Final Project Proposal
"$Know - It - All$"–A Multi-class And Generally Applicable Sentiment Analysis Model For Social Media Texts

Yunzhe Sun(sy2825) , Silvey Yu (ly1164)

1 Introduction

1.1 Introduction to the problem

Valuable information exists on social media–Whether or not people welcome an event, whether or not people approve of a policy, how much people favor or dislike certain products.

When there are needs for public opinion mining, for data analysis on how the population reacts to a certain event/product/incident, etc, and for policy making, sentiment analysis takes on vital importance. It transforms massive text into statistically comprehensible sentiment data. This allows policy makers, economists, and data analysts to have a deeper insight into the problems at hand by monitoring the wider public opinion in social media.

Our project aims to build a useful tool for sentiment extraction on social media texts. We name it "$Know-It-All$", We will train a sentiment analysis model for twitter texts, based on twitter datasets, for the twitter data is widely available in large quantities, and often pre-processed by researchers into helpful corpus. Also, twitter is a platform where instant reactions and comments to social events will take place. Sentiment analysis on this platform can sensitively detect public opinions towards certain major events.

Twitter has a word limit for tweet posts. Therefore, each data entry would not be a huge paragraph, and this allows us to focus more on sentence-level and short-comment level sentiment analysis, which is suitable, because we will be applying our model also on social media texts, which will generally be short.

1.2 Questions we would like to ask:

- **(1) How can we achieve high accuracy in sentiment analysis on social media text?**
- **(2) How detailed can the emotion classification be?**

  Can our model differentiate mild, strong, and stronger emotions? Can it classify detailed emotions such as insult/frustration/anxiety/...etc, apart from merely performing simple classification in "positive/neutral/negative" categories.

- **(3) Can we run our pre-trained model another twitter dataset containing people's comments on CoVid 19? Is the model really helpful and generally applicable for datasets apart form the train/validation/test datasets we selected? Would the result make sense?**

  We would also test our model on other datasets and evaluate whether it is truly generally applicable. In particular, we found on Kaggle some CoVid 19 twitter text datasets. The pandemic has a real-life significance, and we would like to evaluate our model on that to see if our model can be a good and generally applicable tool for social media sentiment analysis.

1.3 Selection of related works

- Felbo, Bjarke, et al. "Using Millions of Emoji Occurrences to Learn Any-Domain Representations for Detecting Sentiment, Emotion and Sarcasm." Its relating application in tweet texts: [Link to the open-source project DeepMoji].
- Nurulhuda Zainuddin, et al. "Sentiment Analysis Using Support Vector Machine" [Link to the paper]

- Yohanssen Pratama, et al. "Implementation of Sentiment Analysis on Twitter Using Naïve Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election" [Link to the paper]
- A Github repository (with code and report) on sentiment analysis using SVM, Naive Bayes, and perceptron [Link to the GitHub repo]
- A GitHub repository on Tweet text pre-processing. [Link to the GitHub repo].
- A GitHub repository for word stemming [Link to the GitHub repo]

2 Datasets

2.1 Twitter/Youtube datasets with specified emotions

We plan to use datasets extracted from online scenarios with different emotions, for the model to learn to classify various emotions. These datasets are found from papers focusing on sentiment analysis and opinion mining. Each dataset is specified in 2 or more emotions, labeled with a Numpy array with 0's and 1's.

We intend to pool them We will expand the label to the same-length numpy arrays, with each "1" indicating an existing emotion. We might drop certain datasets if unsolvable obstacles occur during our data pooling and cleaning or model-building.

- SCv2-GEN Tweet sentence dataset with sarcastic/non-sarcastic emotions. It is a .csv file with pre-processed sentences, with labels 1 and 0 referring to sarcastic and non-sarcastic. It has upper case letters and repeated punctuation marks. We need to unify all letter cases and clean the puncuations. [Link to SCv2-GEN is here]
- kaggle-insults datset with insult/non-insult emotions. It is a .csv file with sentences of insult/non-insult, [Link to Kaggle-insult]
- PsychExp dataset. With various emotions: joy, fear, anger, sadness, disgust, shame, guilt [Lind to PsychExp] indicated by label [1,0,0,0,0,0,0], with 1 indicating the existence of a corresponding emotion. In .csv format.
- SS-YouTube dataset. Comments from YouTube, .csv format. proprocessed and labeled with positive 1 and negative 0. We still need to get rid of "@" symbols and usernames.
- Olympic dataset. Olympic tweets with negative, high control negative, positive, etc. emotions, in .csv format. All sentences are pre-processed into lowercase, but "@" symbol and username are to be cleaned.

2.2 COVID-19-related Tweet datasets (2 in total)

- A labeled dataset with positive/negative/neutral emotions. [Link: Coronavirus Tweets].

  This dataset consists of original tweet text collected from what tweeter users posted, which includes time posted (mainly during 2020, with specification of dates), user region, labeled sentiment (positive/negative/neutral), usernames. It is already pre-processed, with sentiment labels listed above. It is in .csv format, 10.5MB, with 41157 unique tweets.

  After some cleaning of unwanted symbols and processing of stop-words, We will test our model on it, to see if our model correctly detects the sentiments that belong to the three labels. For example, if our model predicts "joy" on a "positive" labeled tweet text, and "joy" is a subtype of "positive", it is counted as a correct prediction.

  Then, based on our detailed emotion classification, we can try displaying a data analysis result with timeline and region.

- An unlabeled dataset with covid vaccine related tweet texts. [Link: Covid Vaccine Tweets]

  This dataset has more columns than the previous one, and also much more rows. It is also in .csv format, 133.62MB, with 312273 unique texts. It includes clumns such as:

  The tweet texts are not pre-processed and cleaned, so we would need to also delete some frequently appearing symbols from them. Without labels, it would be hard to test our prediction accuracy, but could serve as a practical environment for us to apply our model in real life.

2.2 Algorithms

We would first try:

- (1) Naive Bayes for classification.

  This is the most basic solution to text classification. We reserve it as an option to handle certain part of classification task, if needed.

- (2) Support Vector Machine for classification and evaluate the polarity score.

  Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories

  In sklearn, we can also find packages for SVM.

- (3) Perceptron is also an option. Keras and TensorFlow can be used for this implementation.

- (4) If time still allows, we would also try to run the the open-source Long Short-Term Memory(LSTM) Model introduced by the DeepMoji. [DeepMoji link is here]

2.3 Expectation:

We have 3-step goals, we will try to achieve a next goal after achieving the previous one.

- Build a model that accurately marks joy, sadness, anger, sarcasm, etc–the emotions we mark in our pooled dataset

  (with error rate preferably less than 20% in the first stage, and then work to tune parameters and decrease the error rate)

- Improve the model so it can recognize the extend to which an emotion is expressed

  (high, medium, low level of joy for example)

- Apply the model for practical purposes on timely datasets such as the CoVid 19 tweet texts and CoVid 19 vaccine tweet comments from Kaggle.

  (See if the model provides us with valuable insights on people's emotions in a timeline. This could indicate whether or not our model has the potential to be a generally applicable sentiment analysis model for sentence-level/short paragraph level social media texts, and potentially serve as a useful tool for opinion mining and further data analysis.)