Twitter Sentiment Analysis

Rucha Dubbewar
A20373886

Mithil Amin
A20386345

Problem Overview:

Twitter is a social networking service and an online news where users interact with messages and post events and news, known as “tweets”. Twitter can be used for business in expansion of potential as well as to express feeling of an individual or thought of particular event or a subject.

In this report, we will attempt to conduct sentiment analysis on “tweets” using various different machine learning algorithms. We attempt to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label. We use various machine learning algorithms to conduct sentiment analysis using the extracted features. However, just relying on individual models did not give a high accuracy so we pick the top few models to generate a model ensemble. Ensembling is a form of meta learning algorithm technique where we combine different classifiers in order to improve the prediction accuracy.

Data Description:

Live stream twitter data will be used. Data will be collected using Twitter API. Problems related with data collected will mostly be refining it and modifying it in such a way to make it useful for analysis. The train dataset is a csv file of type tweet_id, sentiment, tweet where tweet_id is unique integer identifying the tweet and sentiment is either positive(1) or negative (0). Similarly test dataset is a csv file of type tweet_id, tweet. This data collected contains words, user mentions, emoticons, URLs, unigrams and bigrams. User profile used for this project is Chicago cubs.

Method:

We perform experiments using various different classifiers. We use 10% of the training dataset for validation of our models to check against overfitting. For Naive Bayes, Maximum Entropy, Decision Tree, Random Forest, we use sparse vector representation of tweets. Existing libraries will be used for above stated machine learning algorithms.

For a baseline, we use a simple positive and negative word counting method to assign sentiment to a given tweet. We use dataset of positive and negative words to classify tweets. In cases when the number of positive and negative words are equal, we assign positive sentiment. Using this baseline model, we would try to achieve most accuracy.

Intermediate Experiments:

Project is in feature extraction and feature representation level. We extract two types of features from our dataset, namely unigrams and bigrams. We create a frequency distribution of the

unigrams and bigrams present in the dataset and choose top N unigrams and bigrams for our analysis.

Related Work:

Sentiment analysis: capturing favorability using natural language processing: This paper illustrates a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead of classifying the whole document into positive or negative.

Sentiment analysis of Twitter data: Contributions of this paper are POS-specific prior polarity features, use of a tree kernel to obviate the need for tedious feature engineering.

Opinion mining and sentiment analysis on a Twitter data stream: This paper discusses an approach where a publicised stream of tweets from the Twitter microblogging site are preprocessed and classified based on their emotional content as positive, negative and irrelevant; and analyses the performance of various classifying algorithms based on their precision and recall in such cases.

A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle: This paper describes a system for real-time analysis of public sentiment toward presidential candidates in the 2012 U.S. election as expressed on Twitter, a micro-blogging service. Sentiment analysis can help explore how these events affect public opinion. While traditional content analysis takes days or weeks to complete, the system demonstrated here analyzes sentiment in the entire Twitter traffic about the election, delivering results instantly and continuously. It offers the public, the media, politicians and scholars a new and timely perspective on the dynamics of the electoral process and public opinion.

Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis: This paper presents a new entity-level sentiment analysis method for Twitter using lexicon based approach to perform entity level analysis.

Who does what:

Till now both authors have equal participation in project ideas, data collection, data featuring as well as categorization of data.

Timeline:

Remaining steps we plan to complete are building classifiers and predictin accuracies to estimate the best classifier. We plan to complete them in a month.

References:

- Sentiment analysis: capturing favorability using natural language processing - https://dl.acm.org/citation.cfm?id=945658
- Sentiment analysis of Twitter data - https://dl.acm.org/citation.cfm?id=2021114

- Opinion mining and sentiment analysis on a Twitter data stream- http://ieeexplore.ieee.org/abstract/document/6423033/
- A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle - https://dl.acm.org/citation.cfm?id=2390490
- Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis - https://search.proquest.com/openview/3bb4518ced02bbd25b28594d28b35094/1?pq-origsite=gscholar&cbl=2029261