Twitter Sentiment Analysis

Project report in fulfilment of the requirements of the course of

Spring 2018 - Intro to Machine Learning (CS-580L-01)

By

Shriprasad Bhamare Ulugbek Ergashev

Supervisor
Prof. Arti Ramesh

Computer Science Department

Binghamton University
2017-18

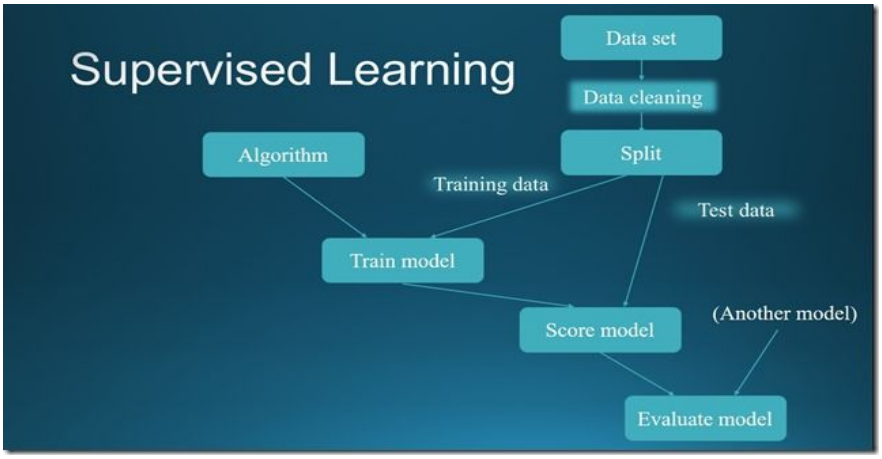
Table of Contents

1. Abstract

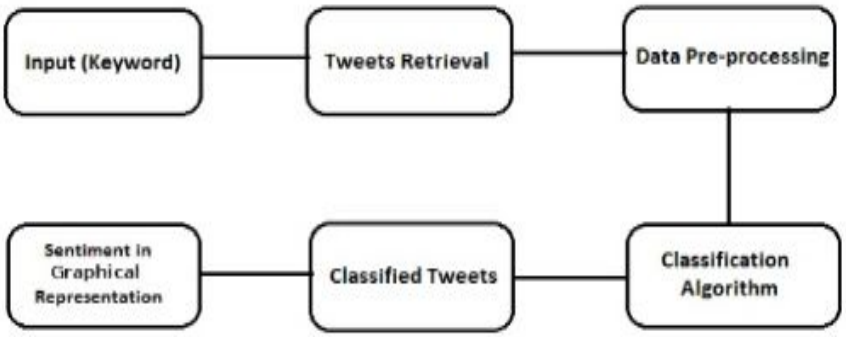
The aim of this project would be to classify tweets to positive or a negative sentiment . A sample input will consist of the labelled dataset in csv format. This input will then be processed for cleaning and feature extraction. Different Machine Learning classifiers(Naive Bayes ,Logistic Regression, Decision Trees ,SVM ,Neural Networks) will then be trained on a dataset containing multiple such sample inputs. Our goal is mainly to select the best result giving model in order to classify real time tweets as a positive or negative sentiment.

2. Scope and Workflow of the project

2.1 Selection of the highest accuracy giving Model

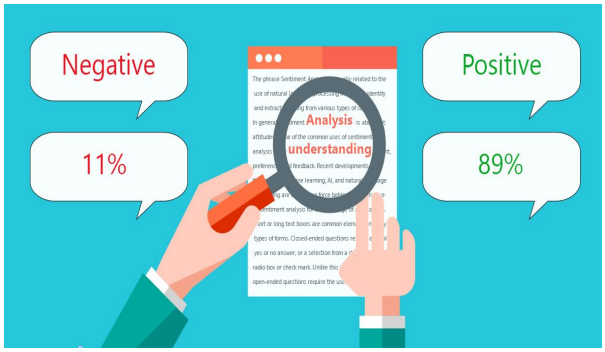


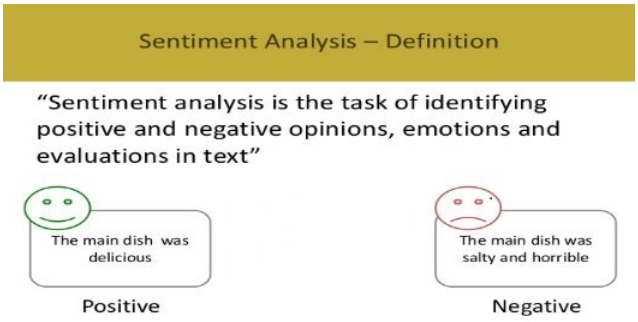
2.2 Classification of Real Time scraped tweets



3. Introduction

Community's view and feedback have always proved to be the most essential and valuable resource for companies and organizations. With social media being the emerging trend among everyone, it paves way for unprecedented analysis and evaluation of various aspects for which organizations had to rely on unconventional, time consuming and error prone methods earlier. This technique of analysis directly falls under the domain of "sentiment analysis". Sentiment analysis encompasses the vast field of effective classification of user generated text under defined polarities. There are several tools and algorithms available to perform sentiment detection and analysis including supervised machine learning algorithms that perform classification on the target corpus, after getting trained with training data. Lexical techniques which performs classification on the basis of dictionary based annotated corpus and Hybrid tools which are combination of machine learning and lexicon based algorithms. In this project, we will use different machine learning algorithms to get the best model which can classify real time tweets as a positive or negative sentiment.





4. Dataset Description

The Twitter Sentiment Analysis Dataset contains 100,000 classified tweets, each row is marked as 1 for positive sentiment and 0 for negative sentiment.Twitter dataset taken from http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/
Dataset is splitted up as 75% for training purpose and 25% for test purpose.We used 10% of data from the train data for validation purpose.The training dataset is a csv file of type tweet_id,sentiment,tweet where the tweet_id is a unique integer identifying the tweet, sentiment is either 1 (positive) or 0 (negative), and tweet is the tweet enclosed in "". Similarly, the test dataset is a csv file of type tweet_id,tweet

4.1 Different Classes of Sentiment Analysis

4.1.1 Positive Sentiments

This refers to positive attitude of the speaker about the text. Words with positive sentiments reflect happiness, joy, smile etc. In case of political reviews, if the positive reviews/sentiments about the politician are more, it means people are happy with his work.

4.1.2 Negative Sentiments

This refers to negative attitude of the speaker about the text. Words with negative sentiments reflect sadness, jealousy, hate etc. In case of political reviews, if the negative reviews/sentiments about the politician are more, it means people are not happy with his work.

5. Data Preprocessing and Feature Engineering

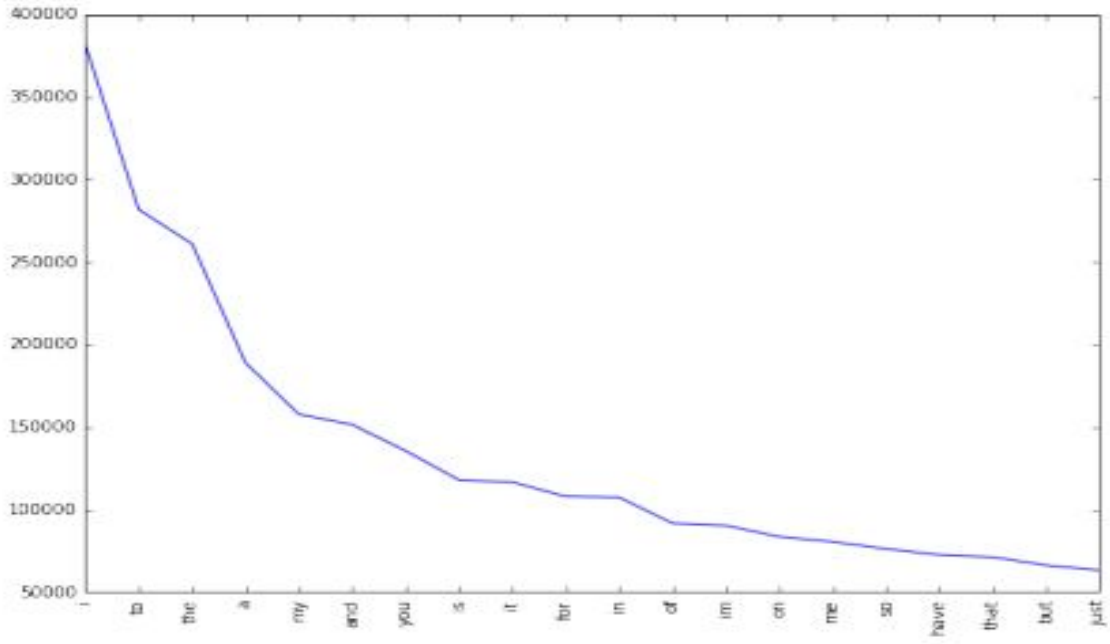
We have applied an extensive number of pre-processing steps to standardize the dataset and reduce its size. We first do some general pre-processing on tweets which is as follows.

- Convert the tweet to lower case.
- Replace 2 or more dots (.) with space.
- Strip spaces and quotes (" and ') from the ends of tweet.
- Replace 2 or more spaces with a single space.
- Replace URL with a word URL using a regular expression ((www\.[\S]+)|(https?://[\S]+))
- Replace all user mentions with the word USER_MENTION using regex @[\S]+
- Hashtags followed by words are removed from the data using regex #(\S+)
- Retweets begin with the letters RT. We remove RT from the tweets as it is not an important feature for text classification. The regular expression used to match retweets is \brt\b.
- Replace the matched emoticons with either EMO_POS or EMO_NEG

Emoticon(s)	Type	Regex	Replacement
:), :), :-), (:, (:, (-:, :')	Smile	(:\s?\)|:-\)|\(\s?:|\(-:|:\'\))	EMO_POS
:D, : D, :-D, xD, x-D, XD, X-D	Laugh	(:\s?D|:-D|x-?D|X-?D)	EMO_POS
;-), ;), ;-D, ;D, (;, (-;	Wink	(:\s?\(|:-\(|\)\s?:|\)-:)	EMO_POS
<3, :*	Love	(<3|:*)	EMO_POS
:-(, : (, :(,):,)-:	Sad	(:\s?\(|:-\(|\)\s?:|\)-:)	EMO_NEG
:,(, :'(, :"(	Cry	(:,\(|:\'\(|:"\()	EMO_NEG

Raw	misses Swimming Class. http://plurk.com/p/12nt0b
Normalized	misses swimming class URL
Raw	@98PXYRochester HEYYYYYYYYY!! its Fer from Chile again
Normalized	USER_MENTION heyy its fer from chile again
Raw	Sometimes, You gotta hate #Windows updates.
Normalized	sometimes you gotta hate windows updates
Raw	@Santiago_Steph hii come talk to me i got candy :)
Normalized	USER_MENTION hii come talk to me i got candy EMO_POS
Raw	@bolly47 oh no :'(r.i.p. your bella
Normalized	USER_MENTION oh no EMO_NEG r.i.p your bella

There are two types of features from our dataset, namely unigrams and bigrams.Unigrams are single words or tokens in the the text. Bigrams occur in succession in the corpus.We have used TFIDF Vectorizer in order to calculate **tf-idf** of features in each model .After plotting, a frequency distribution of the unigrams and bigrams present in the dataset, top N unigrams and bigrams are selected for the analysis.The reason behind selecting top N most common unigrams and bigrams is to reduce the noise present in the dataset to enhance the results of models as most of the irrelevant features and noise is present at the end of the frequency distribution.Extracted features from the dataset are used for the training of the different models.



Frequency distribution graph for features

6. Models

6.1 Decision Tree

We use the DecisionTreeClassifier from sklearn.tree package provided by scikit-learn to build our model.For this model, we got 68.72 % accuracy on the test data.

6.2 Naive Bayes

We used MultinomialNB from sklearn.naive_bayes package of scikit-learn for Naive Bayes classification. We used Laplace smoothed version of Naive Bayes with the smoothing parameter α set to its default value of 1.For this model, we got 77.12 % accuracy on the test data.

6.3 Logistic Regression

We utilise the Logistic Regression classifier available in sklearn.We reported accuracy using different epochs. 76.8% accuracy is achieved using 10 epochs.On the other hand,using 100 epochs ,we got 77.4% accuracy on the test data.

6.4 SVM

We utilise the SVM classifier available in sklearn.Accuracies are calculated using Polynomial Kernel function and Linear Kernel function.Parameter C (slack variable) is set to value 0.1. 77.83% of accuracy is reported using Linear Kernel function.On the other hand,accuracy is decreasing using Polynomial Kernel function which is 59%

6.5 Neural Network

We used keras with TensorFlow backend to implement the Multi-Layer Perceptron model. We used a 1-hidden layer neural network.The sigmoid function is used in the model. 75.93% accuracy is achieved on the test data.

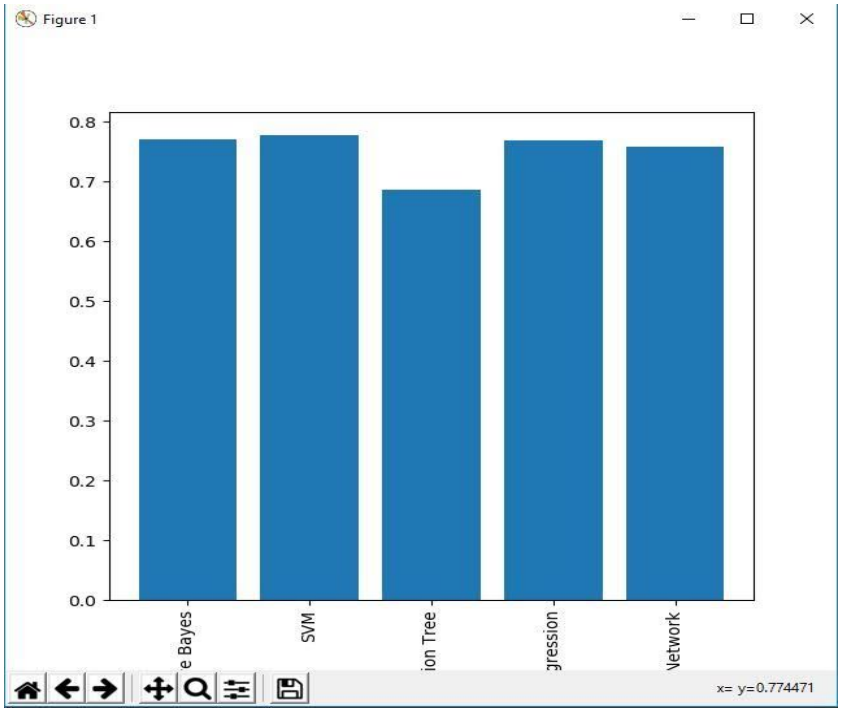
6.6 Testing on Real Time Data

Tweets on a particular topic are scrapped by using Twitter API.A specific number of tweets are scraped using user defined input.We have performed the same steps as mentioned in the preprocessing part on the scraped data in order to bring data in the suitable format for testing. In the first part of the implementation,we got highest accuracy using SVM.Sentiment prediction is performed on scraped preprocessed data with the help of pie chart.

7. Result

In SVM, accuracy decreases using Polynomial Kernel function.In Logistic Regression, accuracy increases as number of epochs increases.Highest accuracy 77.83% is achieved by using SVM model amongst Decision tree, Naive Bayes, Logistic Regression, SVM, and Neural Network.

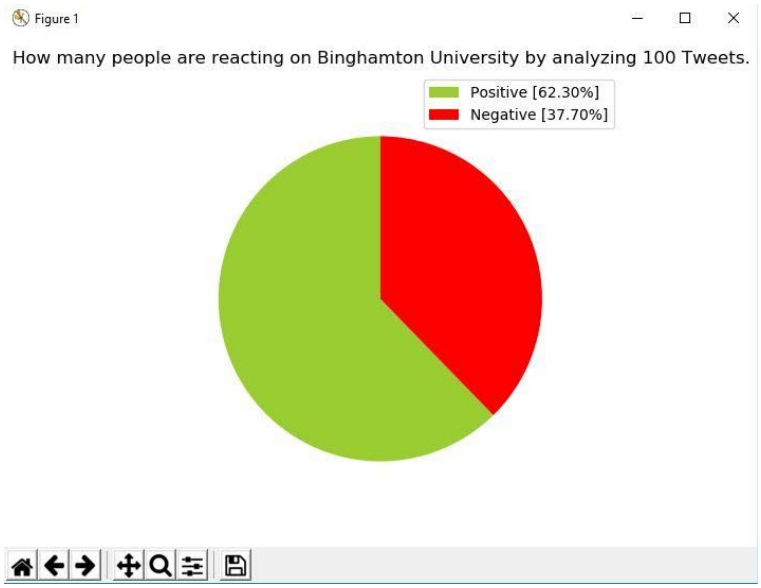
Accuracy Comparison:



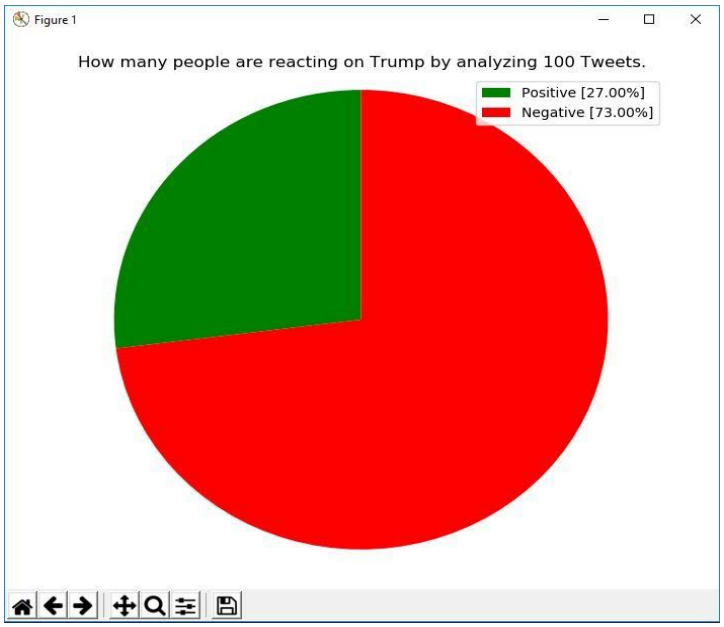
Accuracy for different models:

Model	Accuracy
Decision Tree	**68.72%**
Naive Bayes	**77.12%**
Logistic Regression	**76.8%** **(10 epochs)** **77.4%** **(100 epochs)**
SVM	**77.83%** **(Linear Kernel)** **59.00%** **(Polynomial Kernel)**
Neural Network	**75.93%**

Positive Sentiment Result for scraped data about Binghamton University



Negative Sentiment Result for scraped data about Trump



8. Learning/Key Takeaways

- By implementing “Twitter Sentiment Analysis” project ,we explored different supervised learning algorithms with scikit-learn and keras with tensorflow backend libraries.
- As we calculated tf-idf of each feature and plotted frequency distribution of features, we understood how tfidfvectorizer works in order to select first N important features by getting rid of irrelevant features and noise.
- This is how tf-idf is also useful to get rid of stopwords automatically.
- We got familiarized with when to use Sparse Vector representation or Dense Vector representation and its significance for feature engineering.
- Inclusion of bigrams as features is important to consider negations in the text data.
- We explored many research papers on the internet and as per our analysis,most of the research papers are with SVM as the highest accuracy giving algorithm.In this project,we came up with the same conclusion.
- Also, we understood the significance of Data-preprocessing in order to achieve the correct accuracy.
- We acknowledged, how sparse vector representation tackle with memory issues.

9. References

- Twitter Sentiment Analysis by Abdul Fatir Ansari, Abinaya Seenivasan, Anusha Anandan, Rakkappan Lakshmanan
- https://www.researchgate.net/profile/Shabib_Aftab/publication/321084834_Sentiment_Analysis_of_Tweets_using_SVM/links/5a1497b90f7e9b925cd514b0/Sentiment-Analysis-of-Tweets-using-SVM.pdf
- https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf
- https://thesai.org/Downloads/Volume8No6/Paper_3-Sentiment_Analysis_on_Twitter_Data_using_KNN_and_SVM.pdf
- Sentiment Analysis on Twitter Data Using Support Vector Machine Bholane Savita D., Prof.Deipali Gore