Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review of Data Analysis of Twitter

Manisha Rani, Jyoti Arora
CSE Department, Desh Bhagat University,
Punjab, India

Abstract: ***Twitter is an online social networking site which contains huge amount of data. This data can be a in any form structured, semi-structured and un-structured data. This allows people to share and express their views about issues, have discussion about these issues with different communities, or post blogs across the world. There has been lot of work done in the field of sentiment analysis of twitter data. The term sentiment refers to the feelings or their opinion towards some particular topic. Hence it is also known as opinion mining. It leads to the subjective impressions towards the domain, not facts. This paper reviews the methods of data analysis of twitter. This paper includes three sections.***

Keywords: NLP, API, SVM, LSA

I. INTRODUCTION

Nowadays, the use of Internet has changed the way people express their ideas, opinions. With the advancement and growth of technology, there is a large volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, where people exchange their ideas and share views. Social networking sites like Twitter, Facebook, and Google+ are rapidly gaining popularity. Moreover, social media is providing an opportunity to businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making [8]. For e.g. If someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision. One of the most visited social networking sites by millions of Users is twitter where they share their opinion about various Domains like politics, brands, products, celebrities etc. Twitter is a micro blogging site that offers the opportunity for the analysis of expressed mood, and previous studies have shown that geographical, diurnal, weekly, and seasonal patterns of positive and negative affect can be observed. The amount of content generated by users is too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis (SA) tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements. Many Research works are carried out in the field of sentiment Analysis. But they are only useful in modeling and tracking Public sentiments. They had not found exact reasons behind [9].

II. SENTIMENT ANALYSIS

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction. The words opinion, sentiment, view and belief are used alternatively but there are differences between them. Sentiment analysis has many applications in various domains like political domain, sociology and real time event detection like earthquakes [8].

A. Different Classes of Sentiment Analysis

Sentiments can be classified into three classes' .i.e. positive, negative and neutral sentiments.

a) **Positive Sentiments**: These are the good words about the target in consideration. If the positive sentiments are increased, it is referred to be good. In case of product reviews, if the positive reviews about the product are more, it is bought by many customers.
b) **Negative Sentiments:** These are the bad words about the target in consideration. If the negative sentiments are raised, it is avoided from the preference list. In case of product reviews, if the negative reviews about the product are more, no one intend to buy it.
c) **Neutral Sentiments**: These are neither good nor bad words about the target. Hence it is neither preferred neglected.

B. Architectural Diagram for Sentiment Analysis

A General model for sentiment analysis is as follows,

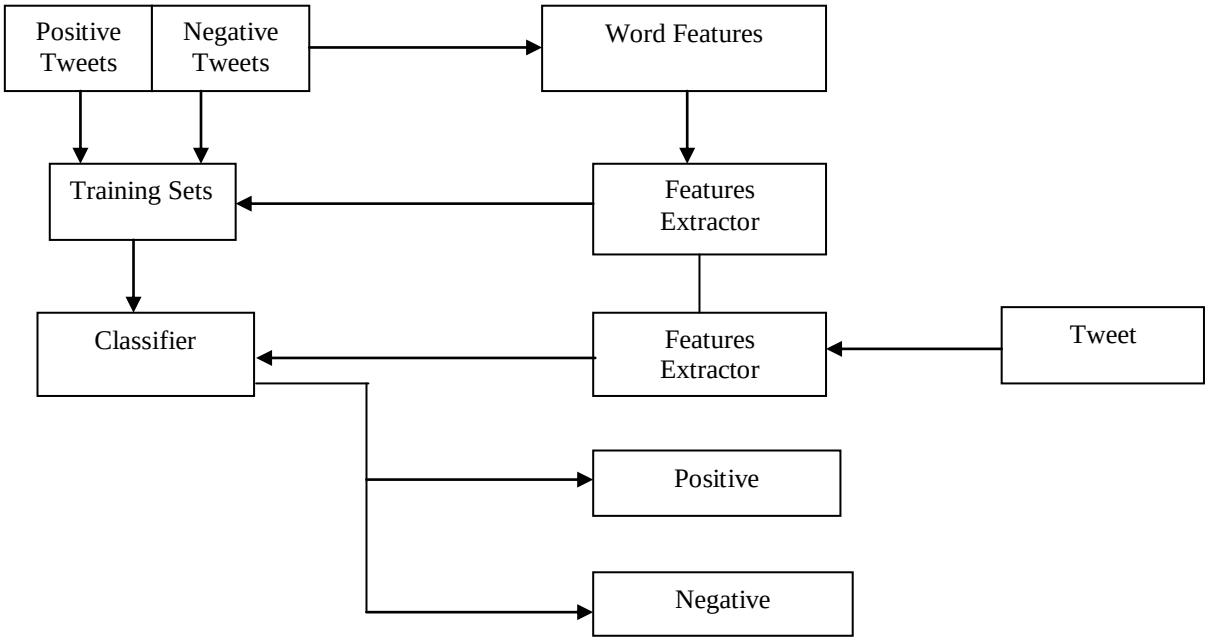

Fig.1. Sentiment Analysis Architecture [8]

III. LITERATURE SURVEY

Luo et. al. [1] highlighted the challenges and an efficient technique to mine opinions from Twitter tweets. Spam and wildly varying language makes opinion retrieval within Twitter challenging task.

Xia et al. [2] used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines) . They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

Bifet and Frank [3] used Twitter streaming data provided by Firehouse API , which gave all messages from every user which are publicly available in real-time. They experimented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They arrived at a conclusion that SGD-based model, when used with an appropriate learning rate was the better than the rest used.

Pak and Paroubek [4] proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Go and L.Huang [5] proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features.

Kotsakos et al. [6] Highlighted on tagging the tweets. They did the hash tag analysis in which a hash (#) symbol used to indicate a special meaning of a word and tag content in social networks like twitter. Users used hashtags for search, annotations or viral conversations often called Memes. They revealed interesting characteristics of some expected hash tags and some not expected hash tags. They also suggested to further investigate features that characterize the behavior of popular topics and to create taxonomies of hash tags that facilitate recommendation or searches.

Mahanaz Roshanaei and Shivakant Mishra [7] emphasize on effect of mood and emotions on a person's behavior. They classified users in to positive, negative and neutral users based on followers and followers. Negative users are not interested in sharing their negativity in social media. The positive users are more likely to make friendships with negative users; also, the negative users retweet more than the positive users. They use Twitter as a tool for social awareness and also to gain emotional support. Retweeting positive tweets makes the negative tweeters feel positive. Both positive and negative users avoid interacting with each other.

IV. TECHNIQUES

The techniques that are used for data analysis of twitter are described as following:-

A. Naïve Bayes Classifier:

The basic mechanism of Naive bayes classifier is done by counting the frequency of words that are related to sentiment in the message. On the basis of these numbers of matches to the sentimental words, tweets are classified and scored. The weight of nodes is adjusted according to the importance of tweets and more accurate result of classified sentiments can be generated.

B. Support Vector Machine:

SVM is generally used for text categorization. SVM gives best results than Naive bayes algorithm in case of text categorization. The basic idea is to find the hyper plane which is represented as the vector w which separates document vector in one class from the vectors in other class.

C. Lexicon-Based Approaches

Lexicon based method uses sentiment dictionary with opinion words and match them with the data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary are. Lexicon-based approaches mainly rely on a sentiment lexicon, i.e., a collection of known and precompiled sentiment terms, phrases and even idioms, developed for traditional genres of communication, such as the Opinion Finder lexicon.

There are two sub classifications for this approach:

a) ***Dictionary-based:*** It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet.
Drawback: Can't deal with domain and context specific orientations.

b) ***Corpus-Based:*** The corpus-based approach have objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques.

- Methods based on statistics: Latent Semantic Analysis (LSA).
- Methods based on semantic such as the use of synonyms and antonyms or relationships from thesaurus like Word Net may also represent an interesting solution.

V. CONCLUSION

In this paper, the different techniques of data analysis of twitter are discussed including machine learning and lexicon-based approaches. The analysis of Twitter data is being done in various perspectives, the presence of words like good, bad and also emoticons in the tweets can be used to infer the sentiment. The Twitter users can be classified into positive, negative and neutral users based on the followers and the followers and their behaviors can be studied based on the tweeting and retweeting activity. The study shows that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require little effort in human-labeled document.

REFERENCES

[1] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013.

[2] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[3] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer, pp. 1-15, 2010.

[4] A.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, pp.1320-1326, 2010.

[5] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[6] L. Jimmy, and A. Kolcz, "Large-scale machine learning at twitter", In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, ACM, pp. 793-804, **2012**.

[7] Jiang, U. Topaloglu and F. Yu, "Towards large-scale twitter mining for drug-related adverse events", In Proceedings of the 2012 international workshop on Smart health and wellbeing, ACM, pp. 25-32, **2012**.

[8] Kharde, Vishal, and Sheetal Sonawane. "Sentiment Analysis of Twitter Data: A Survey of Techniques." arXiv preprint arXiv:1601.06971, 2016.

[9] Ravikumar, Pushpa. "Survey: Twitter data Analysis using Opinion Mining."International Journal of Computer Applications, vol. 128, no. 5, pp. 34-36, 2015.

[10] Sheela, L. Jaba. "A Review of Sentiment Analysis in Twitter Data Using Hadoop." International Journal of Database Theory and Application, vol. 9, no. 1, 77-86, 2016.

