Twitter Sentiment Analysis

Group 23a
CS365A- Project Presentation
Ajay Singh (12056)

Sentiment Analysis

- **Sentiment analysis** (also known as opinion mining) refers to the use of natural language processing, text **analysis** and computational linguistics to identify and extract subjective information in source materials.
- Consumers can use sentiment analysis to research products and services before a purchase. Production companies can use the public opinion to determine acceptance of their products and the public demand. Movie-goers can decide whether to watch a movie or not after going through other people's reviews.

Twitter Sentiment Analysis

- Traditionally, most of the research in sentiment analysis has been aimed at larger pieces of text, like movie reviews, or product reviews. Tweets are more casual and are limited by 140 characters.
- However, this alone does not make it an easy task (in terms of programming time, not in accuracy as larger piece of text tends to be correctly classified) as people rarely give a second thought before posting a tweet. Grammar and content both suffer at the hands of the tweeter.
- The presence of a large dataset is always recommended (for better training of the classifier) and twitter makes it possible to obtain any number of tweets during a desired period. However, various difficulties are faced during processing of raw tweets. (Discussed in coming slides)

Previous Work

- Alec Go, Richa Bhayani and Lei Huang (Students at Stanford University) have done some serious work in twitter sentiment analysis.
- Even though their source code is not publicly available, their approach was to use machine learning algorithm for building a classifier, namely Maximum Entropy Classifier.
- The use of a large dataset too helped them to obtain a high accuracy in their classification of tweets' sentiments. The data set used by them is however public and I too have used the same data set in order to obtain results as close to theirs as possible. Other noteworthy works are by Laurent Luce and Niek Sanders. Both of them used quite smaller datasets, but their work consisted of some insightful approaches.

Challenges

- **Usernames** are mentioned more often than not. Usually they consist of some alphabets and numbers, and do not contribute much towards sentiment classification, except for increasing the size of the feature vector.
- **URLS** too are not required in our task.
- **Repeated letters** People often repeat letters in some words, in order to stress upon a particular emotion. For example:- sad, saaaad, saaaddd. All of them mean the same, yet it is not possible to distinguish between them if guided only by their spellings.
- **Hashtags** Words in hashtags may be read different from the same word without the hash tag
- **Punctuations and additional spaces**.

Preprocessing of tweets.

- All tweets were converted to **lower case**
- All **links and urls** were replaced by generic word URL
- All **usernames** were replaced by generic word USER
- Words with **hashtags** were replaced with the same words without the hashtag
- **Punctuations and additional white spaces** were removed from the tweets.
- All the above work was done in python via regular expression matching. The code for preprocessing will be uploaded along with the main code.

Dataset

- Dataset used in this project is publicly available and can be found at:

http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

It consists of 80,000 positive and 80,000 negatively classified tweets based on the emoticons used by the user.

☺ :-) :) :D =) were used to mark tweets with positive sentiment. ☹ :- (: (were used to mark tweets with negative sentiment.

Approach

- Filtering for Feature Vector

1.) Stop words such as a, an, is, the, you, she, he, it, they etc are removed as they do not indicate any sentiment.

2.) Words starting with anything but alphabets were removed for simplicity sake.

3.) Punctuation and repeating words were removed as they do not serve any purpose.

Naive Bayes Classifier

- Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

  Source: Wikipedia article

Maximum Entropy Classifier

- The Maximum Entropy (MaxEnt) classifier is closely related to a Naive Bayes classifier, except that, rather than allowing each feature to have its say independently, the model uses search-based optimization to find weights for the features that maximize the likelihood of the training data.
- The features you define for a Naive Bayes classifier are easily ported to a MaxEnt setting, but the MaxEnt model can also handle mixtures of boolean, integer, and real-valued features.

  Source: maxent

Support Vector Machines

- A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks.
- Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

  Source: Wikipedia article

Results

Using unigrams as features, Accuracy of:

Naïve Bayes Classifier – 76%

Maximum Entropy classifier – 75.4%

Support Vector Machines – 76.9%

Conclusion

- Even though unigram feature extractor is the simplest, it fails to identify negations. Using bigrams will help a lot in increasing the accuracy of the classifier
- Presence of neutral tweets too causes a dip in the accuracy

Future Work

- Neutral tweets: The current classifier does not consider neutral sentiments, even though many tweets do not exhibit a clear cut positive or negative emotion, especially the ones stating a fact or news.
- Bi-grams in combination with unigrams to handle negations like “not happy”
- Semantics may be employed when sentiment of a tweet depends on the perspective of the reader. For example: “India lost to Australia in the semis ☹” indicates negative sentiment for India, but positive for Australia.

References

- sentiment140, Go, Bhayani and Huang, Stanford University.
- Twitter sentiment classier using Python and NLTK Laurent Luce.
- Naive Bayes Classier Jacob Perkins.
- Twitter Sentiment Niek Sanders