Project 2: Sentiment Analysis

Prepared by:
Kartikey Pradhan
Jay Dave

University of Illinois Chicago

Fall 2015

Table of Contents

1 Abstract

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". These tweets enable people to share and express their views about topics, or post messages. There has been a lot of work in the Sentiment Analysis of twitter data. This project involves classification of Obama and Romney tweets into two main sentiments: positive and negative (and neutral tweets). In this project we demonstrate, the use of features such as unigram, bigram, and effects of data pre-processing like filtering, stop words removal and stemming. Main classifier algorithms used are Naive Bayes multinomial and Multivariate Bernoulli model. As we shall see in the sections below, Naive Bayes multinomial with features of unigrams, bigrams and stemming, outperforms Multivariate Bernoulli model.

2 Introduction

Sentiments analysis is the task of finding opinions and proclivity of people towards specific topics of interest. This project concentrates on a political topic of interest between Obama and Romney. Sentiments of tweets can be categorized into various categories like positive, negative, neutral, extremely positive, extremely negative, and so on. The three types of sentiments considered in this classification experiment are positive, negative and neutral sentiments. The data in social networking services are sourced by humans and has lots of noise, thus gets difficult to achieve good accuracy.

Currently, the best results are obtained by Naive Bayes multinomial for a feature set containing filtering, stop words removal, stemming, and Bigram that gives an accuracy of 57.56. The main algorithms used in this project are Multivariate Bernoulli model and Naive Bayes multinomial and we would be comparing these in the upcoming sections.

3 Technique

The main approaches involved in this project are the various data pre-processing steps, the machine learning classifiers and feature extraction. The main machine learning algorithms used are Multivariate Bernoulli model and Naive Bayes multinomial. The main data pre-processing steps include Duplicates or repeated characters filtering, stop-words removal and stemming. Feature extraction includes unigram and bigram (all the above in various combinations) and numeric features all of which are described below.

3a Data Pre-processing

(a) Filtering

Duplicates or repeated characters

People use a lot of casual language on twitter. For example, 'hey' is used in the form of 'heyyyyyy'. Though this implies the same word 'hey', the classifiers consider these as two different words. To improve this and make words more similar to generic words, such sets of repeated letters are replaced by two occurrences. Thus heyyyyyy would be replaced by hey.

Removal of unnecessary characters

We also removed various unnecessary elements from each tweets before making unigram/bigram bag of words. We removed the html tags, additional white spaces, remove underscores, remove unicode and emoticons, remove hash, and remove special symbols, replaced punctuation. After removing all above junk form the tweet, then we created bag of words using unigram and bigram encoding.

(b) Stop-words removal

In information retrieval, there are many words that are added as conjunctions in sentences. For example, words like the, and, before, while, and so on do not contribute to the sentiment of the tweet. Also these words do not help in classifying the tweets as they appear in all classes of tweets. These words are removed from the data so as to avoid using them as features. The stopwords corpus was obtained from NLTK. Some modifications were required to this as the corpus also had some negative words such as nor, not, neither which are important in identifying negative sentiments and should not be removed.

(c) Stemming

In information retrieval, stemming is the process of reducing a word to its root form. For example, stems, stemmer, stemming all these words are derived from the root word stem. Hence, the stemmed form of all the above words is stem. NLTK provides various packages for stemming such as the PorterStemmer, LancasterStemmer and so on. The PorterStemmer was used in this project which uses various rules for suffix stripping. In addition to stemming the train and test data, the positive, negative and neutral word corpus was also stemmed. Stemming reduces the feature space as many derived words are reduced to the same root form. Multiple features now point to the same word and hence it increases the probability of the word.

By stemming, different derived words are mapped to their root words and this allows more matching between the tweets in the test and training set.

3b Classification Algorithms used

(a) Naive Bayes Multinomial model

The Naive Bayes classifier is one of the basic text classification algorithms. It is a simple classifier based on Bayes theorem and makes naive independence assumptions of the feature variables. Despite this very naive assumption, it is seen to perform very well in many real-world problems.

By Bayes theorem, we have,

$$P(Y|X_i) = \frac{P(X_i|Y)P(Y)}{P(X_i)}$$

Using Bayes theorem in the previous equation, we can find the probability of predicting the class Y given the features Xi. The class that gives the maximum probability that the given features predict it, is the class that the tweet will belong to. In this experiment, the Naïve Bayes Classifier from NLTK was used to train and test the data.

(b) Multivariate Bernoulli model

An alternative to the multinomial model is the multivariate Bernoulli model or Bernoulli model. It is equivalent to the binary independence model, which generates an indicator for each term of the vocabulary, either 1 indicating presence of the term in the document or 0 indicating absence. The Bernoulli model has the same time complexity as the multinomial model.

3c Feature Extraction

(a) Unigram

Unigrams are the simplest features that can be used for learning tweets. The bag-of-words model is a powerful technique in sentiment analysis. This technique involves collecting all words in the document and using them as features. The features can either be the frequency of words, or simply 0s and 1s to indicate if the word is present in the document or not. In this project, 0s and 1s are used to indicate the absence or presence of a word in the tweet.

(b) Bigram

Bigrams are features consisting of sets of two adjacent words in a sentence. Unigram sometimes cannot capture phrases and multi-word expressions, effectively disregarding any word order dependence. For example, words like 'not happy', 'not good' clearly say that the sentiment is negative, but a unigram might fail to identify this. In such cases, bigrams help in recognizing the correct sentiment of the tweet.

3d Tools

(a) Natural Language Toolkit

The NLTK is platform for building python programs to work with text data. It provides a variety of corpora and resources and various libraries for text classification, tagging, stemming, tokenization and parsing. In this project, NLTK was used extensively for tokenizing (tokenizing the tweets).

(b) Sci-kit Learn

Scikit-Learn is an open source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, logistic regression, naive Bayes, random forests, gradient boosting and k-means, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

4 Evaluation

4a Naïve Bayes Multinomial Distribution

(a) Romney Tweets

	Negative	Neutral	Positive
Precision	69.24	45.85	49.83
Recall	65.73	41.24	60.72
F1-Score	67.43	43.41	53.48
Accuracy	57.56		

(b) Obama Tweets

	Negative	Neutral	Positive
Precision	55.20	55.49	53.46
Recall	64.09	41.72	58.78
F1-Score	59.31	47.62	55.99
Accuracy	54.70		

4b Multivariate Bernoulli model

(a) Romney Tweets

	Negative	Neutral	Positive
Precision	68.52	44.39	48.77
Recall	60.27	48.55	61.09
F1-Score	65.38	46.37	54.14
Accuracy	57.01		

(b) Obama Tweets

	Negative	Neutral	Positive
Precision	60.67	56.86	46.54
Recall	50.42	37.86	74.00
F1-Score	55.06	45.44	57.14
Accuracy	53.08		

5 Conclusion

Naïve Bayes multinomial distribution has better performance than Bernoulli model. Multinomial distribution has better precision and accuracy for Romney tweets for all, positive, negative and neutral classes. In case of Obama tweets, the Bernoulli model has better precision for negative class, but for neutral and positive class, multinomial model outperforms Bernoulli model and hence has a better accuracy than Bernoulli model.