JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SENTIMENT SCORE OF TWITTER USING MACHINE LEARNING: REVIEW

Shaurya Vir Singh Pathania*
Prof. K. L Bansal* *

* *Shaurya Vir Singh Pathania, M.Tech Student of Department of Computer Science, Himachal Pradesh University, Shimla-171005.*

*** Prof. K L Bansal, Department of Computer Science, Himachal Pradesh University Shimla-171005.*

Abstract

Three centuries ago, Oil companies ruled the world and still do explains the monopoly of Middle-East countries enjoy petrodollar surplus. Oil, the world's most precious resource in the 18th century, was an enormous, undiscovered valuable asset. It was the backbone of everything from the government to small businesses. Without it, development would come to a halt, and economies would contract. Today Social Media outlets such as Twitter analogous to oil is a gold mine for various companies and advertisers which has recently opened up a lot of doors for them that want to track and control the credibility of their brands, products and businesses, as well as policymakers and leaders by helping them in evaluating the public opinions on their policies or political matters. This paper presents a review to comprehend the various components and basic information required to conduct Twitter sentiment analysis

Keywords Sentiment analysis, Machine learning, Tweet Sentiment Analysis, Twitter

1. Introduction

The Internet is a wide virtual place where people may express and share their personal opinions, which has an impact on all aspects of life and has implications for marketing and communication. Social media has an impact on consumer habits by altering their beliefs and behaviours. Social media monitoring of consumer loyalty and attitude toward companies or products is an useful technique to measure customer loyalty and keep track of their sentiments.

We can comprehend a lot from the massive amount of data that is being shared across multiple social media platforms. This information comes in the form of online journals, comments, and reviews, among other things. People nowadays choose to share their opinions on various topics using social media sites, such as Twitter, Facebook, Pinterest & Quora etc. Few decades ago, people used to express their opinions by writing or speaking in public places. These reviews were also used as suggestions for improvement. This data would then be processed manually. With the advent and increase of the usage of internet, people began to share their opinions about various things via emails or social media platforms.

Sentiment Analysis (SA), also known as opinion mining, is the process of classifying the emotions, conveyed by a text, as negative, positive or neutral. The data made available by social media has contributed to a lot of research activities within SA in recent times. Information gained by applying SA to social media data has many potential usages, for instance, to help marketers evaluate the success of an ad campaign, to identify how different demographics have received a product release, to predict user behaviour, or to forecast election results.

Sentiment analysers are increasingly being used by businesses to identify flaws in their products or services. An emotion analyser that works as rationally as humans is the best. So, the goal is to close the research gaps in effective sentiment processing.

Microblogging websites like Twitter allow users to write textual entries of up to 280 characters. Commonly referred to as tweets, there have seen a significant increase in their popularity in recent years. Companies and media organisations are increasingly looking for ways to mine Twitter for information about what people think and feel about their goods and services as a result of this development.

Though, a reasonable amount of research has been carried out on how sentiments are expressed in online reviews and news articles, very less research has been done on how sentiments are expressed in microblogging due to the informal language and message-length constraints.

2. Related Work

There is a plethora of related works for sentiment analysis but, we are only interested in contributions for Twitter Sentiment Analysis. The task of categorizing or recognizing the text as good, negative, or neutral can be termed as sentiment analysis. It's a multifaceted activity that uses Natural Language Processing and Machine Learning approaches to conduct numerous detection tasks at various text granularity levels. There are three approaches to sentiment analysis, lexicon based, machine learning based and a hybrid one. For the purpose of this research, the review of literature is confined to the machine learning technique.

G. Shobana, et.al [1] analysed the famous person's id's (@realdonaldtrump) or hash tags (#IPL2018) for understanding the mindset of people in each situation when the person has tweeted or has acted upon some incidents. The proposed system is to analyze the sentiment of the people using python, twitter API, Text Blob (Library for processing text). As the results it helps to analysis the post with a better accuracy.

Brian Heredia, et.al [2] conducted an empirical study using sentiment data from two sources, online reviews and tweets. We first test the performance of sentiment analysis models built using a single data source for both in-domain and cross-domain classification. Then, we evaluate classifiers trained using instances randomly sampled from both sources. Additionally, the experts evaluated sampling different quantities of instances from both data sources to determine how many instances should be included in a training data set. We apply statistical tests to verify the significance of our results and find that using a combination of instances from reviews and tweets is similar to, or better than any model trained from a single domain.

Rajkumar S. Jagdale, et.al [3] elaborated different approaches of Sentiment Analysis and Opinion Mining for different dataset and find out which approach is best for which dataset which will help to researchers to select approach and dataset. In proposed work we collected tweets using R tool of different events from twitter and did pre-processing and calculate sentiment score from that events. We plot Wordcloud of particular event which highlight the frequent term from tweets and also calculated numbers of positive, negative and neutral tweets from each events.

Aishwarya Kotwal, et.al [4] explained that Twitter represents a microblogging site where people post and read views about various topics. These tweets contain people's opinion, emotions, sentiments, appraisals, evaluations regarding entities consisting of movies, politics, research, business, sports etc. This data can be obtained by using Twitter API services. The sentiments of this collected data can be studied, analysed and categorized as positive, negative or neutral. Thus the popularity of the topic can be detected from the statistics of the opinions and emotions which is achieved by classifying the data to the trained form. The size of the data obtained from the twitter is humungous. To handle such data the Hadoop framework is used to store, process and manage it so that it can be time efficient.

Bogdan Batrinca, et.al [5] presented a study to analyze the wealth of social media now available. It presents a comprehensive review of software tools for social networking media, wikis, really simple syndication feeds, blogs, newsgroups, chat and news feeds. For completeness, it also includes introductions to social media scraping, storage, data cleaning and sentiment analysis. Although principally a review, the paper also provides a methodology and a critique of social media tools**.**

Zohreh Madhoushi, et.al [6] aimed to categorize SA techniques in general, without focusing on specific level or task. And also to review the main research problems in recent articles presented in this field. We found that machine learning-based techniques including supervised learning, unsupervised learning and semi supervised learning techniques, Lexicon-based techniques and hybrid techniques are the most frequent techniques used.

Sareh Aghaei, et,al [7] provided an overview from the evolution of the web. Web 1.0, web 2.0, web 3.0 and web 4.0 were described as four generations of the web. The characteristics of the generations are introduced and compared. It is concluded web as an information space has had much progress since 1989 and it is moving toward using artificial intelligent techniques to be as a massive web of highly intelligent interactions in close future.

Huifeng Tang, et.al [8] discussed four problems, i.e., subjectivity classification, word sentiment classification, document sentiment classification based on machine learning techniques, and opinion extraction problem. Although we were able to obtain fairly good results for the review classification task through the choice of appropriate features and metrics, but we identified a number of issues that make this problem difficult.

Bo, Pang, et.al [9] considered the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods we employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. The experts concluded by examining factors that make the sentiment classification problem more challenging.

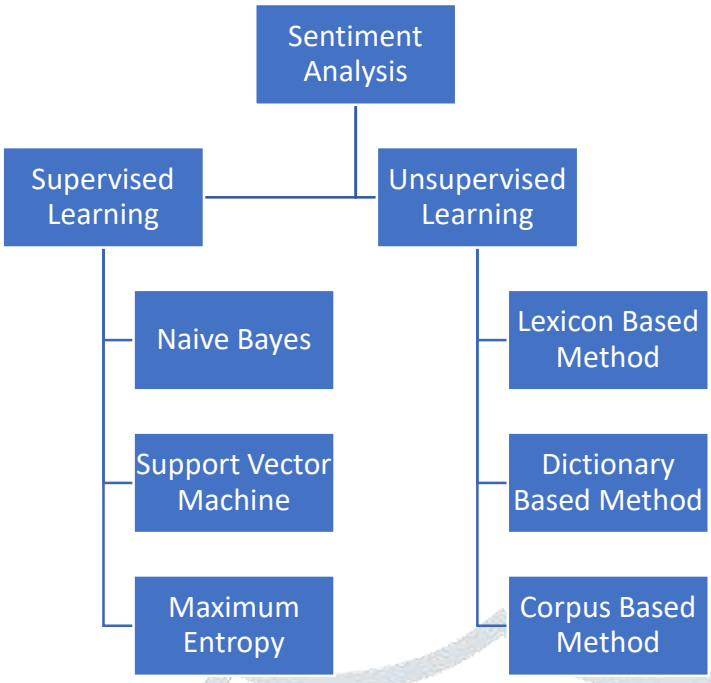
3. Methodology

"Sentiment Analysis is the computational study of people's feelings and opinions expressed in the text". In other words, it is the task of identifying the mood of people about a particular subject. It is also referred as Opinion Mining, Subjectivity Analysis, Appraisal Extraction and Review Mining with some associations to Affective Computing. It automates the retrieval of text from appropriate sources, extraction of relevant sentences, understanding its contents, summarizing it and presenting the results in an appropriate format. In Computational Terms it is defined "as a data mining technique that uses Natural Language Processing, Computational Linguistic and Text Analytics to identify and extract content of interest from a body of textual data"

A new system was proposed which uses machine learning approach and Lexicon based approach to provide the polarity for sentences present in the Twitter Datasets. Online communities are considered as a good resource of texts in the form of tweets, posts, blogs, comments etc. Text is the words, sentences or paragraphs that the users use to convey their opinions or share their views. Extracting opinions and sentiments from online text requires different levels of text preprocessing, text classification and text mining approaches. The text classifiers can be broadly classified into different classes based on their approaches viz. Machine learning approach and the Lexicon based approach

3.1 Machine Learning Approach

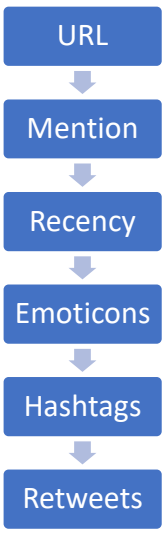
Machine learning algorithms can be addressed as a combination of methods to automatically detect the available pattern in the given set of data. It makes use of undiscovered patterns to forecast the future data (or) to implement the decision making under uncertainty. Machine learning can be performed in 2 ways such as supervised and unsupervised. Supervised learning is performed by considering the target value and unsupervised learning is conducted by not considering the target value



3.2 Twitter

Twitter is an online social networking and micro blogging service that enables users to share and discuss their thoughts and views in 140 characters without being constrained by space and time. A tweet is not only a simple text message but it is a combination of text data and Meta data associated with the tweet. These attributes are the features of tweets. They express the content of the tweet or what is that tweet about. The Metadata can be utilized to find out the domain of the tweet. The Metadata of tweet are some entities and places. These entities include user mentions, hashtags, URLs, and media Users, Twitter userID. RT stands for retweet,'@' followed by a user identifier report the user, and'#' followed by a word characterizes a hashtag. Work on the Twitter in this paper is limited up to text data

Features of Twitter:



- URL

Many tweets share a link along with the introduction to the links. The sharing link is initiated as URL. Presence of URL, gives its feature value as 1, else is 0.

- Mention

In a tweet when user want to refer to another user he can write his name starting with @ symbol. It is called as Mention and it also represented as "@username". If tweet encloses mention the binary feature representing it will have value 1, else is 0

- Recency

When the query is fired to get a tweet, it is better to get most recent tweet about that matter. Thus Recency feature measures the age of tweet in seconds after its generation.

Emoticons

are face-based and symbolize sad or happy feeling, even though there exists wide range of non-facial variations. For extracting the feeling polarity from an emoticon, various set of common emoticon can be used

- Hashtags

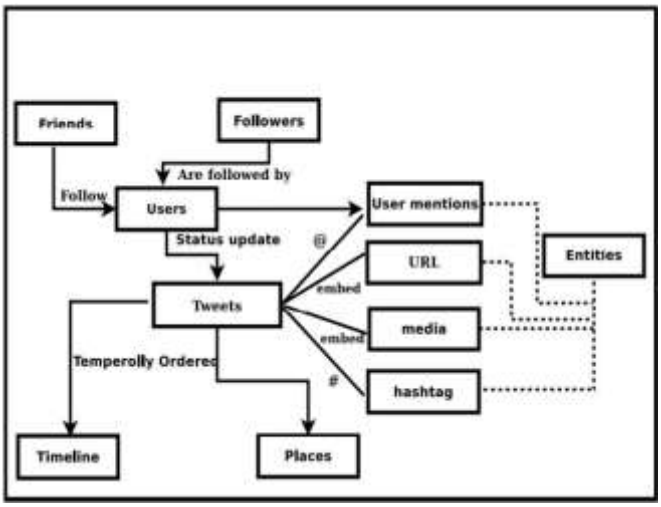
A hashtag is a word which consists of the hash symbol ("#") and a phrase. It represents a group or a topic associated with a tweet. We replace hash symbols in hashtags in a tweet by the hash word.

- Retweets

A tweet can be just a statement made by a user, or could be a reply to another tweet. Retweets are marked with either "RT" followed by '@user id' or "via @user id". Retweet is considered the feature that has made Twitter a new medium of information dissemination as well as direct communication.

3.3 Twitter Architecture

Twitter adheres to the REST (REpresentational State Transfer) style architecture proposed by Fielding (2000). REST style architecture consists clients and servers where clients initiate request and servers process the requests. Twitter API (Application Program Interface) allows three kinds of HTTP requests: GET, POST and DELETE. There are three kinds of Twitter APIs



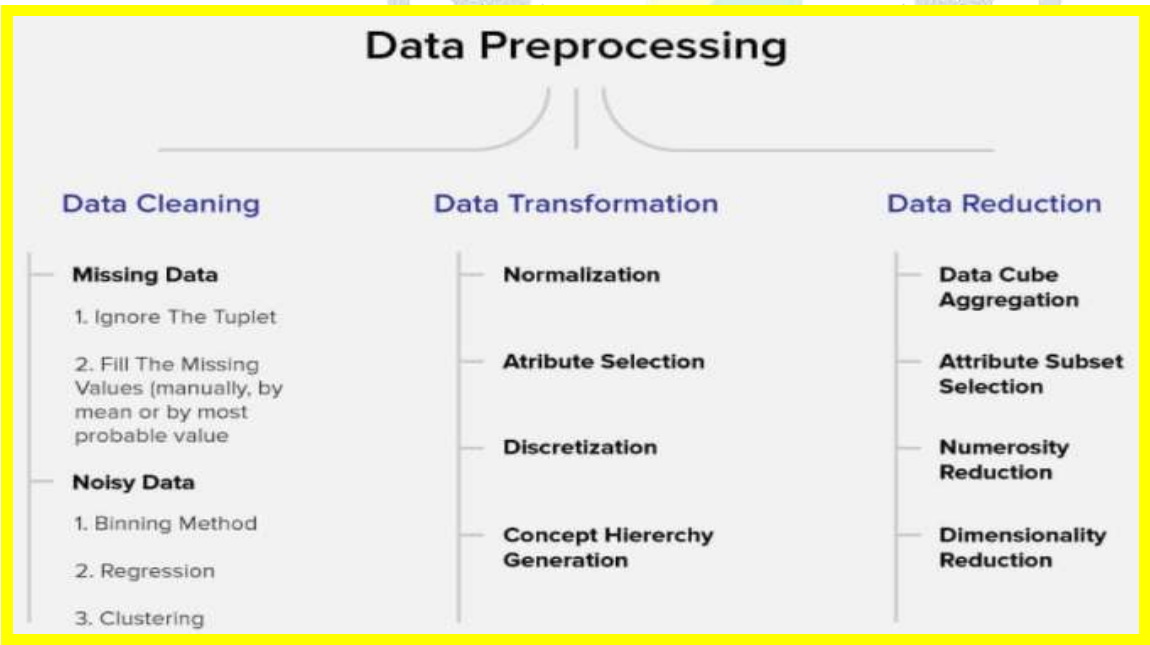
Search API allows a user to query for Twitter content. These types of APIs are used to find tweets with specific keywords, finding tweets referencing a specific user, or finding tweets from a particular user. But there is limit for the rate of access. REST API provides access to timeline, status and user objects. Through the REST API, the user can create and post tweets back to Twitter, reply to tweets, favourite certain tweets, retweet other tweets etc. Streaming API allows large quantities of keywords to be specified and tracked, retrieving geo-tagged tweets from a certain region, or have the public statuses of a user set returned. Streaming API is required to mine and analyse real time tweets

4. Sentiment Analysis in Twitter

Sentiment Analysis (SA), also known as opinion mining, is the process of classifying the emotions, conveyed by a text, as negative, positive or neutral. The data made available by social media has contributed to a lot of research activities within SA in recent times. Information gained by applying SA to social media data has many potential usages, for instance, to help marketers evaluate the success of an ad campaign, to identify how different demographics have received a product release, to predict user behaviour, or to forecast election results.Sentiment analysers are increasingly being used by businesses to identify flaws in their products or services. An emotion analyser that works as rationally as humans is the best. So, the goal is to close the research gaps in effective sentiment processing.

4.1 Data Pre-Processing

Data pre-processing is done to eliminate the incomplete, noisy and inconsistent data. Data must be pre-processed in order to perform any data mining functionality.



4.2 Feature Extraction

is a process in data mining that involves the steps for reducing the amount of data available to describe a large set of data. While performing the sentiment analysis of complex texts one of the major problems stems from the number of variables involved. Analyzing huge and complex texts generally requires a large amount of memory and computation power. This may result in over fitting of the classification algorithm to training samples and

generalize poorly to new samples. Applying feature extraction techniques to the input data before passing the input data to the classification algorithm results in improving the accuracy of the classifier model.

4.3 Feature Selection

This is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict. If you are using supervised learning or some hybrid that includes that approach, your data will be enriched with data lab. This is used to measure the performance, such as accuracy or efficiency, of the algorithm we are using to train the machine. Test data will help us to see how well our model can predict new answers, based on its training. Both training and test data are important for improving and validating machine learning models.

4.4 Sentiment Polarity

Final output of sentiment analysis process is obtained as identification of sentiment polarity. Possible polarities in a given statement can be positive, negative or neutral which can be listed by the results obtained by the classification algorithms used in sentiment classification.

5. Conclusion

We reviewed a variety of studies and found that this review paper clearly states the basic information required to conduct Twitter sentiment analysis. What is Sentiment Analysis in terms of levels of sentiment analysis, approaches to sentiment analysis, sentiment analysis methodology, features to be extracted from text, and applications where it may be used are all discussed hierarchically. If we want to undertake Twitter sentiment analysis, we must first learn about Twitter, including how to extract tweets, their structure, and their significance. This document provides a quick overview of tweets. When it comes to sentiment analysis of tweets, one must specialise on a certain aspect of sentiment analysis. This paper provides a basic understanding of Twitter Sentiment Analysis. Different methods and procedures are compared and contrasted. We can imagine the future based on the accuracy/results of each approach.

References

[1] Shobana G, Vigneshwara B, Maniraj Sai A. (2018). Twitter Sentimental Analysis. International Journal of Recent Technology and Engineering (IJRTE). 7(4s) (pp: 2277-3878)

[2] Heredia, B., Khoshgoftaar, T. M., Prusa, J., & Crawford, M. (2016, November). Integrating multiple data sources to enhance sentiment prediction. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)* (pp. 285-291). IEEE.

[3] Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... & Socher, R. (2016, June). Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning* (pp. 1378-1387).

[4] Kotwal, Aishwarya, Jadhav, Dipali & Fulari, Priyanka. (2016). Improvement in Sentiment Analysis of Twitter Data using Hadoop. *International Conference on "Computing for Sustainable Global Development. pp: 0973-*7529.

[5] Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, *30*(1), 89-116.

[6] Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In *2015 Science and Information Conference (SAI)* (pp. 288-291). IEEE.

[7] Aghaei, S., Nematbakhsh, M. A., & Farsani, H. K. (2012). Evolution of the World Wide Web: From WEB 1.0 TO WEB 4.0. *International Journal of Web & Semantic Technology*, *3*(1), 1.

[8] Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*(7), 10760-10773.

[9] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up? : Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*(pp. 79-86). Association for Computational Linguistics.