figure 6, we find that the predicted class distributions of both models are similar. We want to investigate whether the two sets of predictions disagree with each other. We define a "large" disagreement as two predicted sentiment classes have a difference greater than 2 (eq. 1).

$$Disagreement : abs(C_{Covid} - C_{Election}) \geq 2 \qquad \textit{(Eq. 1)}$$

The number of tweets with large disagreement is about 80 out of the 2200 test data. We show some examples of large disagreements in Figure 7. The first example is a Covid-hashtagged tweet. We can sense the positiveness of the sentence, and the Covid model rates a three. However, the election model rates one maybe because of the mention of "racism". The other example has both Covid and election hashtags with a slightly positive tone, but the Covid model rates it one. Most of the model disagreements have one model mistakenly lowly rate the sentences. Hence, we assume our results are slightly negative-biased.

Example 1: Covid: 3, Election: 1

RT @nanosounds: If you want to help tackle racism against ESEA people and you're based in England, this is a good way to do so! Talk to you...

Example 2: Covid: 1, Election: 3

RT @BarackObama: 102 never looked better! Grateful for all the folks like your great aunt who continue to show up and vote in this importan...

Examples of large disagreements (Figure 7)

Conclusion and Future Work

When learning how people feel due to some events through sentiment analysis, it is helpful to exclude effects from other events. In the year of 2020, people in the US can be impacted by both Covid-19 and the election. From the project, we investigate how the two major events in the year affect the sentiment level using BERT sentiment classifiers and Covid tweets in Canada and election tweets from 2016 as controlling groups. Our analysis shows negativeness in tweets in 2020. In this project, the lack of data labelling is the major issue we have. We suggest using tools such as AWS crowdsourcing service to label text. Besides that, we also find it is hard to correctly label the sentiment of election tweets. People tend to use more sarcasm in political tweets. Sentiments analysis on political context will be helpful in this analysis. Another possible future work is to label the data and train an aspect-based sentiment model, which may improve the performance when a sentence is related to both Covid and election. When there is a labeled dataset with proper size, other interesting problems can be analyzed as well, such as which model has a larger impact on twitter users during a specific event by comparing the accuracies of both models.

Reference

1. Kevin Schaul, Kate Rabinowitz, Ted Mellnik (2020) ‘2020 Turnout Is The Highest In Over A Century’ *The Washington Post* , 5 November. Available at: https://www.washingtonpost.com/graphics/2020/elections/voter-turnout/
2. Twitter, ‘Getting Started With Premium Search Tweets: Full Archive API’, available at https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive
3. Twitter, ‘Developer Policy’, available at: https://developer.twitter.com/en/developer-terms/policy
4. Littman, Justin; Wrubel, Laura; Kerchner, Daniel, 2016, "2016 United States Presidential Election Tweet Ids", https://doi.org/10.7910/DVN/PDI7IN, Harvard Dataverse, V3

5. Emily Chen, Ashok Deb, Emilio Ferrara. #Election2020: The First Public Twitter Dataset on the 2020 US Presidential Election. Arxiv (2020)
6. Chen E, Lerman K, Ferrara E Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set JMIR Public Health Surveillance 2020;6(2):e19273 DOI: 10.2196/19273 PMID: 32427106
7. M. Taboada, J. Brooke,M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis", *2011 Association for Computational Linguistics*. 2011.
8. S. Hochreiter, J. Schmidhuber, "Long short-term memory", *Neural Computation*, pp. 1735–1780. 1997.
9. J-H Wang, T-W Liu, X. Luo, and L. Wang, "An LSTM Approach to Short Text Sentiment Classification with Word Embeddings", *The 2018 Conference on Computational Linguistics and Speech Processing*, pp. 214-223. 2018.
10. J. Devlin, M-W Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv:1810.04805*. 2018.
11. R. Khandelwal, "Multi-class Sentiment Analysis using BERT". [Online]. Available: https://towardsdatascience.com/multi-class-sentiment-analysis-using-bert-86657a2af156. [Accessed: 18-Dec-2020].
12. "Papers with Code - Sentiment Analysis," Sentiment Analysis | Papers With Code. [Online]. Available: https://paperswithcode.com/task/sentiment-analysis. [Accessed: 18-Dec-2020].
13. D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets", *arXiv:2005.10200*, 2020.