IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

A Survey on Twitter Sentiment Analysis

Eriq-Ur Rahman[1*], Rituparna Sarma[2], Rajesh Sinha[3], Priyankar Sinha[4], Adarsh Pradhan[5]

[1,2,3,4,5]Dept. of CSE, Girijananda Chowdhury Institute of Management and Technology, Guwahati, India

**Corresponding Author: eriqrahman007@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 22/Nov/2018, Published: 30/Nov/2018

***Abstract*-** Twitter sentiment analysis offers organizations an ability to monitor public feeling towards the products and events related to them in real time. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous tweets. It offers organizations a fast and more effective way to analyze customer's perspectives towards the success in the market place. Sentiment analysis is an approach to be used to computationally measure customer's perceptions to a vast extent. This is a survey on the design of a sentiment analysis. After extraction of a vast amount of tweets, it classifies perspectives of customers via tweets into positive and negative sentiments. Which is obtained after classifying the data by using classification approaches like for example Bayes Naïve, Linear Regression, etc.

Keywords—Twitter, sentiment analysis, datasets, pre-processing, feature extraction, classification,

I. INTRODUCTION

Internet has changed the way people express their views and opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. We can see that millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. at a rate of millions each day.

Twitter-

Twitter is a popular real time microblogging service that allows users to share short information known as tweets which are limited to 140 characters [1,2], [3]. Users write tweets to express their opinion and feelings about various topics relating to their daily lives making it an ideal platform for the extraction of general public opinion on specific issues [4, 5]. A collection of tweets is used as the primary corpus for sentiment analysis, which refers to the use of opinion mining or natural language processing [6]. Twitter with more than 500 million users and million messages per day, has quickly become a valuable asset for organizations to invigilate their reputation and brands by extracting and analyzing the sentiment of the tweets by the public about their products, services market and even about competitors [7]. [1] highlighted that, from the social media generated opinions with the mammoth growth of the world wide web, huge volumes of opinion texts in the form of tweets, reviews, blogs or any discussion groups and forums are available for analysis, thus making the world wide web the fastest, most comprising and easily accessible medium for sentiment analysis

Behind all these websites, textual information retrieval techniques mainly focus on processing, searching or analyzing the factual data present. Facts have an objective component but, there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis. It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks. For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items by making use of sentiment analysis.

Here the rest of the paper is organized as follows. In Section II problem statements for analyzing tweets has been mentioned. The pre-processing of tweets and the methods to proceed are shown in section III, Section IV consists of datasets and examples of popularly known datasets, Section V and Section VI consists of feature extraction and classifications respectively, Section VII explains why python language is recommended and at last Section VIII contains the conclusion.

II. Problem Statements.

The availability of software to extract data regarding a person's sentiment on a specific product or service,

© 2018, IJCSE All Rights Reserved

organizations and other data workers still face issues regarding the data extraction.

• Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [8]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract sentences, read them, analyze tweet by tweet, summarize them and organize them into an understandable format in a timely manner [8].

• Difficulty of Sentiment Analysis with inappropriate English

Informal language refers to the use of colloquialisms and slang in communication, employing the conventions of spoken language [9] such as 'would not' and 'wouldn't'. Not all systems are able to detect sentiment from use of informal language and this could hamper the analysis and decision making process. Emoticons, known to be the pictorial representation of human facial expressions [2], which in the absence of body language and prosody serve to draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, improving and changing its interpretation [10]. For example, ☺ indicates a happy state of mind. Systems currently in place do not have sufficient data to allow them to draw feelings out of the emoticons. As humans often turn to emoticons to properly express what they cannot put into words [10]. Not being able to analyze this puts the organization at a loss. Short-form is widely used even with short message service (SMS). The usage of short-form will be used more frequently on Twitter so as to help to minimize the characters used. This is because Twitter has put a limit on its characters to 140[11]. For example, 'tba' refers "to be announced".

III. Pre-processing

The pre-processing methods that are assessed in this paper are as follows:

- Replacing negative mentions. Tweets consist of various notions of negation. In simple words, negation has a very important role in determining the sentiment of the tweet. Here, the process of negation is transforming "won't", "can't", and "n't" into "will not", "cannot", and "not", respectively.
- Removing URL links in the corpus. URL's do not carry much information regarding the sentiment of the tweet considered by almost all researchers. Twitter's short URLs are expanded to URLs and are tokenized. Then, the URL matching the tokens are removed from tweets to refine the tweet content.
- Reverting words containing repeated letters to their original English form. Words with repeated letters, e.g., "coooool", are common in tweets, and people tend to use this way to express their sentiments. Here, a sequence of more than three similar characters is replaced by three characters. For example, "cooooool" is replaced by "coool". Using three characters distinguish words like "cool" from "cooooool".
- Removing numbers. In general, numbers are of no use when measuring sentiment and are removed from tweets to refine the tweet content.
- Removing stop words. Stop words usually refer to the most common words in a language, such as "the", "is", and "at". Almost all researchers consider that stop words play a negative role in the task of sentiment classification, and they are removed before feature selection by researchers. The classic method of removing stop words is the method based on pre-compiled lists. Multiple lists exist in the literature [7], [12].
- Expanding acronyms to their original words by using an acronym dictionary. Acronyms and slang are common in tweets but are ill-formed words. It is necessary to expand them to their original words. Terms have originated from various sources, including Bulletin Boards, AIM, Yahoo, IRC, Chat Rooms, Email, and Cell Phone Text Messaging. Each acronym corresponds to an explanation. Example, "*4 u" is "Kiss for you", "2 mro" is "tomorrow".

IV. Datasets

Pre-processing may have different impacts in various contexts. Words and URLs that do not provide any discriminative power in one context may carry some semantic information in another context. The effects of pre-processing on five different Twitter datasets that have been used in other sentiment analysis literature focusing our selection on those datasets which are:

- publicly available to the research community,
- manually annotated, providing a reliable set of judgements over the tweets and,
- Used to evaluate several sentiment analysis models. Tweets in these datasets have been annotated with diff erent sentiment labels including: Negative, Neutral, Positive, Mixed, Other and Irrelevant.

Some of the popularly recognized datasets available in the internet are-

- The Stanford Twitter Sentiment Test (STS-Test) dataset was introduced by Go et al. [13]. It has been manually annotated and contains 177 negative, 182 positive and 139 neutral tweets. Although the Stanford test set is relatively small, it has been widely used in the literature [14], [15] for different evaluation tasks.
- SemEval2014 dataset was provided in SemEval2014 Task95. The dataset consists of tweet id's which have been annotated with positive, negative and neutral

© 2018, IJCSE All Rights Reserved

labels. Some of the tweets were not available for downloading. This leaves us with 11042 tweets for testing.

- The Stanford Twitter Sentiment Gold (STS-Gold) dataset was introduced by Saif et al. [16].The dataset has been manually annotated both the tweet-level and the entity-level by three graduate students.
- The Sentiment Strength Twitter Dataset (SS-Twitter) consists of 4242 tweets manually labeled with their positive and negative sentiment strengths. The dataset was constructed by Thelwall et al. [14] to evaluate SentiStrenth.
- The Sentiment Evaluation Dataset (SE-Twitter) was introduced by Sacha Narr et al. [17]. The dataset consists of 6745 tweets that have been human-annotated with sentiment labels by three Mechanical Turk workers.

TABLE I. TOTAL NUMBER OF TWEETS AND THE TWEET SENTIMENT DISTRIBUTION IN ALL DATASETS

Dataset	No. Of Tweets	#Negative	#Neutral	#Positive
STS-Test	498	177	139	182
SemEval2014	11042	1650	5150	4242
STS-Gold	2023	1402	--	621
SS-Twitter	4242	1037	1953	1252
SE-Twitter	6745	990	4097	1658

V. Feature Extraction

There are many distinctive properties in the preprocessed dataset where the aspects are extracted from. Later these aspects text are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram [7]. Machine learning techniques require representing the key features of texts or documents for processing. These key features are considered as feature vectors which are used for the classification task. Some examples of features that have been reported in literature [18] are:

1. Frequency of words: Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Pang et al [18] showed better results by using presence instead of frequencies.

2. Speech Tag and its parts: Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3. Phases and Opinion Words: Apart from specific words, some phrases and idioms which convey sentiments can be used as features. E.g. cost someone an arm and leg.

4. Position of Terms**:** The position of a term within a text can effect on how much the term makes difference in overall sentiment of the text.

5. Negation: It is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.

6. Syntax: Syntactic patterns like collocations are used as features to learn subjectivity patterns by many of the researchers.

Many statistical feature selection methods for document level classification can also be used for sentiment analysis. The simplest statistical approach for feature selection is to use the most frequently occurring words in the corpus as polarity indicators. The majority of the approaches for sentiment analysis involve a two-step process:

• Identify the parts of the document to contribute the positive or negative sentiments.

• Join these parts of the document in ways that increase the odds of the document falling into one of these two polar categories.

In [24] presented a system in which the movie reviews are summarized at a level which helps a user to easily find out which aspects of movie are liked and disliked by the user. To find the highest information features, we need to calculate information gain for each word. Information gain for classification is a measure of how common a feature is in a particular class compared to how common it is in all other classes. A word that occurs primarily in positive movie reviews and rarely in negative reviews is high information. For example, the presence of the word “magnificent” in a movie review tweet is a strong indicator that the review is positive. That makes “magnificent” a high information word. The point is to use only the most informative features and ignore the rest

VI. Classification

Classification using machine learning can be in two steps:

1. Learning the model using the training dataset. 2. Applying the trained model to the test dataset.

Considering the tweets tweeted by users each row of the dataset contains the text of the tweet, based on the tone of the tweet we need to classify as positive or negative. Next few instances should be trained with an algorithm using the reviews and classifications and then make predictions on the

© 2018, IJCSE All Rights Reserved

reviews. It will be able to calculate error using the actual classification and see how good the predictions were. [19].

Sentiment analysis is a text classification problem and thus any existing supervised classification method can be applied. Naïve Bayes classifier is a simple probabilistic classifier that is based on the Bayes theorem. This classification technique assumes that the presence or absence of any feature in the document is independent of the presence or absence of any other feature. Naïve Bayes classifier considers a document as a bag of words and assumes that the probability of a word in the document is independent of its position in the document and the presence of other word.

Naive Bayes classification algorithm, A Naive Bayes classifier works by figuring out the probability of different attributes of the data being associated with a certain class. This is based on Bayes' theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

This basically states "the probability of A given that B is true equals the probability of B given that A is true times the probability of A being true, divided by the probability of B being true."

VII. Python

Python was found by Guido Van Rossum in Netherlands, 1989 which has been public in 1991[20]. Python is a programming language that's available and solves a computer problem which is providing a simple way to write out a solution [20]. [21] Mentioned that Python can be called as a scripting language. Moreover, [21] and [21] also supported that actually Python is a just description of language because it can be one written and run on many platforms. In addition, [23] mentioned that Python is a language that is great for writing a prototype because Python is less time consuming and working prototype provided, contrast with other programming languages. Many researchers have been saying that Python is efficient, especially for a complex project, as [22] has mentioned that Python is suitable to start up social networks or media steaming projects which always are web-based which is driving a big data. [23] Gave the reason that because Python can handle and manage the memory used. Besides Python creates a generator that allows an iterative process of things, one item at a time and allows program to grab source data one item at a time to pass each through the full processing chain

VIII. Conclusion

Twitter sentiment analysis is developed to analyze customer's perspectives towards the critical to get success in the marketplace. The program is using a machine-based learning approach which is more accurate for analyzing a sentiment; together with natural language processing techniques will be used. The proposed work presents an approach for sentiment analysis by comparing the different combination with various feature selection schemes.

REFERENCES

[1] A. K. Jose, N. Bhatia, and S. Krishna, "TwitterSentimentAnalysis". NationalInstituteof TechnologyCalicut, 2010.

[2] M. Comesaña, A. P.Soares, M.Perea, A.P. Piñeiro, I. Fraga, and A. Pinheiro, " Author ' s personal copy Computers in Human Behavior ERP correlates of masked affective priming with emoticons," Computers in Human Behavior, 29, 588–595, 2013

[3] S.Lohmann, M. Burch, H. Schmauder and D. Weiskopf, "Visual Analysis of Microblog Content Using Time-Varying Co-occurrence Highlighting in Tag Clouds," Annual conference of VISVISUS. Germany: University of Stuttgart, 2012.

[4] D. Osimo, and F. Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis," Proceeding of the 12th conference of Fruct association, 2010, United Kingdom.

[5] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Special Issue of International Journal of Computer Application, France: Universitede Paris-Sud, 2010.

[6] M.Rambocas, and J. Gama, "Marketing Research: The Role of Sentiment Analysis". The 5th SNA-KDD Workshop'11. Universityof Porto, 2013.

[7] H. Saif, Y.He, and H. Alani, "SemanticSentimentAnalysisof Twitter," Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. United Kingdom: Knowledge Media Institute, 2011.

[8]P.Lai,"ExtractingStrongSentimentTrendfromTwitter". Stanford University, 2012.

[9] Y. Zhou, and Y. Fan, "A Sociolinguistic Study of American Slang," Theory and Practice in Language Studies, 3(12), 2209–2213, 2013. doi:10.4304/tpls.3.12.2209-2213

[10] A.H.Huang, D.C. Yen, & X. Zhang, "Exploring the effects of emoticons," Information & Management, 45(7), 466–473, 2008.

[11] D. Boyd, S. Golder & G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," System Sciences (HICSS),2010.

[12] Fox C. (1992). "Information retrieval data structures and algorithms". Lexical Analysis and Stoplists, pp.102–130, 1992.

[13] Go A, Bhayani R, & Huang L. "Twitter sentiment classification using distant supervision". CS224N Project Report, Stanford, 2009

[14] Saif H, He Y, & Alani H. "Alleviating Data Sparsity for Twitter Sentiment Analysis". In Proc. CEUR Workshop Proceedings, 2012.

[15] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, & V. Varma. "Mining sentiments from tweets". In Proc. the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, Jeju, Korea, pp.11–18, 2012.

[16] Saif H, Fern M, & He Y. "Evaluation Datasets for Twitter Sentiment Analysis A survey and a new dataset the STS-Gold". Proc. 1st ESSEM Workshop. Turin, Italy, 2013.

[17] Sascha Narr, Michael Hulfenhaus & Sahin Albayrak. "Language-Independent Twitter Sentiment Analysis." 2014.

[18] V. Nareyko, "Why python is perfect for startups,"

[19] H.M.Wallach,"Topic modeling: beyond bag-of-words," in Proceedings of the 23rd International Conference on Machine Learning (ICML '06), pp. 977–984, Pittsburgh, Pa, USA, June 06.

[20] A. Sweigart, "Invent your own computer games with Python. 2nd edition, 2012.

[21] C. Seberino, "Python. Faster and easier software development," Annual Conference. California: San Diego, 2012.

© 2018, IJCSE All Rights Reserved

[22] A.Lukaszewski, "MySQL for Python. Integrate the flexibility of Python and the power of MySQL to boost the productivity of your applications," UK: Birningham. Packt Publishing Ltd, 2010.
[23] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., etal. ―Part-of-speech tagging for twitter: Annotation, features, and experiments‖. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers - volume 2 HLT '11,pp. 42–47,2011.
[24] P. Bhoir, S. Kolte, "Sentiment Analysis of Movie Reviews using Lexicon approach", IEEE International Conference on Computational Intelligence and Computing Research, Madurai pp.1-6, 2015.

BIOGRAPHIES

Eriq-ur Rahman (4th year student) pursing B.Tech in Computer Science and Engineering at GIMT, Guwahati.

Rituparna Sarma (4th year student) pursing B.Tech in Computer Science and Engineering at GIMT, Guwahati.

Rajesh Sinha (4th year student) pursing B.Tech in Computer Science and Engineering at GIMT, Guwahati.

Priyankar Sinha (4th year student) pursing B.Tech in Computer Science and Engineering at GIMT, Guwahati.

Mr. Adarsh Pradhan working as an Assistant Professor at GIMT, Guwahati has published many research papers in the areas of Machine learning, Image Processing. His area of interest is Machine learning.

© 2018, IJCSE All Rights Reserved