Introduction:

Social networking service is a platform to build social relations among people.Twitter is one of the significant service that allows the user to share their thoughts and information and to get in touch with other friends.

Film Industry is one of the biggest industry where large amount of money is invested. Though the expense involved in making movies are large, all the movies are not successful. The success and failure of a movie varies on many factors. Based on the reviews of people, ratings of a film can be detertmined. There are lot of systems for movie ratings and to analyse block buster hit. For instance, an actor may want to know how does their promotion news spread and the individual would like to care about how popular they are among public. Public review is more important for them to improve their skills and to go a long way in their career.

So we came with an idea to analyse the tweets in twitter about an actor and to predict their popularity specific to regions in the United States of America. Analysis is done by sentiment analysis, one of the profound research area for prediction and classification. Different machine learning methods are used to find the accuracies. The accuracies vary for each method and it is compared. The popularity of the actor in a region can be determined from the number of positive tweets. These data reveal the current popular trend through different perspectives. Futhermore , this model can be used for various purpose. The popular actors can be used for any type of promotions and awareness among people in that region. The collection of a particular actor's movie in that region can also be determined. It can be used for long term tracking of popular actors.

We assume that people who post the tweets know about the actor and post their true feelings about them. These tweets are used for the sentiment analysis. Some attributes like location, images, videos, retweets will affect the total tweets and the predictions.

Data:

Actors text file was manually created with 10 actors name. Regions was also classified manually into Midwest, Northeast, Southeast, Southwest, West and saved as text files.The tweets about the actors are collected from www.twitter.com. The tweets for these actors are collected using tweepy. Since the popularity of the actor in each region has to be determine location of the users were considered.

Tweepy is open sourced that enables Python to communicate with Twitter platform and use its API. Jsonpickle library was used to collect the data as json. Tweets per query, max tweets to be downloaded, max –id , since –id were set to collect the tweets from Twitter. The max tweets were set to 10,000.

The actors names are passed as a parameter in the search query. Folders for each actor is created. The tweets which are written in English was collected for easy analysis and classification. The users tweets are checked by the users location. If the location of the user falls in any one of the regions set then the region folder is created for the actor and the tweet text is saved in the folder as a separate file. If the user location does not belong to any of the regions then it is neglected. This is done for all the actors.

By collecting tweets like this we had many difficulties in classification. Large amount of tweets were collected but a lot of retweets and duplicate tweets were present in the data. In order to collect unique tweets about the actors, the conditions for saving a tweet into text file was increased. So the retweet status is checked and only if it is null the tweets are saved in a text file. The duplicate tweets are checked using MD5Hash Algorithm, and the duplicates are removed. Finally after satisfying all the conditions the total number of files were nearly 3500.

Limitations:

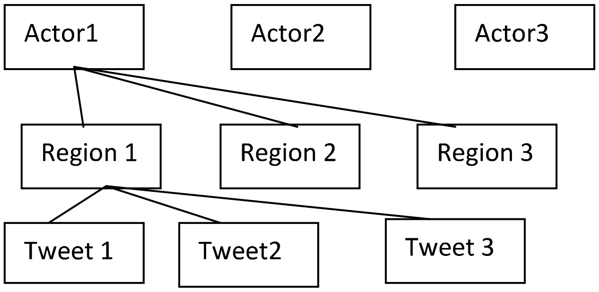
Though the duplicate tweet texts are removed using MD5 Hash Algorithm, some tweets may have the same text but a different URL in it. In such cases, the tweets are not duplicated and could not mark it as duplicate.

Method:

Data Collection and Processing Data:

3500 tweets was collected for actors based on user location. According to the user location the tweets are saved in separate text files.

The structure of it would be like the following figure:



These tweets has to be classified into positive, negative. The data is separated into training data set and testing data set.

Sentiment Analysis:

Training Data:

The collected tweet text files are manually classified into positive and negative for training data set. Since the training data set is to train the classifiers , some amount of collected tweets are taken and classified into positive and negative. 775 positive files and 546 negative files are present.

Testing Data:

Since the amount of data is large, manually categorising the tweets was a tedious process. The testing data set was also manually classified into positive and negative tweets. Neutral tweets are not considered.The testing data has 2262 files.

Tokenization:

The accuracy will be altered even because of tokenization. Tokenization is done for all the files. The URL's , actors names, stop word , white space and punctuations are eliminated for a better feature list. Based on it top positive and negative coefficients are also found.

Classifiers:

Logistic Regression:

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable.The logistic regression can be understood simply as finding the β parameters that best fit: y=1 if $\beta_0 + \beta_1 x + \epsilon > 0$; y =0, otherwise

Sklearn.linear_model.LogisticRegression is used as a classifier. The classifier is trained using the training data set by using Count Vectorise and Cross Validation methods.The testing accuracy for the testing data is found. The cross validation accuracy and testing accuracy are compared.

Gaussian Naive Bayes:

Gaussian implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

sklearn.naive_bayes. GaussianNB is used as a classifier. The classifier is trained using the training data set by using Count Vectorizer and Cross Validation methods.The testing accuracy for the testing data is found. The cross validation accuracy and testing accuracy are compared.

Count Vectorizer:

Convert a collection of text documents to a matrix of token counts. Various min-df and max-df values are considered and experiments have been done.

K-Folds cross validation:

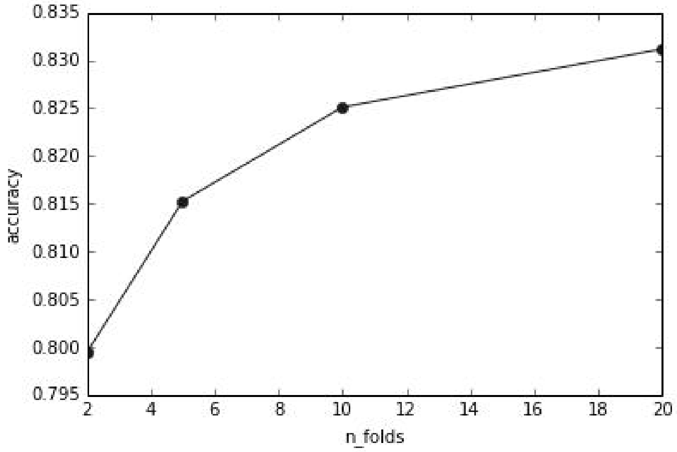
Provides train/test indices to split data in train test sets. Split dataset into k consecutive folds (without shuffling by default).Each fold is then used a validation set once while the k - 1 remaining fold form the training set. K- fold = 5 and the average cross validation accuracy is found. Various cross validation accuracy is found by differing the values of k-fold.

Prediction:

The popularity of actors in each region is simply determined by the number of positive tweets for the user in that region.

Experiment:

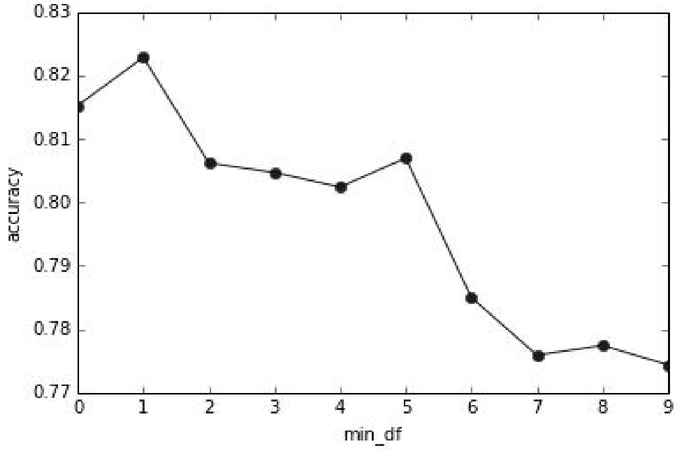
Logistic Regression:



n=2	accu =0.79939141796176583
n=5	accu=0.81526014865637497
n=10	accu =0.82510822510822501
n=20	accu=0.83117367706919953

For K-fold, the data is broken into K-blocks. Then, for K = 1 to X,the Kth block is the test block and the rest of the data becomes the training data. Bigger number of folds - bigger the training set and smaller the testing one. Increasing the k value decreases the variance and increases the bias- underfit.Decreasing the k value increases the variance and decreases the bias- overfit. After experimenting with the value of *k* the best accuracy is 0.815260

The average training accuracy is 0.8153



Min_df range(1-10)

1 acc= 0.81526014865637497

2 acc=0.82283018867924529

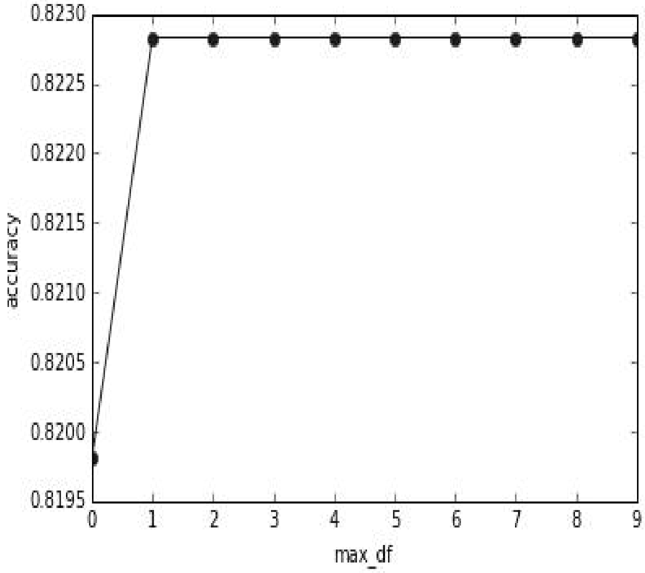
3 acc=0.80619496855345907

4 acc=0.80469125214408233

5 acc=0.80242424242424248

6 acc=0.806969696969697

For different min_df values the accuracies varies. There is an increase and decrease in the values.When building the vocabulary it ignore terms that have a document frequency strictly lower than the given threshold. This value is also called cut-off in the literature.If min_df = 2 it ignores all the terms which has a frequency less than the threshold.The number of terms decreases and only terms which occur more than the threshold value is taken into consideration for calculating accuracy.



Max_df range (.1-1)

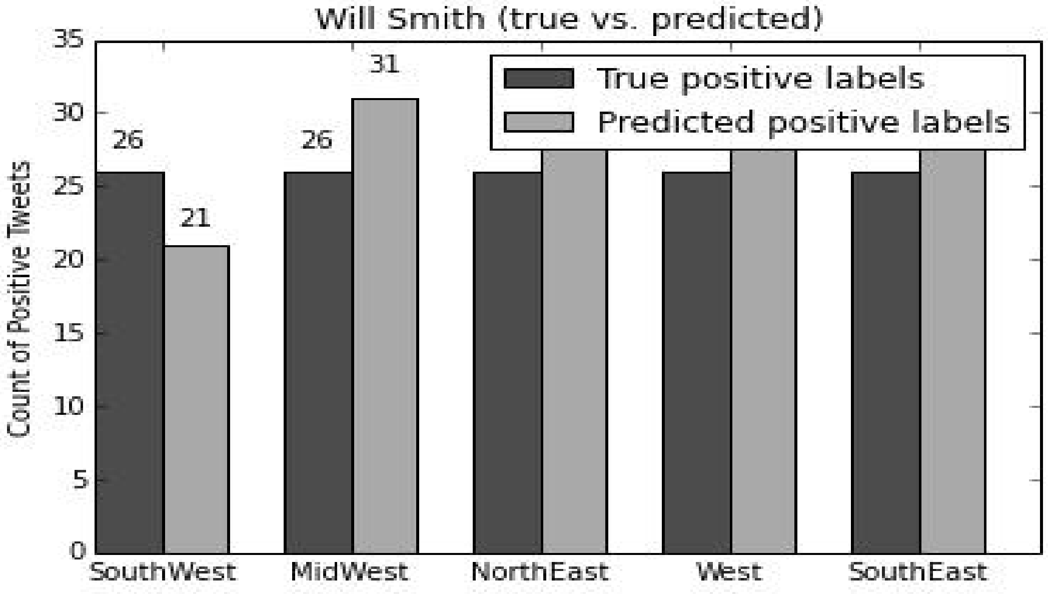
.1 acc=0.81981132075471697

.2 acc=0.82283018867924529

.3 acc=0.82283018867924529

For different max_df values the accuracy remains the same after certain point.

When building the vocabulary ignore terms that have a document frequency strictly higher than the given threshold (corpus-specific stop words). If float, the parameter represents a proportion of documents, integer absolute counts.



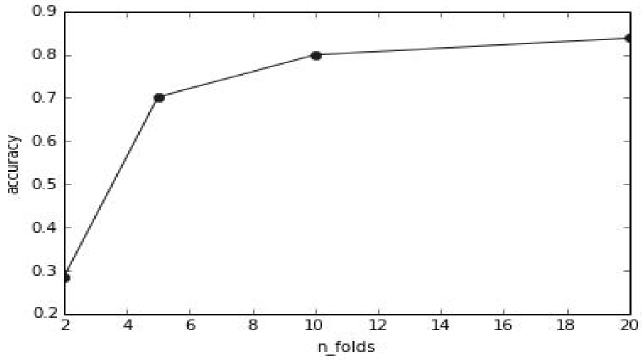
This is a sample graph which compares the true positive tweets vs predicted positive labels.

The testing accuracy is 0.82456432595361917

Our testing accuracy is 82.4%, which is pretty close to our estimated accuracy of 81.5%.

The popularity is found by the total number of positive tweets over negative tweets for each actor in each region and the popular actor in the region is finally displayed.

Gaussian Naive Bayes:



n=2 acc=0.2845482051998973469

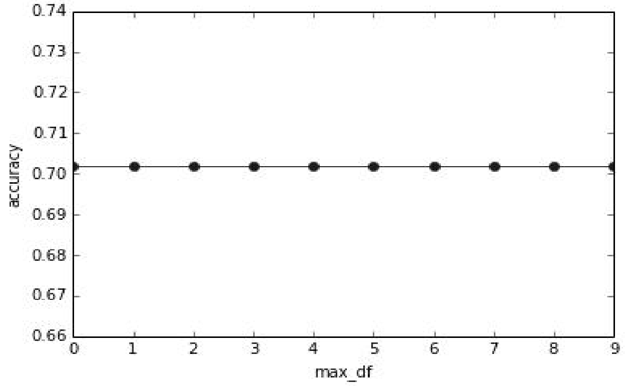
n=5 acc=0.70180388793596338

n=10 acc= 0.79943039416723616

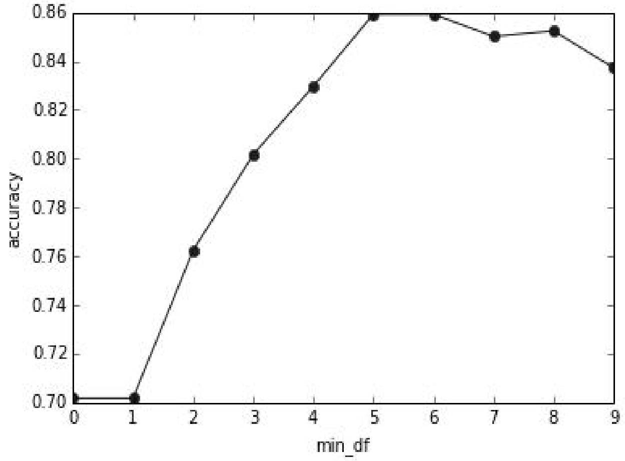
n=20 acc = 0.83799185888738137

For various n values the accuracies increases if n value is more.There is a decrease in variance and increase in bias. After analysing with various k-fold n=5 is assumed as best fit for our model.

The average training accuracy is 0.7018



For various values of max_df the accuracy remains the same for our model.



For various min_df there are increase and decrease in accuracies.

Result: When compared to both the models logistic regression worked well for us.But the accuracy value for Gaussian Naive Bayes was less. This may be due to some error in classification.

Related Work:

Several researchers have done sentimental analysis of social networks such as Twitters, Facebook. These works deal with the comments, tweets and other metadata collected from the social network profiles of users. Ratings of movies, Mobile Apps are also done based on Sentiment Analysis.

Automatically Detecting and Rating Product Aspects from Textual Customer Reviews[1] is a paper where the authors Wouter Bancken, Daniele Alfarone and Jesse Davis have followed a new approach to aspect- based sentimental analysis. . The goal of their algorithm is to obtain a summary of the most positive and the most negative aspects of a specific product, given a collection of free-text customer reviews.

Pang, Lee and Vaithyanathan [2] have performed sentiment analysis on movie reviews.They examined the hypothesis that sentiment analysis can be treated as a special case of topic-based text classification. The standard machine learning techniques such as Naive Bayes or Support Vector Machines (SVMs) is used and outperform manual classification techniques that involve human intervention. However, the accuracy of sentiment classification falls short of the accuracy of standard topic-based text categorization that uses such machine learning techniques.

Our approach is different from the first paper but its similar to the second one. We have done the sentiment analysis based on the standard machine learning techniques Logistic Regression and Gaussian Naive Bayes. The training data is classified manually and the classifier is trained. The cross validation accuracy and the testing accuracy is compared. The accuracy values are good since the positive and negative text are separated to a correct level and the error would have been less.

Conclusion and Future Work:

The sentiment analysis is done based on the Logistic Regression model and Gaussian Naive Bayes classifier and the accuracies are compared. The various experiments shows that the k-fold, min-df, max-df values affect the accuracies. The popularity of the actors are predicted based on the number of positive tweets in each region.

The sentiment analysis and prediction is done using users tweets. More data can be collected and training data set can be improved. Different classifiers should be used and the parameters k –folds, min-df, max-df values can be varied for each model and the comparison of accuracies should be done studied. Further a model can be built to predict the popularity of the actors. This system can then be used for various purposes like predicting the salary of the actors based on popularity.

Reference:

1.Bo Pang, Lillian Lee, and ShivakumarVaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002.

2. http://ceur-ws.org/Vol-1202/paper1.pdf