Project Final Report

29, March, 2018

Natural Language Process

Md Lutfar Rahman(mrahman9@memphis.edu)

Title: Sentiment Classification by Linear-chain CRF

Introduction

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. In this project, I worked on detecting sentiment in a given text. The real motivation is, automatically classify twitter posts based on sentiments it imply. Out of endless tweets we may look for tweets that represent a particular public sentiments. On the other hand we may dig down politicians dialogue in terms of sentiment and factual relation. On a bigger scale if we can ensure high accuracy we may understand the current demography of emotion and sentiments on certain issues.

Related Work

There have been numerous works on sentiment analysis. Most of them focus on twitter sentiments analysis as the labelled data on twitter is fairely available. Other source of data is not that common. Companies such as Twitratr (twitrratr.com), tweetfeel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are just a few who advertise Twitter sentiment analysis as one of their services. In the recent year there have been a number of papers looking at Twitter sentiment and buzz (Jansen et al. 2009; Pak and Paroubek 2010; O'Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010; Davidov, Tsur, and Rappoport 2010). Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data

Approach

At the beginning of my journey, I was looking for context based sentiment analysis i.e. if I can take account of context/aspect of emotion, I might get something better. After initial digging, I found Linear-chain CRF has the exact same idea that I was looking for. But going forward, reading more about Linear-chain CRF, I found that the initial design of Linear-chain CRF is to model sequential data, such as labels of words in a sentence**.** So, it can detect If a word falls into sentimental or not from a labeled data. But I was looking for it to work for the whole sentence. I found no theoretical approach to go forward with Linear Chain CRF. This is where I

got stuck. Rather I was digging aspect based sentiment analysis (ABSA). Found some papers that talk about it. In order for ABSA to work the train data needs integer rating value to show the polarity of positivity or negativity. The data initially I had just three labels positive, negative, neutral. Finally, I used nltk naive bayes classifiers to do the sentiments analysis.

Data:

To train a sentiment entity recognition model, we need some labelled data. There are lots of twitter sentiment labelled data is available. I have decided to use sentiment140 dataset.

It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment .

It contains the following 6 fields:

1. target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
2. ids: The id of the tweet (2087)
3. date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. flag: The query (lyx). If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted (robotickilldozr)
6. text: the text of the tweet (Lyx is cool)

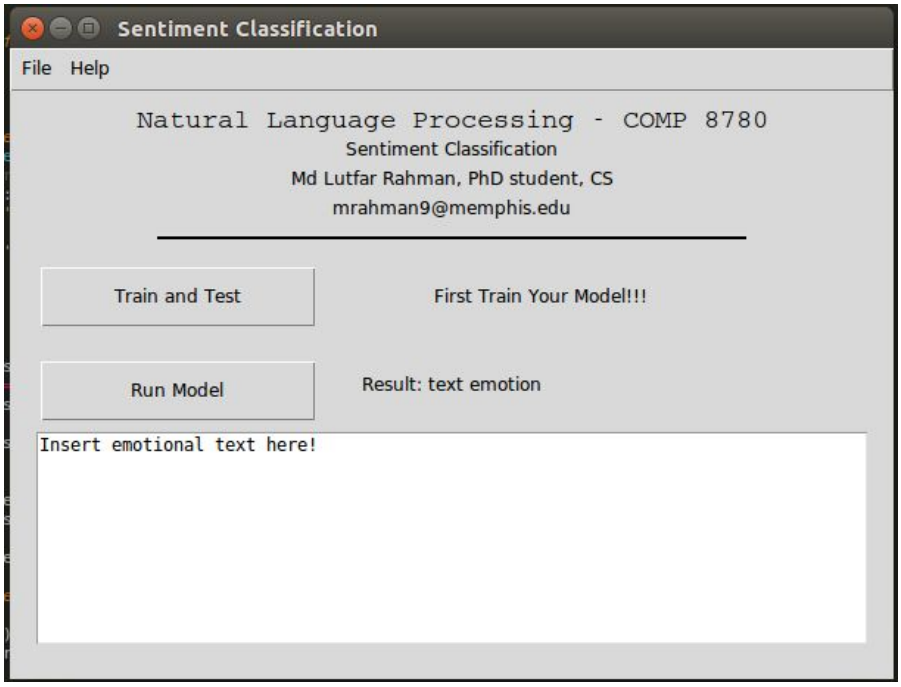
Implementation Details

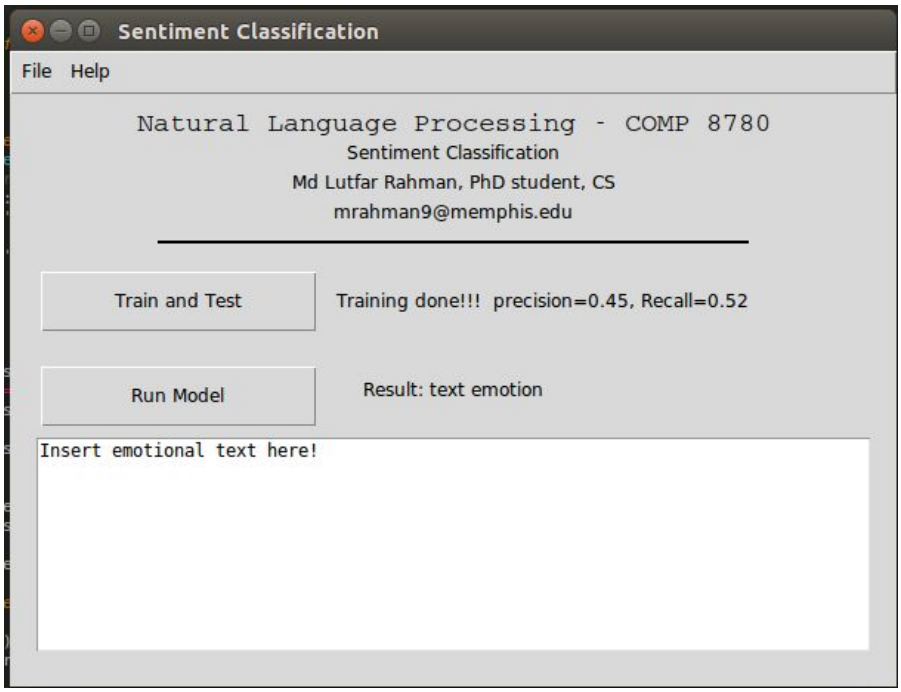
I extracted dataset in such a way that I just have a pair of string (“sentence”, emotion) for each twitter data. Emotion can have three states: positive, negative, neutral. Then I converted a sentence as a sequence of words having the same label. Then I extracted features by taking unique words. Whenever I get a new test data I convert it to sequence of unique words, that becomes unique features. Finally classifiers give the result for a new data.

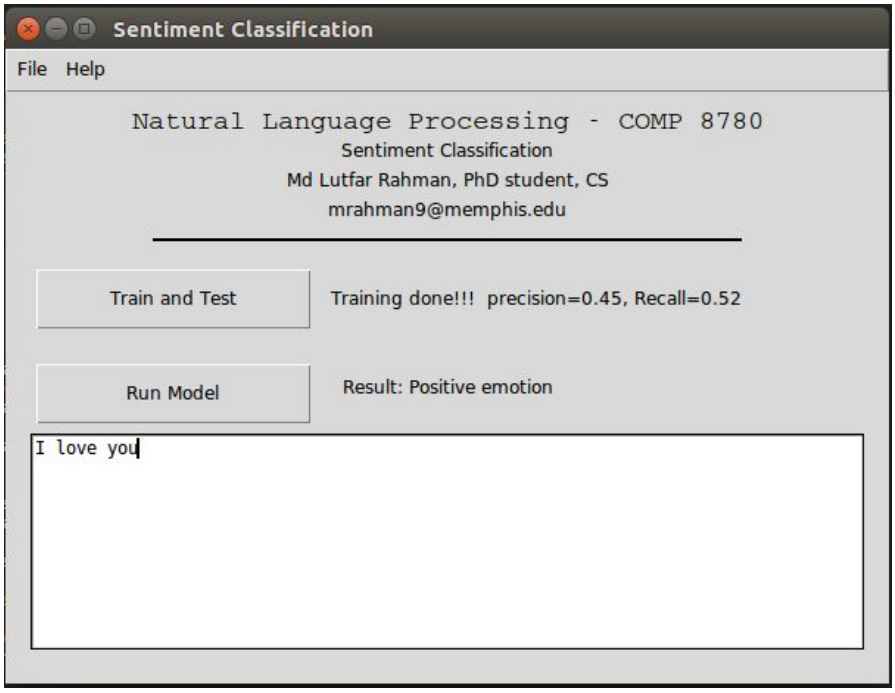
The following Python 3.5 package was necessary:

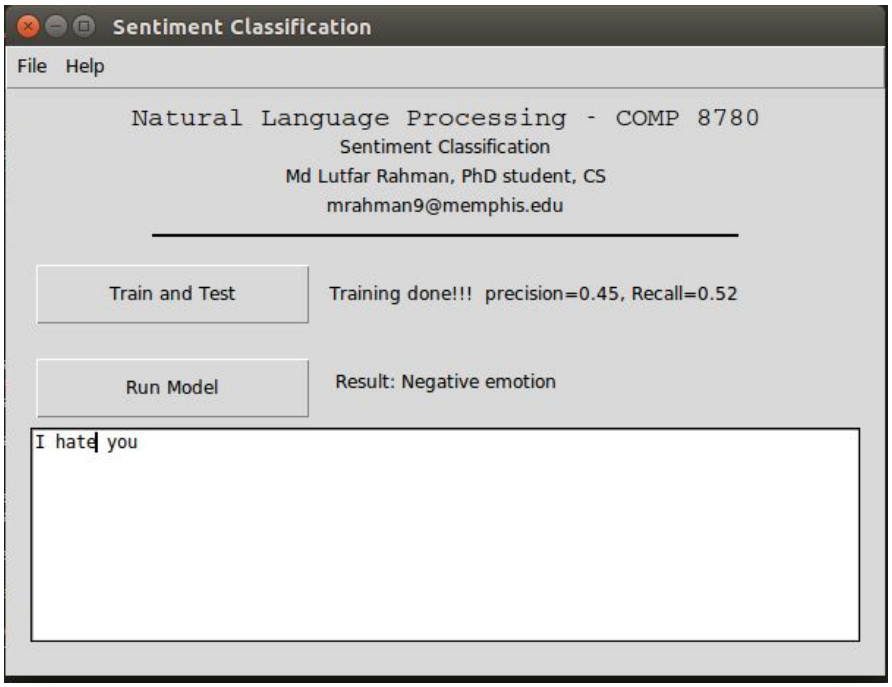
1. Numpy
2. Nltk
3. Pycrfsuite
4. Scikit-learn
5. BeautifulSoup
6. Tkinter

I used python tkinter and designed a GUI for the project:


Sentiment Classification
File Help
Natural Language Processing - COMP 8780
Sentiment Classification
Md Lutfar Rahman, PhD student, CS
mrahman9@memphis.edu
Train and Test
First Train Your Model!!!
Run Model
Result: text emotion
Insert emotional text here!


Sentiment Classification
File Help
Natural Language Processing - COMP 8780
Sentiment Classification
Md Lutfar Rahman, PhD student, CS
mrahman9@memphis.edu
Train and Test
Training done!!! precision=0.45, Recall=0.52
Run Model
Result: text emotion
Insert emotional text here!





Experiment and Results

In order to compare and test models I prepared subset of original dataset.

The test dataset has 496 tweets
Precision = 44.7% , Recall=51.8%

Conclusion

It appears that sentiment analysis is one of the hardest task is the natural language processing. It needs to understand context of a certain text. The advancement towards this fields still needs a lot of improvement. The precision and recall are not still good enough.

Future Work

My original goal is to design a model that works closer to a human mind. How a human mind understand emotion in a given text. I definitely believe, if we can improve contextual analysis we will also improve sentiment analysis. That may lead us to something bigger in future.

References

1. Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11.538-541 (2011): 164.
2. Sarlan, Aliza, Chayanit Nadam, and Shuib Basri. "Twitter sentiment analysis." Information Technology and Multimedia (ICIMU), 2014 International Conference on. IEEE, 2014.
3. Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.
4. Jansen, Bernard J., et al. "Twitter power: Tweets as electronic word of mouth." Journal of the Association for Information Science and Technology 60.11 (2009): 2169-2188.