ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Semantic Data Classification of Twitter Data

Gowtham A S, Yamuna B R

PG Student, M. Tech, CSE, Rajeev Institute of Technology, Hassan, India

Assistant Professor, Dept. of CSE, Rajeev Institute of Technology, Hassan, India

ABSTRACT*:* The best option to be stayed tuned with followers is through Micro blogging sites. Micro blogging is elevating their fame all over the online world. Truly micro blogging web sites are use for private or legit promotional reason. Twitter is one of the online social site that allows the users share the messages of about 140-character which are usually called as tweets. More than 500 million people use Twitter. Semantic analysis tries to determine the linguistic input meaning. In different words, language is processed to be able to produce knowledge of concerning the world. In this project we classify the tweet words into good or bad words. To accomplish this task we have carried classification which is helpful to analyze the information in the form of the no of tweets where opinions are highly unstructured and are either good and bad words .For this first we pre-processes the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector using TF-IDF, then select feature vector list and there after applied machine learning based twitter.

KEYWORDS*:* Micro blogging, Twitter, Classification, Feature Vector, TF-IDF.

I. Introduction

Day by day micro blogging is increasing its popularity over the world. The micro blogging sites are used for official and personal reason. The users of these sites updates their activities, sell their products, top news in the world and many more. The updates, breaking news, about the product are given within 140 characters in micro blogging sites. The people who do online business get more benefit from these sites. It helps in promoting a webpage or production, back link and drive visitor. It is without doubt one of the high-quality platform for online marketing.

The Benefits from the micro blogging sites are: 1) Provide good quality backlink.2) Connects with whole world 3) Stay tuned together with your followers and friends. 4) Get new ideas from invention, e-newsletter and research. 5) Be updated with patron, acquaintances, loved ones and follower 6) Create on-line fame 7) Force traveler on your website online (which you could get extra traffic with utilizing hash tag) 8) Worthwhile for getting social signal. 9) First-rate strategy for on-line product advertising and marketing 10) It's also a part of off web page see for an website.11) It upward thrust industry faith. 12) Industry relation grows up faster to a brand new customer.13) Develop brand cognizance.14) Instantaneous suggestions in your product or publications.15) Micro blogging websites giant enough to look tourist of your product or your site.

Twitter data analysis for current events, companies, products and people thus, leading a way to shape history. The sentiment found within comments, feedback or critiques provide useful indicators for many different purposes. These sentiments can be categorized either into two categories: positive and negative; or into an n-point scale, e.g., very good, good, satisfactory, bad, very bad. In this respect, a sentiment analysis task can be interpreted as a classification task where each category represents a sentiment. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others.

The sudden eruption of activity in the social networking domain has drawn the attention of analysts, social media as well as general public to area of sentiment analysis to extract invaluable information from public opinion. Data miners use Twitter as the source of its opinionated data. One way to perform Twitter data analysis is to directly exploit traditional Sentiment Analysis methods. However, tweets are quite different from other text forms like product reviews and news articles. Firstly, tweets are often short and ambiguous because of the limitation of characters. Secondly, there are more misspelled words, slang, modal particles and acronyms on Twitter because of its casual form. Thirdly, a huge amount of unlabeled or noisy labeled data can be easily downloaded through 2 Twitter API. We propose a method for modeling the data and analyzing it to solve the complex real world problem solving.

DOI: 10.15680/IJIRCCE.2016. 0404304

Sentiment analysis over Twitter data and other similar micro blogs faces several new challenges due to the typical short length and irregular structure of such content. Two main research directions can be identified in the literature of sentiment analysis on micro blogs. First direction is concerned with finding new methods to run such analysis, such as performing sentiment label propagation on Twitter follower graphs, and employing social relations for user-level sentiment analysis. The second direction is focused on identifying new sets of features to add to the trained model for sentiment identification, such as micro blogging features including hash tags, emoticons, the presence of intensifiers such as all-caps and character repetitions etc., and sentiment-topic features.

II. LITERATURE SURVEY

Cornelian et.al [2], In the context of an architecture of Information Retrieval, there research aims on implementing a semantic extraction agent for the Web environment, allowing information finding, storage, processing and retrieval, such as those from the Big Data context produced by several informational sources on the Internet, serving as a basis for the implementation of information environments for decision support. Using this method, it will be possible to verify that the agent and ontology proposal addresses this part and can play the role of a semantic level of the architecture.

Agarwal et.al [3] examined sentiment analysis on Twitter data. The contributions of this paper are: (1) we introduce POS-

Specific prior polarity features (2) we explore the use of a tree kernel to obviate the need for tedious feature engineering. The new features (in conjunction with previously proposed features) and the tree kernel perform approximately at the same level, both outperforming the state-of-the-art baseline.

David et.al [4], introduced a generic, automatic classification method that exploits Semantic Web technologies to assist in several phases in the classification process; defining the classification requirements, performing the classification and representing the results. Using Semantic Web technologies enables flexible and extensible configuration, centralized management and uniform results. This approach creates general and maintainable classifications, and enables applying semantic queries, rule languages and inference on the results.

III. PROPOSED SYSTEM

Our proposed system is based on Twitter specific sentiment analysis. The research on sentiment analysis mainly focused on two things: identifying whether a given textual entity is subjective or objective, and identifying polarity of subjective text. In our system we are identifying good words and bad words or irrelevant. Most sentiment analysis studies use machine learning approaches. We have employed SVM classifier.

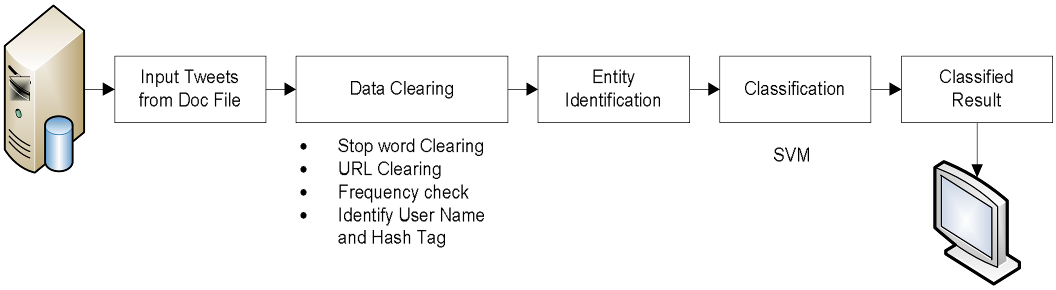

Figure 1: Block Diagram of Proposed Work

a) Data Cleaning/Data Processing

In this we clean the data by removing irrelevant and unwanted data. If we specify a keyword it is possible that the keyword is used in some other context which is not required for us. Cleaning of data is followed by sentiment analysis on the tweets. After sentiment analysis we can retrieve knowledge from the tweets.

b) User Name and Hash tag

The hash tag data set is a subset of the Edinburgh Twitter corpus. The Edinburgh corpus contains 97 million tweets collected over a period of two months. To create the hash tagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain hash tags. From the remaining set (about 4 million), we investigate the distribution of hash tags and identify what we hope will be sets of frequent hash tags that are indicative of good, bad, and irrelevant messages. These hash tags are used to select the tweets that will be used for development and training. Messages with these hash tags were included in the final dataset, and the polarity of each message is determined by its hash tag.

c) Entity identification

An entity is a physical or non-physical thing that can be identified by its properties (e.g., my university). A named entity is an entity that has been assigned a name (e.g., “good boy, bad boy, smart”). Thus, the mention of a named entity in a text is defined as a named entity mention.

d) Feature extraction

Text, smiles, emotions and punctuations are combination of tweets. The collected tweet class labels are annotated manually. These features are extracted using TF-IDF. Given a query q composed of a set of wordsw_i, we calculate$w_{i\,d}$, for each w_i for every document d ϵ D. In the simplest way, this can be done by running through the document collection and keeping a running sum of $f_{w,d}$ and $f_{w,D}$. Once done, we can easily calculate $w_{i\,d}$ according to the mathematical framework presented before. Once all $w_{i,d}$s are found, we return a set D* containing documents d such that we maximize the following equation

$$\sum_i w_{i,d} \qquad (1)$$

Either the user or the system can arbitrarily determine the size of D* prior to initiating the query. Also, documents are returned in a decreasing order according to equation (1).

e) *Classification*

Training Data – A hand-tagged collection of data is prepared by most commonly used crowd-sourcing method. This data is the fuel for the classifier; it will be fed to the algorithm for learning purpose. Next step: Classification – This is the heart of the whole technique. Depending upon the requirement of the application SVM is deployed for analysis. The classifier (after completing the training) is ready to be deployed to the real time tweets/text for sentiment extraction purpose.

IV. EXPERIMENTAL RESULT

In our proposed system Semantic we consider good and bad words.

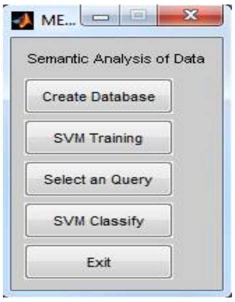

Figure 2: Menu created for the Execute of the Proposed Model.

First step is to training phase, so that later when query is given it will classify based on the data we train. So we create the database i.e., text for important data and useless data are read and their features are extracted and trained using SVM training Next is to testing phase, Query data is selected when select an query is opted, the data is read from the

text file is passed to data cleaning model where URL, username, hash tags and stop words are eliminated. After data cleaning, features are extracted from the data which are classified using SVM classification.
Consider the query twitter data as follows:

[#MLUC09] Customer Innovation Award Winner: Booz Allen Hamilton.

The hash tags identified in input data is:

When the above hash tag is removed the data now looks as follows:

Customer innovation award winner Booz Allen Hamilton -- http://ping.fm/c2hpp

Next step is identifying the URL and remove it. The URL found is shown as follows:

http://ping.fm/c2hpp

The URL eliminated is as follows

Customer innovation award winner Booz Allen Hamilton [#MLUC09]

V.Conclusion

Opinion classification is an important and challenging task using twitter data set. A twitter micro blog suffers from various linguistic and grammatical errors. This paper provides how to pre-process the tweets for maximum information extraction from short text message and classify the opinion using machine learning algorithms. Machine learning approaches have been so far good in delivering accurate results. Using hash tags to collect training data did prove useful, as did using data collected based on good and bad words. However, which method produces the better training data and whether the two sources of training data are complementary may depend on the type of features used.

ReferenceS

1. Schulz, "Semantic Abstraction for Generalization of Tweet Classification: An Evaluation on Incident-Related Tweets", IOS Press, pp. 1-1, 2015.
2. Alec," Twitter Sentiment Classification using Distant Supervision", IEEE, pp 234-123, 2008.
3. Harsh," Approaches for Sentiment Analysis on Twitter: A State-of-Art study", IEEE, pp121-345, 2015.
4. Panasyuk," Extraction of Semantic Activities from Twitter Data".
5. Bautin, M," sentiment analysis for news and blogs". In Second Int. Conf. on Weblogs and Social Media ICWSM, 2008.
6. Turney, "Semantic orientation applied to unsupervised classification of reviews". In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp 417–424, 2002.
7. Pang, " Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval", Vol. 2, pp 1-21-135, 2008.
8. Pang, "Using very simple statistics for review search", an exploration. In Proceedings of the International Conference on Computational Linguistics (COLING), 2008
9. Pang, "Using very simple statistics for review search", an exploration. In Proceedings of the International Conference on Computational Linguistics (COLING), 2008
10. Pang, "Using very simple statistics for review search", an exploration. In Proceedings of the International Conference on Computational Linguistics (COLING), 2008
11. Dave, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", pp 519-528, 2008.
12. God bole, "Large-scale sentiment analysis for news and blogs", 2007.
13. Kennedy, "Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters, Computational Intelligence", pp 110–125, 2006.
14. Kamp, "Using WorldNet to Measure Semantic Orientation of Adjectives." LREC, Vol. IV, pp 1115–1118, 2008.
15. Hatzivassiloglou, "Effects of Adjective Orientation and Grad ability on Sentence Subjectivity". Proceedings of the 18th International Conference on Computational Linguistics, New Brunswick, NJ, 2000
16. Andreevskaia, "All Blogs Are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs". In: International Conference on Weblogs and Social Media (ICWSM-2007), Boulder, CO, 2007