International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 1, Volume 2 (January 2015) www.ijirae.com

Sentiment Analysis on Twitter Data

Varsha Sahayak Vijaya Shete Apashabi Pathan
BE (IT) *BE (IT)* *ME (Computer)*
Department of Information Technology, Savitribai Phule Pune University, Pune, India.

Abstract – Now-a-days social networking sites are at the boom, so large amount of data is generated. Millions of people are sharing their views daily on micro blogging sites, since it contains short and simple expressions. In this paper, we will discuss about a paradigm to extract the sentiment from a famous micro blogging service, Twitter, where users post their opinions for everything. In this paper, we will discuss the existing analysis of twitter dataset with data mining approach such as use of Sentiment analysis algorithm using machine learning algorithms. An approach is introduced that automatically classifies the sentiments of Tweets taken from Twitter dataset as in [1]. These messages or tweets are classified as positive, negative or neutral with respect to a query term. This is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product before purchase. We will use machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision which is discussed in [8]. The training data consists of Twitter messages with emoticons, acronyms which are used as noisy labels discussed in [4]. We examine sentiment analysis on Twitter data. The contributions of this survey paper are: (1) we use Parts Of Speech (POS)-specific prior polarity features. (2) We also use a tree kernel to prevent the need for monotonous feature engineering.

Keywords – Micro blogging, Twitter, Sentiment, Classifiers, Sentiment Analysis.

I. Introduction

We know that there are almost 111 micro blogging sites. Micro blogging websites are nothing but social media site to which user makes short and frequent posts. Twitter is one of the famous micro blogging services where user can read and post messages which are 148 characters in length. Twitter messages are also called as Tweets. We will use these tweets as raw data. We will use a method that automatically extracts tweets into positive, negative or neutral sentiments. By using the sentiment analysis the customer can know the feedback about the product or services before making a purchase. The company can use sentiment analysis to know the opinion of customers about their products, so that they can analyze customer satisfaction and according to that they can improve their product. Sentiment analysis has become one of popular research area in computational linguistics, because of the explosion of sentiment information from social web sites (i.e., Twitter and Facebook), online forums, and blogs as in paper [10].

We are going to use three models namely unigram model, tree kernel model and feature based model. Sentiment Classification has been researched for better result. Traditionally, Sentiment classification concentrated for classifying larger pieces of text which includes reviews or feedback. But in Twitter which includes tweets are different from reviews. Both Twitter and reviews are differentiated by their purpose. Tweeter's emotion or feeling on particular topic can be express by using tweets. While, summarized thoughts of authors are represented by reviews. On the other hand, tweets are more casual with the limited 140 characters text in length.

In paper [1], there is use of two resources : 1) a hand annotated dictionary for emoticons 2) an acronym dictionary gathered from web. The approach is the use of different machine learning classifiers and feature extractors. Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM) are the machine learning classifiers. Unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags are the feature extractors. In paper [1] and [2], one of the best uses of Sentiment Analysis is that the organization knows their own business progress by user's feedback. Sentiment Analysis is highly domain centered; the application developed for twitter can't be used for facebook. When looking at Twitter, it is particularly problematic. For example: "The meal was awesome but the service was terrible". In this case, computer gets confused for the result of sentiment.

Machine Learning Methods:
There are three different machine learning algorithms who achieved great success for text categorization as in paper [3] which are as follows:

1) Naive Bayes:
Naive Bayes model is a simplest model. For the categorisation of the text this model works well. Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. As in [6], it is made to simplify the computation and in this sense considered as "Naive".

Class c^* is assigned to tweet which is denoted by d,
Where, $c^* = argmacc P_{NB}(c|d)$

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)P(f|c)^{ni(d)})}{P(d)}$$

In this formula, f represents a feature and $ni(d)$ represents the count of feature fi found in tweet d. There are a total m features. Parameters $P(c)$ and $P(fjc)$ are obtained through maximum estimates, and add-1 smoothing is utilized for unseen features.

2) Maximum Entropy (MaxEnt):
This model is Feature based model. MaxEnt do not make any independence assumption for its features, therefore MaxEnt is different than Naive Bayes. MaxEnt can handle features overlapping problems better than Naïve Bayes. Stanford classifier is used for classification in MaxEnt model. In practical scenarios different types of problems can be resolved by MaxEnt easily as compared to Naive Bayes.

3) Support Vector Machines (SVMs):
Support Vector Machines are theoretically well motivated algorithms and has been developed from statistical learning theory since the 60s. The class of algorithms called SVMs which are used for pattern recognition. They are effective and famous classification learning tool. Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik. The SVM algorithm is based on the statistical learning theory and the Vapnik-Chervonenkis (VC) dimension introduced by Vladimir Vapnik and Alexey Chervonenkis. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression.

A few methods were devised and analysed because of centrality of the SVM optimization problem which are discussed in [9]. We can build models using Naive Bayes, MaxEnt and SVMs same as in [4] and [6].

Using these machine learning algorithms, three models are developed in Weka namely Unigram Model, tree kernel model and feature based model. These models will used for feature extraction.

As in paper [11] which presents SentiView tool. It is an interactive visualization system and it focuses on analysis of public sentiments for popular topics on the Internet. Uncertainty modeling and model-driven adjustment is combined in SentiView, it mines and models the changes of the sentiment on public topics, by searching and correlating frequent words in text data.

II. Proposed System

Sentiments are the words or sentences that represent view or opinion that is held or expressed that can be positive, negative or neutral. We are going to propose a novel hybrid approach involving both corpus-based and dictionary-based techniques, which will find the semantic orientation of the sentiments words in tweets. We will also consider features like emoticons, neutralization, negation handling and capitalization as they have recently become a huge part of the internet language.

The proposed Sentiment Analysis on twitter data is based on two important parts viz Data Extraction, pre-processing of extracted data and classification.

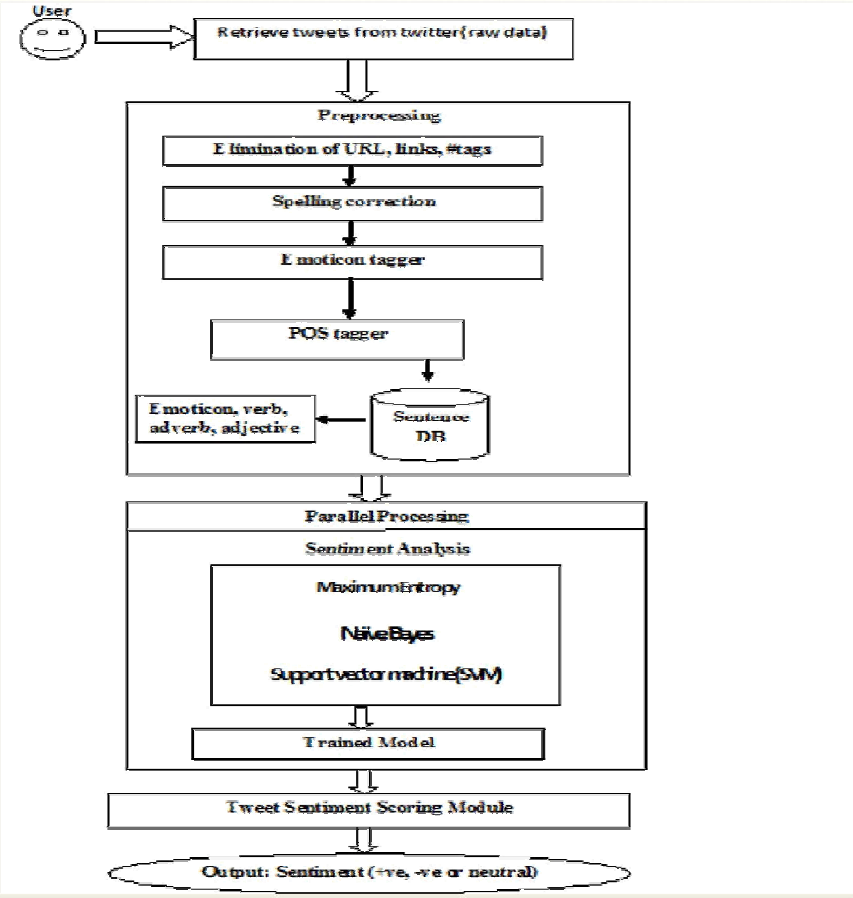

Fig (1): System architecture of proposed system

To uncover the sentiments, we will first extract the opinion words from tweets and then we find out their orientation, i.e., to decide whether each sentiment word reflects exaggerated and self-indulgent feelings of tenderness, sadness, or nostalgia.

The following steps will expound the process of the proposed system which is discussed in paper [2] and [6] shown in fig [1]:

1. Retrieval of tweets
2. Pre-processing of extracted data
3. Parallel processing
4. Sentiment scoring module
5. Output sentiment

These steps are explained below:

1. Retrieval of tweets :

As twitter is the most exaggerated part of social networking site, it consists of various blogs which are related to various topics worldwide. Instead of taking whole blogs, we will rather search on particular topic and download all its web pages then extracted them in the form of text files by using mining tool i.e. Weka which provides sentiment classifier..

2. Pre-processing of extracted data:

After retrieval of tweets Sentiment analysis tool is applied on raw tweets but in most of cases results to very poor performance. Therefore, preprocessing techniques are necessary for obtaining better results as given in [12]. We extract tweets i.e. short messages from twitter which are used as raw data. This raw data needs to be preprocessed. So, preprocessing involves following steps which constructs n-grams:

i) Filtering:

Filtering is nothing but cleaning of raw data. In this step, URL links (E.g. http://twitter.com), special words in twitter (e.g. “RT” which means ReTweet), user names in twitter (e.g. @Ron - @ symbol indicating a user name), emoticons are removed.

ii) Tokenization:

Tokenization is nothing but Segmentation of sentences. In this step, we will tokenize or segment text with the help of splitting text by spaces and punctuation marks to form container of words.

iii) Removal of Stopwords:

Articles such as “a”, “an”, “the” and other stopwords such as “to”, “of”, “is”, “are”, “this”, “for” removed in this step.

iv) Construction of n-grams:

Set of n-grams can make out of consecutive words. Negation words such as “no”, “not” is attached to a word which follows or precedes it. For Instance: “I do not like remix music” has two bigrams: “I do+not”, “do+not like”, “not+like remix music”. So the accuracy of the classification improves by such procedure, because negation plays an important role in sentiment analysis. Paper [3] represents that negation needs to be taken into account, because it is a very common linguistic construction that affects polarity.

3. Parallel processing:

Sentiment classifier which classifies the sentiments builds using multinomial Naïve Bayes Classifier or Support Vector Machines (SVMs). Training of classifier data is the main motive of this step. Every database has hidden information which can be used for decision-making. Classification and prediction are two forms of data analysis which can be used to extract models describing important data and future trends. Classification is process of finding

a set of models or functions that describe and distinguish data classes or concepts, for the purpose of being able to use the model for predicting the class of objects whose class label is unknown.

The derived model is based on the analysis of a set of training data. Training data consists of data objects whose class labels are known. The derived model can be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

Classification process is done in a two step process. First step is Model Construction in which we will build a model from the training set. And step2 is Model Usage in which we will check the accuracy of the model and use it for classifying new data.

4. Sentiment scoring module:

Prior polarity of words is the basic of our number of features. The dictionary is used in [1] in which English language words assigns a score to every word, between 1 (Negative) to 3 (Positive). So, this scoring module is going to determine score of sentiments in the sentiment analysis of data.

5. Output sentiment:

Based on the dictionary assignment of score, the proposed system interprets whether the tweet is positive, negative or neutral.

III. Conclusion

Twitter is a demandable micro blogging service which has been built to discover what is happening at any moment of time and anywhere in the world. In the survey, we found that social media related features can be used to predict sentiment in Twitter. We will use three machine learning algorithms which will contribute to outperform three models namely unigram, feature based model and tree kernel model by using Weka. So, our proposed system concludes the sentiments of tweets which are extracted from twitter. The difficulty increases with the nuance and complexity of opinions expressed. Product reviews, etc are relatively easy. Books, movies, art, music are more difficult. We can also implement features like emoticons, neutralization, negation handling and capitalization/internationalization as they have recently become a huge part of the internet.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, *"Sentiment Analysis of Twitter Data"* Department of Computer Science, Columbia University, New York, 2009.

[2] Akshi Kumar and Teeja Mary Sebastian, *"Sentiment Analysis on Twitter"* department of Computer Engineering, Delhi Technological University, Delhi, India, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[3] G. Vinodhini, R. M. Chandrasekaran *"Sentiment Analysis and Opinion Mining: A Survey"* Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002, Volume 2, Issue 6, June 2012, IEEE paper.

[4] Luciano Barbosa and Junlan Feng, "*Robust sentiment detection on twitter from biased and noisy data*." Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44, 2010.

[5] Adam Bermingham and Alan Smeaton, "*Classifying sentiment in microblogs: is brevity an advantage?"* ACM, pages 1833–1836, 2010.

[6] Pak and P. Paroubek. *"Twitter as a Corpus for Sentiment Analysis and Opinion Mining"*, In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010.

[7] R. Parikh and M. Movassate, *"Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques"*, CS224N Final Report, 2009

[8] Go, R. Bhayani, L.Huang. *"Twitter Sentiment Classification Using Distant Supervision", Stanford University,* Technical Paper, 2009

[9] Shai Shalev-Shwartz, Yoram Singer, Nathan, Srebro, Andrew Cotter *"Pegasos: Primal Estimated sub-GrAdient SOlver for SVM"*, 2000.

[10] Chuan-Ju Wangz, Ming-Feng Tsaiy, Tse Liuy, Chin-Ting Changzy, *"Financial Sentiment Analysis for Risk Prediction"* Department of Computer Science & Program in Digital Content and Technology National Chengchi University Taipei 116, 2013.

[11] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, *"SentiView: Sentiment Analysis and Visualization for Internet Popular Topics"* IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 43, NO. 6, NOVEMBER 2013.

[12] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, "Interpreting the Public Sentiment Variations on Twitter", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, MAY 2014

