Twitter Popularity v/s Yelp Ratings: Predictive and Sentiment Analysis of Food Chains

Rekha K. Karangiya[1], Mehul C. Parikh[2], Rahevar Mrugendrasinh L.[3]

[1] *ME student, Department of Computer Science and Engineering, GEC Modasa, Gujarat, India*
[2] *Professor, Department of Computer Science and Engineering, GEC Modasa, Gujarat, India*
[3] *Professor, Department of Computer Science and Engineering, Charusat Changa, Gujarat, India*

ABSTRACT

Yelp connects people to great local businesses. Twitter is another important social media portal where twitter users express their experiences in a few sentences called tweets. This paper, focus on the reviews for restaurants. This model predict the star ratings for the restaurant from the Tweets extracted from Twitter. Through current work here, this paper perform a predictive analysis between the Yelp ratings of among popular food chains vs their equivalent twitter recommendations. Three machine learning algorithms are used with the sentiment analysis. After analyzed the performance of each models, the best model for predicting the ratings from reviews is the Naïve Bayes algorithm. Also, found that sentiment features are very useful for rating prediction.

Keyword: - *Sentiment, Rating, Classifier, Twitter, Yelp.*

1. INTRODUCTION

Today, the textual data on the net is growing at a rapid pace and creating any sense out of that data is a tedious task. Different industries are trying to use this huge textual data for extracting the people's views towards their products. Social media is a vital source of data in this case. It is impossible to manually analyze the large amount of data. This is where the necessity of automatic categorization becomes apparent [4]. Lots of research is done recently to automate this task of sentiment analysis of the data. Opinion mining, sentiment analysis, and subjectivity analysis are related fields sharing common goals of developing and applying computational techniques to process collections of opinionated texts or reviews. Sentiment analyses refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Twitter is an important social media portal where twitter users express their experiences in a few sentences called tweets [6].

Short message/tweet context analysis is a challenging task which has been investigated for several years. Traditional text mining techniques lose accuracy when applied to tweet mining. Among the challenges for tweet mining are: 1) limited information of fewer than 140 characters; 2) informal expressions and word variations caused by spelling errors, tweet slang and abbreviations. 3) Data volumes: only a tiny proportion of the whole tweet stream will be relevant to any given topic. Thus a major challenge is how to identify and analyze the tweets accurately, instantly and automatically. At this point, pre-processing [8] are applied.

Yelp is a huge website for food lovers that uses an active customer base to review restaurants worldwide. The reviewers write their comments and rate these restaurants and a generic rating (up to 5 stars) appears on the website for restaurants [7].

Table 1: Yelp Dataset Attributes

Name	Attributes
Business	Business Name, Id, Category, Location, etc.
User	Name, Review Count, Friends, Votes, etc.
Review	Date, Business, Stars, Text, etc.

The reason why Yelp data set is selected is that it has more information among the users, reviews and businesses that could be investigated for feature selections and models building to help much more accurate predictions.

Yelp is a heavily used portal by general public before we eat out at any restaurant. However, we seldom realize that these ratings are also dependent on the number of people writing the reviews. Thus, these ratings are highly sensitive to any single comments. The experiments were designed in a way to test the null hypothesis that Twitter is likely a good reflector of the general public consensus for popular food chains by capturing the positive/negative/neutral feelings of people expressed in tweets. Through current work here, this paper perform a predictive analysis between the Yelp ratings of 12 popular food chains vs their equivalent twitter recommendations. The sentiment analysis of the tweets is performed through techniques that have been acclaimed for their precision and accuracy in text classification like Support Vector Machines, Naive Bayesian and Maximum Entropy.

2. RELATED WORK

Sentiment Analysis has been one of fastest growing domain in Data Analysis which has widely accepted applications in social media monitoring and VOC to track customer reviews, survey responses, and competitors. This study uses Twitter Sentiment Analysis and we have analyzed some previous work that has been done in this field to build this models. This paper get a strong evidence from Kim et al. [1] to use Twitter data despite its small size (140 characters) to deduce the sentiment of a user writing it. Pang et al. [2] have demonstrated that machine learning through the three standard machine learning algorithms; Nave Bayes, maximum entropy (MaxEnt), and support vector machines (SVMs) can be used to analyze movie reviews and thus deduce the sentiments. Based on this three algorithms are used to build a classifier for the tweets and thus analyze the sentiment of the tweet as positive/negative/neutral.

As discussed in section 1, a lot of research has been performed on the text mining and sentiment analysis and thus a lot of methods have been identified to perform the classification. This paper chose primarily three methods for text classification because of their relative popularity and success in prediction of sentiments:

- **Naive Bayes:** This works on the assumption of conditional independence and despite this oversimplified assumption, Naive Bayes performs well in many complex real-world problems. Naive Bayes classifier is superior in terms of CPU and memory consumption.

- **Support Vector Machines:** SVM also provides a robust approach to build text classifiers and was picked because of its ability to handle High dimensional input space. When learning text classifiers, many (morethan10000) features can been countered. Since SVMs use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.

- **Maximum Entropy:** MaxEnt Naïve Bayes is based on conditional independence assumption, hence to ensure that this paper covers an alternative, it uses Maximum Entropy that does not assume conditional independence. It is based on the Principle of Maximum Entropy and from all the models that fit the training data, selects the one which has the largest entropy. Although it takes more time than Naïve Bayes to train

the model, this method has proven to be useful in cases where we do not know anything about the prior distribution.

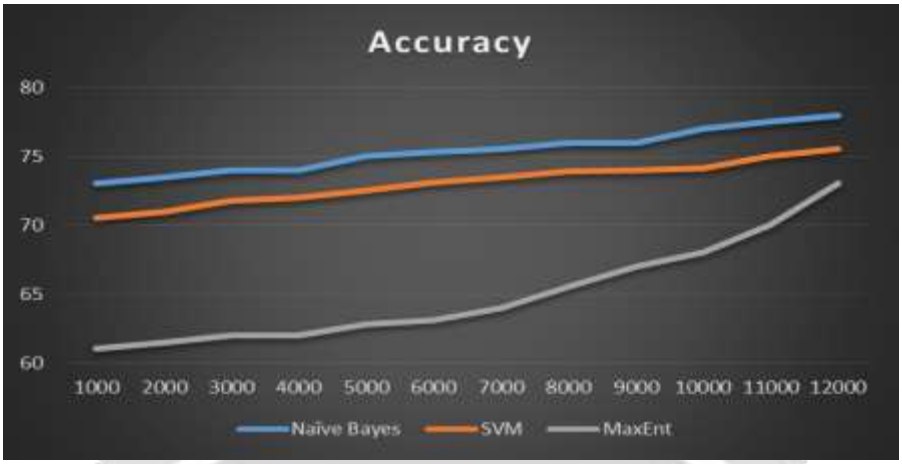

Chart -1: Comparison of Accuracy of Different Learning Algorithms

Chart -1 shows the comparison of Accuracy of different learning algorithms and the best accuracy was obtained by Naïve Bayes and this model is then used to predict the labels for other food chains as +1/0/-1.

Neutral Class: In real world applications, neutral tweets cannot be ignored. Proper attention needs to be paid to neutral sentiment [3].
It has been argued that the neutral class is requires in real world applications [3], including education [5]. Peoples may have positive, negative or neutral opinions. Discarding the neutral class and focusing on only on positive and negative opinions doesn't show a whole picture of the class opinion and will distort the true proportions of the positive and negative opinions when the neutral proportion is not known. Many of the researchers have enclosed the neutral category when analyzing data from the educational domain [5]. Consequently, the role of neutral class deserves additional investigation.

REST v/s STREAMING APIs:

- **REST APIs** are based on the REST architecture now popularly used for designing web APIs. These APIs use the pull strategy for information retrieval. To collect information a user must explicitly request it.

- **Streaming APIs** provides a continuous stream of public data from Twitter. These APIs use the push strategy for information retrieval. Once a re-quest for information is created, the Streaming APIs give a continuous stream of updates with no further input from the user

3. PROPOSED METHODOLOGY

The goal of this dissertation is to first extract live twitter data, then analyze these and extract root words in each tweet. The final sentiment analysis consists of calculating the numerical attributes like the sum of positive/negative/neutral terms. These counts are then scaled and compared to their Yelp ratings to present a correlation. In order to achieve the above, this paper designed an experiment following the below series of steps:

3.1: Data Collection

- In step 3.1 it uses twitter oAuth API to extract live twitter data.
- This data will be in JSON format. So json data will be retrive and store in database and then converted into CSV format.

3.2: Data Preprocessing

After storing data in database. Before done sentiment analysis pre process the data it includes

- Covert uppercase letter to lowercase letter.
- Removing stop words like (a,an,the,where…etc)
- Removing extra character like (, # $ % ^etc)
- Stemming process will be considered in which each word is reduced to its root words using Porter's Stemmer.

3.3: Feature Determination/Sentiment Labelling

- The data set that is achieve after pre-processing needed to be labelled with its respective category of a Positive/Negative/Neutral sentiment. So it used the counts of the positive and negative words in a tweet to categorize it to Positive/Negative/Neutral
- If the count of positive words is greater than the negative words, then the tweet is classifies as 1, else its value is -1, and it has value 0 for a tie.
- This data is then split into 70% and treated as training data and 30% which is treated as test data.
- This paper used Unigram model.

3.4: Training and Prediction

- This paper used 3 experiments to test for the classification to choose the best out of the three: I. SVM II. Naïve Bayes III. Maximum Entropy
- The labelled data was split into a 70-30 percentage as Training and Testing and each of the above mentioned methods were implemented to build 3 models. The test data was passed through these models and their accuracy was predicted. **10 fold cross validation** was used to validate these models.
- The **best accuracy** was obtained by **Naïve Bayes** and this model is then used to predict the labels for other food chains as +1/0/-1.

3.5: Rating Calculation

- For each food chain, a count of these classifiers is taken to get a total rating. It take a sum to ensure both positives and negatives are considered while calculation of the ratings. These are then scaled to a rating of 5 using the following:

$$X/N = Y/5 \text{ where}$$

X = Sum of +1/-1/0 for each food chain, N = Total tweets for the food chain, Y = equivalent rating to compare with Yelp.

4. PERFORMANCE STUDY

Dataset: For this experiments we extracted data for 12 popular food chains using their respective handles through twitter API and also downloaded dataset from the Yelp Academic Dataset Challenge Website.

System Setup: The experiments were conducted on laptop based on an Intel(R) Pentium(R) N3540 CPU@2.16GHZ processor with 4GB main memory and 500GB, The operating system as Windows10

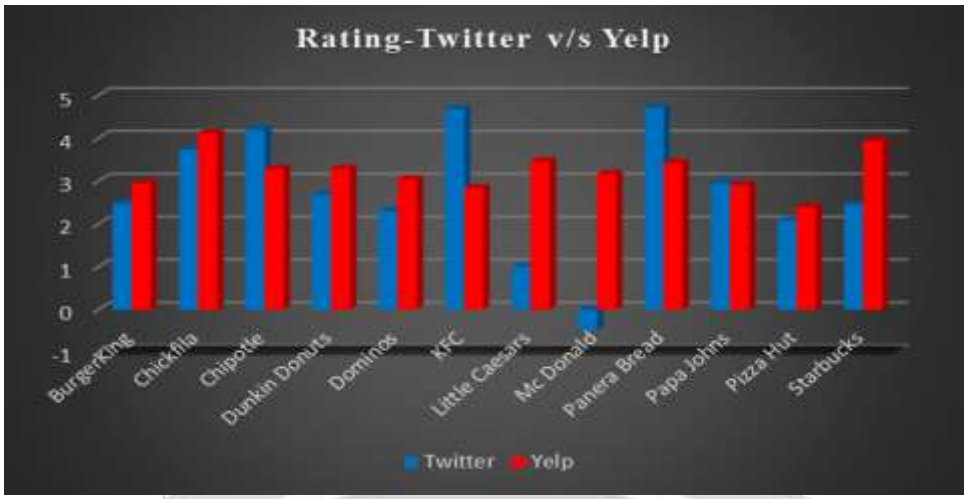

Chart -2: Final Comparison Graph between Twitter and Yelp

Chart -2 shows the comparison graph between Twitter and Yelp. After plotting the graph between the Twitter and Yelps Rating this model represents the difference in both the Ratings. For Example, Chipotle, KFC and Panera Bread is considerably more popular on Twitter as compared to Yelps. A drastic difference was seen in the food chains like Little Caesars and Mc Donald's, where the Twitter rating were not only less but went to negative.

5. CONCLUSION

High customer base of Twitter gives an improved representation of public sentiments and the most recent updates on the popularity in comparison to Yelp. This model can be further extended to perform a comparative analysis in areas like elections, movies, music, sports and so on. Further, the emoticons that are repeatedly used by twitter users can also be explored to determine the sentiments of the responses.

6. REFERENCES

[1]. D. Kim, Y. Jo, I-C. Moon, and A. Oh, Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users, Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHI 2010).

[2]. B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proc. Of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), July 2002, pp. 79-86.

[3]. A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009.

[4]. Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis"

[5]. A. Ortigosa, J. M. Martin, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," Computers in Human Behavior, vol. 31, pp. 527 – 541, 2014.

[6]. Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, AShehan Perera, Nadarajah Prasath, Thiruchittampalam Ragavan, "Opinion Mining and Sentiment Analysis on a Twitter Data Stream" International Conference on Advances in ICT for Emerging Regions Dec 2012.

[7]. Yun Xu, XINHUI Wu, QINXIA WANG "Sentiment Analysis of Yelp's Ratings Based on Text Reviews, Stanford University.

[8]. I. Hwmalatha, Dr. G. P saradhi varma, Dr. A. Govardhan "preprocessing the informal text for efficient sentiment Analysis" International Journal of Emerging Trends of Technology in Computer Science (IJETTCS), July-August 2012.