Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Sentiment Analysis for Social Media

Asst. Prof. A Kowcika*, Aditi Gupta, Karthik Sondhi, Nishit Shivhre, Raunaq Kumar
Department of CSE
R.V. College of Engineering
Bangalore, India.

Abstract— The proposed system is able to collect useful information from the twitter website and efficiently perform sentiment analysis of tweets regarding the Smart phone war. The system uses efficient scoring system for predicting the user's age. The user 'gender is predicted using a well trained Naïve Bayes Classifier. Sentiment Classifier Model labels the tweet with a sentiment. This helps in comprehensively analyzing the data based on various consumer parameters such as location, gender and age group.

Keywords— Sentiment Analysis, Social media, Twitter Streaming, Entity Extraction.

I. INTRODUCTION

Social media is a great medium for exploring developments which matter most to a broad audience and it is the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks. Social media technologies take on many different forms including magazines, Internet forums, weblogs, social blogs, micro blogging, wiki, social network, podcasts, photographs or pictures, video, rating and social bookmarking. Micro blogging websites have evolved to become source of varied kind of information. This is due to nature of micros blogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these micro blogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on micro blogs. Social media continues to gain increased presence and importance in society. Public and private opinions about a wide variety of subjects are expressed and spread continually via numerous social media, with twitter being among the timeliest. Social media has become one of the biggest forums to express ones opinion. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author when writing), or the intended emotional communication (the emotional effect the author wishes to have on the reader. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

II. LITERATURE SURVEY

Social media continues to gain increased presence and importance in society. Opinions about a wide variety of subjects are expressed and spread continually via numerous social media, with twitter being among the timeliest. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document and the effect the user wants to have on the reader. Sentiment analysis has become popular in judging the opinion of consumers towards various brands [1]. The way in which consumers express their opinion on social networking websites helps to judge this opinion [2]. The main issue is to understand this sentiment and being able to classify it appropriately [3]. The tweets are obtained from the twitter website using the twitter API. This will provide us with a large source of information for conducting the sentiment analysis [4]. Since data is being retrieved from a microblogging website an appropriate approach is to be used [5]. The tweets are first checked for relevance to Smartphones by using a list of keywords [6]. The system is trained using the training dataset which makes it capable to analyze the input tweets [7]. The input tweets are then checked word by word and the words expressing opinion are taken into account [8]. The sentiment analysis of the tweets is performed by the system [9] after which the tweets are classified into positive, negative and neutral categories [10]. Age prediction is performed on the data by the analyzing the similarity between the Name and the Screen Name, the user's description given on the twitter website and the previous tweets by the user [11].A list of keywords is provided for words commonly used by people below 30 and above 30 which are matched with these Parameters[12]. A scoring scheme is adopted which takes into account these three parameters and then according to a threshold score divides the person tweeting into above the age of 30 and below the age of 30 [13]. Predicting the gender of the person tweeting is also equally important as it helps understand the Smartphone market in a better way. This is done by using a training dataset of all the names being used in the United States in the past one twenty years [14]. This then trains the Naïve Bayes classifier to be able to predict the gender of the person tweeting [15]. The

system takes into consideration the last two characters, the last character and whether the last character is a vowel or not. The patterns observed for males and females with respect to these parameters are taken into account while making the prediction.

III. SENTIMENT ANALYSIS

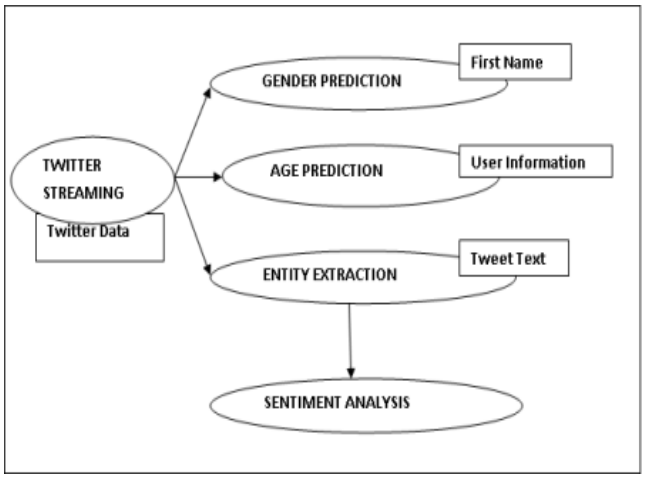

Fig. 1 Architecture

Fig.1 describes the entire architecture of the "Sentiment Analysis".

A. Twitter Streaming

The proposed system connects to the Twitter Streaming API using the developer's username and password. The streaming rules being one tweet per second. As soon as the username and password is authenticated, the program sends a list of keywords which acts as a filter of the stream of tweets and the result is the desired stream of tweets. The resultant stream of tweets is then checked for English language. If the tweet is not in English language then it is rejected.

B. Gender Prediction

The proposed system predicts the gender using the first name of the user. For the process of Gender Prediction, there is a training data set which has names from the past one hundred and twenty years labeled with their genders. The feature extractor extracts the last two alphabets, the last alphabet and checks whether the last alphabet is a vowel or not, from the name. Then a Naïve Bayes Classifier is trained using the extracted features. The first name of the user is then extracted from the twitter data and the trained classifier is used to classify the gender of the user. Fig.2 explains the methodology used for Gender Prediction.

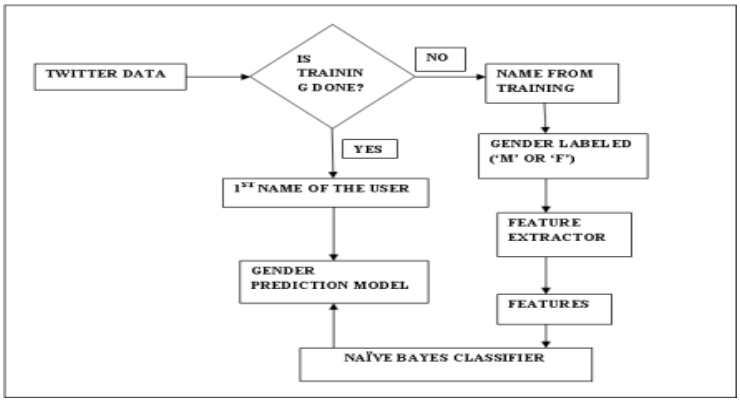

Fig.2 Gender Prediction

C. Age Prediction

Fig.3 shows how the Age Prediction Model works. Firstly, the screen name of the user and the full name of the user are compared and the ratio of the longest common sub-sequence ratio of the two is found out. The longest sub-sequence problem is to find the longest sub-sequence common to all sequences in a set of sequences (often just two). A sub-sequence is a sequence that can be derived from another sequence by deleting some elements without changing the order of the remaining elements. The longest common subsequence ratio between a name*n* and a screen name *sn*is the ratio of their L.C.S. to the maximum of the lengths of the two. The ratio is returned as *simScore*.

$$LCSR(n,sn) = \frac{LCS(n,sn)}{\max(lengths(n,sn))} \qquad \dots(1)$$

It has been observed that people who are more mature put their full name as their screen name. Secondly, the user description is checked with a list of keywords used most often by youngsters and old people. When in the description, if any keyword matches from either of the two lists, then the score is given in the favour of the matching list. The score for description, *desScore*, is 1 if there are more matches in the list for youngsters and 0 if there are more matches in the list for older people. And thirdly, the previous 100 tweets of the user are obtained using Twitter User timeline API and checked with the list of keywords used by younger people. The score, *slangScore*, is given infavour of the user being a youngster if there is a matching keyword. Finally all these scores are integrated and checked with a threshold score using equation (2). If the score is more than the threshold score then the age of the user is classified as below 30, otherwise it is classified as above 30.

$$ageScore = \frac{(3 * desScore + 2 * slangScore + 0.75 * simScore)}{6} \quad \ldots (2)$$

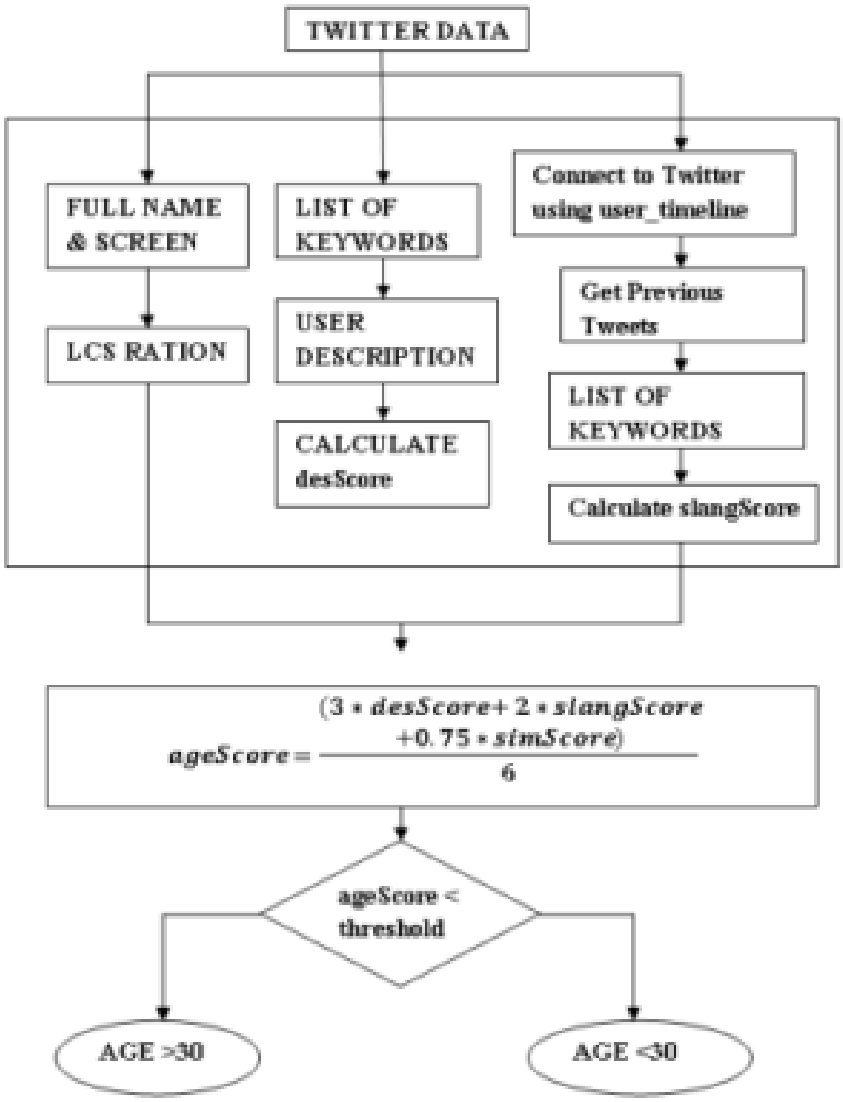

Fig.3 Age Prediction

D. Entity Extraction and Sentiment Analysis

Fig.4 is the model of Entity Extraction. Data from stream is the input for the Entity Extraction module. There is a list of entities specified with respect to this particular case. The data from stream is then matched with the list and if there is a match then that entity becomes the entity for that particular data. Now, sentiment analysis is performed. For this, the tweet goes through refining processes namely, Stop Word removal and Noise Cleaning. In Stop Word removal, words which don't add any meaning to the sentence, example of, the, is, are removed from the tweet. In Noise Cleaning, characters such as '@', '#' hashtags, extra white spaces and repeated characters are removed. After the refining process the features are extracted. The Naïve Bayes Classifier is trained for sentiment analysis with a training data set labeled with sentiments positive, negative or neutral. The dataset consists of around 1 million labeled tweets. Then the Sentiment Classifier Model labels the tweet with a sentiment. Using this trained Naïve Bayes classifier after performing the same processes of Stop Word removal, Noise Cleaning and feature extraction on the input tweet.

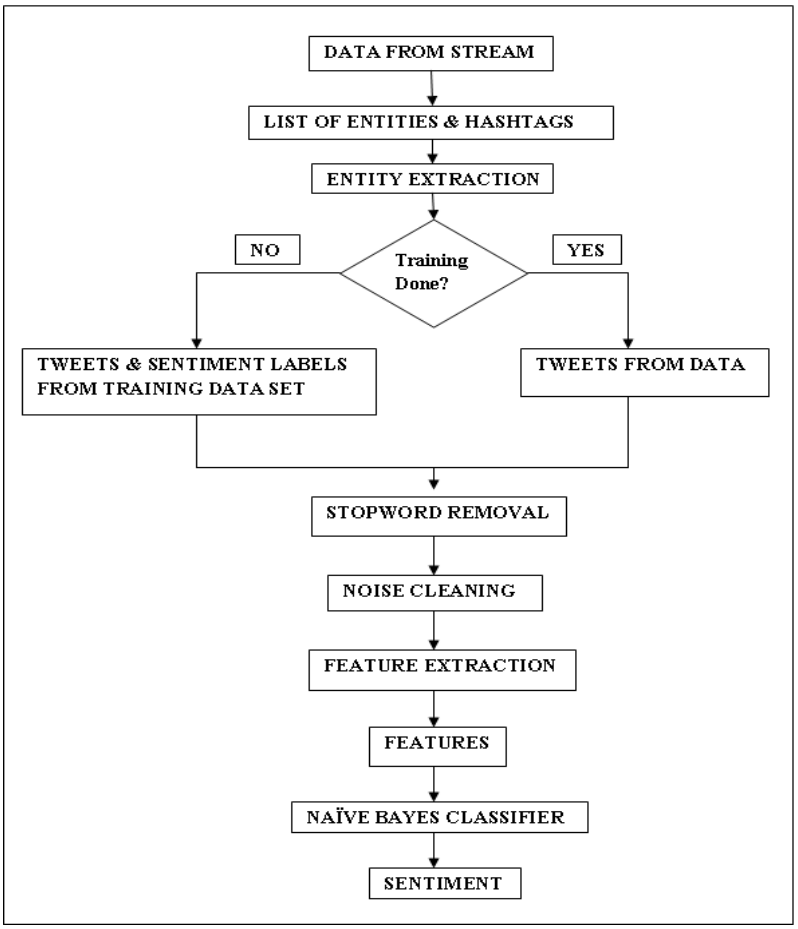

Fig.4 Entity Extraction and Sentiment Analysis

E. Experimental Analysis and Results

There are three basic functions being performed in our project which are sentiment analysis, gender prediction and age prediction. All three have their own different methods hence they are to be evaluated individually. The first being sentiment analysis where the tweets after being obtained from the twitter website are classified into positive, negative and neutralusing the Naïve Bayes classifier. However sometimes certain tweets maybe misinterpreted for this purpose evaluation has to be performed on the data. A Sample of 200 tweets is used which is then tested manually for accuracy. The second is age prediction where the age of the person tweeting is predicted on the basis of certain keywords being used in the description, keyword usage in the tweets, similarity between the display name and the actual name. A scoring scheme then allots a score depending on these parameters and hence predicts the age as below 30 or above 30. There can however be discrepancies as these are not definite methods of predicting a person's age and it is subject to dependence on person to person. The third being gender prediction done on the basis of the Name provided by the user on twitter. A training dataset of names for citizens of United States is used where the last two characters, the last character and if the last character is a vowel are tested. This dataset is used to train the Naïve Bayes classifier. This then performs gender prediction on the basis of patterns observed for males and females.

A threshold is set and the actual classification into positive, negative and neutral is done for sentiment analysis. Similarly, this approach is used for age prediction and gender prediction as well.

1) *Matching Matrix (Confusion Matrix)*: A matching matrix is a specific table layout that allows visualization of the performance of an algorithm, typically an unsupervised learning one (in supervised learning it is usually called a confusion matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

TABLE I
SENTIMENT ANALYSIS CONFUSION MATRIX

Predicted class → Actual Class ↓	**Positive**	**Negative**	**Neutral**
Positive	58	2	3
Negative	3	41	12
Neutral	5	7	69

TABLE III
GENDER PREDICTION CONFUSION MATRIX

Predicted class → Actual Class ↓	**Male**	**Female**
Male	56	11
Female	22	111

TABLE IIIII
AGE PREDICTION CONFUSION MATRIX

Predicted class → Actual Class ↓	**Above 30**	**Below 30**
Above 30	80	36
Below 30	13	71

2) Performance Measures:

		Predicted class		
		A	B	C
Known class (class label in data)	A	tp_A	e_{AB}	e_{AC}
	B	e_{BA}	tp_B	e_{BC}
	C	e_{CA}	e_{CB}	tp_C

- *Precision:* Precision is a measure of the accuracy provided that a specific class has been predicted.

  Precision=tp/(tp+fp)

  where tp and fp are the numbers of true positive and false positive predictions for the considered class. The result is always between 0 and 1.

In the matching matrix above, the precision for a class is calculated as:

- For sentiment analysis
  - Positive = 58/(58+2+3) = **0.92**
  - Negative = 41/(41+3+12) = **0.73**
  - Neutral = 69/(69+12+0) = **0.85**
- For gender prediction
  - Males = 56/(56+11)=**0.835**
  - Females= 111/(111+22)= **0.834**
- For age prediction
  - Above 30=80/(80+36) = **0.690**
  - Below 30 = 71/(71+13) = **0.845**

- *Accuracy:* Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications.
  - *Accuracy for sentiment=168/200 =0.840*
  - *Accuracy for gender prediction= 167/200 = 0.835*
  - *Accuracy for age prediction=151/200=0.755*

The models for sentiment analysis and gender prediction were also tested on their respective training datasets. The training dataset for sentiment analysis consists of around 1 million tweets labeled as '0', '2' and '4', where '0' stands for negative, '2' stands for neutral and '4' stands for positive. 10% of these tweets (approx. 10000) were used for testing. The Naïve Bayes Classifier used for sentiment analysis displayed an accuracy of 83.5%. The gender prediction model's

training dataset consisted of names from each of the previous 130 along with labels for male and female. 15% of this dataset, i.e., names from 20 years were used for testing this Naïve Bayes Classifier.

IV. CONCLUSIONS

Thus, the proposed system is able to collect useful information from the twitter website and efficiently perform sentiment analysis on the data and predict the user's age and gender using an efficient scoring system and a well trained Naïve Bayes Classifier, respectively. This helps in comprehensively analyzing the data based on various consumer parameters such as location, gender and age group. Also, the user friendly GUI makes this analysis efficient and easy.

REFERENCES

[1] Bernard J. Jansen, Mimi Zhang, Kate Sobel and AbdurChowdury,*Micro-blogging as online word of mouth branding*, 27th International Conference Extended Abstracts on Human Factors in Computing Systems, New York, 2009,pages 3859-3862.

[2] J.C. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou.*Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews*,AdvancesIn Knowledge andorganization, 2004, pages 49-54.

[3] R. Prabowo and M. Thelwall, *Sentiment analysis: A combined approach*, Journal of Informetrics, 2009, pages 143-157.

[4] OnurKucuktunc,B. BarlaCambazoglu, Ingmar Weber and HakanFerhatosmanoglu, *A large-scale sentiment analysis for Yahoo! Answers*, Fifth ACM International Conference on Web Search and Data Mining, Seattle, Washington, USA, 2012, pages 633-642.

[5] Adam Bermingham and Alan F. Smeaton, *Classifying Sentiment inMicroblogs:is brevity an advantage?*, Nineteenth ACM International Conference on Information and knowledge management, Toronto, Canada, pages 1833-1836.

[6] Ana-Maria Popescu, Marco Pennacchiotti, *Detecting Controversial Events from Twitter*, Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto, ON, Canada, 2010, DOI: 10.1145/1871437.1871751, pages 1873-1876.

[7] Pang and L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, 2008, pages 1-135.

[8] Eugene Agichtein, Luis Gravano, Jeff Pavel, Viktoriya Sokolova, and Aleksandr Voskoboynik, *Snowball: a prototype system for extracting relations from large text Collections*, Proceedings of the fifth ACM conference on Digital Libraries, Bremen.

[9] ApoorvAgarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneaua , *Sentiment Analysis of Twitter Data*, Proceedings of the Workshop on Languages Social Media in, Portland, Oregan, 2011, pages 30-38.

[10] Luciano Barbosa, JunlanFeng, *Robust sentiment detection on twitter from bayes and noisy data*, Proceedings of the 23rd International Conference on Computational Lisguistics, Beijing, China, 2010, page 36-44.

[11] Collin F. Baker, Charles J. Fillmore, and John B. Lowe, *The Berkeley framenet project*, In Proceedings of the 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada, 1998,pages 86–90.

[12] G. Forman, *An extensive empirical study of feature selection metrics for text Classification*, Journal of Machine Learning Research, vol.3, 2003, pages 27-35.

[13] DelipRao, David Yarowsky, AbhishekShreevats, Manaswi Gupta, *Classifying Latent User Attributes in Twitter*, In Proceedings of the 2nd International Workshop onSearch and Mining User-generated Contents (SMUC), Toronto, ON, Canada, 2010, DOI: 10.1145/1871985.1871993, pages 37-44.

[14] Bo Pang,LillianLee,ShivakumarVaithyanathan, *Thumbs up? Sentiment Classification using Machine LearningTechniques*, Proceedings Conference of Emperical methods in natural language, Association for Computational Linguistics Language Processing (EMNLP), University of Cornell ,Philadelphia, 2002, pages 79-86.

[15] Ashraf M.Kibriya, EibeFrank, BerhardPfahringer and Geoffrey Holmes, *Multinomial Naïve Bayes for text categorization revisited*, Proceedings of the 17th Australian Joint conference on Advances in Artificial Intelligence,Australia,2004, pages 488-499.

