Various Approaches on Sentiment Analysis Using Social Media Data

I. Mohan
Assistant Professor, Information Technology, Prathyusha Engineering College, India.

R. Vigneshwar
Student, Information Technology, Prathyusha Engineering College, India.

G. Sureshkumar
Student, Information Technology, Prathyusha Engineering College, India.

B. Udhaykumar
Student, Information Technology, Prathyusha Engineering College, India.

B. Ajithkumar
Student, Information Technology, Prathyusha Engineering College, India.

Abstract-Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language. It is one of the most active research areas in recent years. Its popularity is mainly due to Social media like Facebook, Twitter, YouTube review videos, is the platform, where people express their frank opinions and thoughts about any event or program. To understand the feedback of the viewers, social media is employed as the best tool. Twitter is the world's most powerful and the 9th widely viewed social media platform. To analyse the sentiments for a channel program, a twitter API is created to extract tweets and preprocess them, to analyse the sentiments through the back-end process. Finally, the sentiment analysis is visually represented in different perspectives of sentiments, which can easily aid the enhancement of quality of the program, to provide the best entertainment and satisfy the audience.

1. INTRODUCTION

Sentiment analysis is the process of analysing the opinions, feelings and attitude of the author about a particular product, topic, task, organization etc. Hence, it is known as opinion mining. Social media has made people to express their emotions, feelings and suggestions as comments voluntarily.

It is complex to find out the overall opinion and suggestions of the people. Sentiment analysis has been a popular research area in the past few years. Many approaches and algorithms have been introduced for this analysis. We are using Naïve Bayes algorithm to predict the sentiment of people on a television program.

2. DIFFERENT CLASSES OF SENTIMENT ANALYSIS

Sentiments can be classified into three different class i.e. positive, negative and neutral sentiments.

a. Positive Sentiments: These are considered as the good words about the object/subject in concern. If there is lot of positive sentiments, it is denoted as good.

b. Negative Sentiments: These are the bad words about the object/subject in consideration. If there is lot of negative sentiments, it is rejected from the preference list.

c. Neutral Sentiments: These are neither good nor bad words about the product. Hence it is neither preferred nor ignored.

3. LEVELS OF SENTIMENT CLASSIFICATION

There are three different levels of sentiment classification i.e. word level, phrase level and document level sentiment classification .

a. Word Level Classification: This type of classification is done on the basis of the words that indicate the sentiment about the target. Word level classification is based on lexicon-based approach. It only classifies the words which expresses sentiment. The word may be noun, adjective or adverb. Word level classification does not give more accurate classified sentiments.

b. Phrase Level Classification: This classification results in positive as well as negative category. The phrase signifying the attitude is found out from the sentence and the classification is done. But then it sometimes gives incorrect results if a negative word is added in front of the phrase. Aspect level sentiment analysis can be a better way to achieve phrase level classification with accuracy. The phrase is a group of words which builds a meaningful sentence.

c. Document Level Classification: In this type of classification, single document is considered about the prejudiced text. A single evaluation about the single subject from the document is considered. The document may consist of sentences which don't look like an opinion. So the document level classification will not be effective to know the overall opinion.

4. CHALLENGES IN SENTIMENT ANALYSIS

A tweet contains entity, feature/aspect and sentiment. The sentiment word refers to the words used by the user to express their feeling. The feature/aspect is the perspective of how the tweet can be classified whether it is positive, negative or neutral. For example, "The food was good but the restaurant was not great", with this statement if we take it in the aspect of food, it is considered as positive. But if we take it in the aspect of restaurant, it is negative.

- In most of the time, tweets are highly unstructured and non-grammatical which will make it difficult to fetch the tweets by keywords. It will also be difficult to preprocess.
- The use of out of vocabulary words is also a problem which makes it difficult to classify the polarity or emotion of the words.
- Sarcastic sentences always tend to be difficult for the sentiment classification.
- Extensive use of acronyms (asap, omg, lol, rofl, idk, btw) which is also a challenge during classification.

5. LITERATURE SURVEY

Sentiment analysis is the most critical study location in commercial enterprise fields. Formerly research was done for sentiment evaluation in diverse domains like company product, movie critiques, politics and so on. Previous studies like pang et al. has furnished with the baseline for wearing out research in various domains. It uses superstar ratings as polarity signals of their education information. Even many authors have used the equal concept supplied via pang et al.

1) *Earthquake Shakes Twitter Users: Real-Time Event Detection By Social Sensors:*

T. Sakaki et al. [1] developed an event notification machine which monitors the tweets and can provide notifications thinking about the time constraint. They come across real-time events in twitter which include earthquakes. They have got proposed an set of rules to monitor tweets detecting target Occasion. Each twitter consumer is taken into consideration as a sensor. Kalman Filtering and particle filtering are used for estimation of location.

Record set:

For type of tweets, we organized 597 nice examples which record earthquake incidence as a training set.

Positive aspects:

1. Primary project of earthquake detection is carried out using the machine. Customers are registered with it and e-mail messages are sent to them.

2. The two filtering strategies come across and provide estimation for location.

Negative aspects:

1. More than one activities cannot be detected at a time.

2. It can't provide advanced algorithms to expand queries.

3. Constrained to most effective one goal occasion detection at an unmarried time event.

4. It makes use of svm as a classifier into fantastic and bad sentiments which isn't always relevant to small statistics sets.

2) Event summarization using tweets:

Previous studies could not be contributing to come across hidden time activities in repeating events consisting of sports activities. Right here the purpose is to extract some tweets that describe the most essential degrees in that occasion.

Chakrabarti and Punera [2] have defined the variation in hidden markov model (hmm) in summarizing the occasion from tweets. It offers the continuous tweet illustration for intermediate stages applicable for an event. Right here three algorithms are used to summarize the applicable tweets. Hmm offers hidden occasions.

Statistics set:

Tweets between the durations Sep12th, 2010 to Jan 24th, 2011 containing the names of NFL teams.

Positive aspects:

1. It offers advantages for preceding techniques of matching queries.

2. It is far maximum useful for one shot occasions like earthquakes.

3. It tackled problems like construction of actual time summaries of activities.

4. It identifies an underlying hidden country representation of an occasion.

Negative aspects:

1. It isn't always applicable to discover non-stop time photographs found in tweets.

2. It does not provide the minimum set of tweets which can be applicable to an event.

3. It cannot provide summary of unknown occasions which cannot be anticipated.

4. In this model noises and historical past subjects cannot be removed.

3) Et-Lda: Joint Topic Modeling For Aligning Events And Their Twitter Feedback:

Twitter has become the widely used micro-running a blog website to percentage the evaluations. In this work, y.hu et.al [3] has proposed a joint bayesian model et-lda this is event-subject matter lda which plays the project of topic modeling and occasion segmentation if you want to carry out sentiment analysis quantitatively and qualitatively. Here y.hu et.al, has taken into attention two big scale data units from two one-of-a-kind domain names related to two activities. The work finished right here is most beneficial for subject matter modeling due to the fact the subject matter might also includes many paragraphs wherein the tweet may additionally belong to specific occasion in a paragraph or widespread event in the topic. So as to do the sentiment analysis correctly without misconceptions the event subject matter version is very beneficial.

Statistics units:

Big scale information sets related to activities from two exceptional domains :(1) president obama's speech on 19 may also 2011 and (2) republican number one debate on sept 7, 2011. Above datasets include 25,921 and 121,256 tweets respectively.

Positive aspects:

1. The baseline ldatr eats events and tweets separately even as et-lda treats them relating to each

Different. Hence the venture of locating polarity and sentiment evaluation gives more correct outcomes.

2. The simple task of occasion modeling and segmentation of events is accomplished efficaciously.

Negative aspects:

Tweets are modeled as binomial mixture wherein tweets in which most phrases belong to standard topics are taken into consideration as well known tweets and tweets wherein maximum

Phrases belong to particular occasion as specific tweets. It's far totally unreasonable for tweets having brief lengths.

6. METHODOLGY

In machine learning, sentiment analysis can be done using unsupervised learning and supervised learning.

1) Unsupervised learning

Unsupervised learning is more likely known as the real artificial intelligence – The ability of the computer to learn and know complex processes and patterns without human guidance. Though unsupervised learning is complex for some simpler use cases, it helps on solving problems that humans cannot normally tackle. Some examples of unsupervised machine learning algorithms include k-means clustering and association rules.

Unsupervised learning problems can be grouped into clustering and association problems.

Clustering: A clustering problem is discovering the inherent groupings in the data, such as grouping employers by their performance.

Association: An association rule learning problem is discovering rules that describes large portions of our data, such as people that buy A also tend to buy B.

Some well-known examples of unsupervised learning algorithms are:

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

2) Supervised learning

Supervised learning is the generally used method among the two. It includes algorithms such as linear and logistic regression, multi-class classification, and support vector machines. It is named as supervised learning because the data scientists will act as guide to teach the algorithm what results it should conclude with. For supervised learning, the algorithm's possible outputs should be already known and the data used to train the algorithm should be already labeled with correct answers. For example, a classification algorithm will learn to identify persons after being trained on a dataset of images that are properly labeled with the names of the persons and some identifying characteristics.

Supervised learning problems can be grouped into regression and classification problems.

Classification: A classification problem is when the output variable is a category, such as "red" or "blue" or "present" and "absent".

Regression: A regression problem is when the output variable is a real value, such as "pounds" or "rupees".

Some eminent examples of supervised machine learning algorithms are:

- Linear regression (for regression problems).
- Random forest (for classification and regression problems).
- Support vector machines (for classification problems).

Choosing to use either supervised or unsupervised machine learning algorithm typically depends on the factors related to the structure and amount of your data and the use case of the issue at hand [6]. Both types of algorithms can be used to build predictive data models that help us make decisions across various challenges.

7. ACCURACY OF ALGORITHM

Accuracy of any method or approach is determined by the correctness of the prediction i.e., If a positive tweet is correctly predicted as positive, it is called True Positive (TP). But if it gets it wrong, it is called False Positive (FP). Similarly, If a negative tweet is correctly predicted as negative, it is called True Negative (TN). But if it gets it wrong, then it is treated as False Negative (FN).

$$Accuracy = (TN + TP)/(TN + TP + FN + FP)$$

8. TWITTER

Twitter is a widely famous and well-known social networking platform and micro blogging service which lets the user to express their opinions and feeling through their posts, which are commonly known as Tweets. Tweets are relatively small messages which had a limit bound of 140 characters previously but now it has been doubled to 280 characters for all languages except Japanese, Korean and Chinese on November 7, 2017. People also use a lot of acronyms, emoticons, short words in their tweets to express their feeling.

Following are the some of the terminologies that used in tweets

Target: Twitter users uses the symbol "@" to refer the target user or micro blogger that will automatically alert the target user.

Emoticons: Emoticons are the pictorial representations of the feeling that are used to convey the feeling of the user quickly.

Hash tags: Hash tags are usually used to mark up the important topics. Hash tags increase the visibility of their tweets.

Sentiment Analysis

Word level and document level classification may also produce inaccurate results. Sometimes it is insufficient in many applications. Hence, we need to understand the sentiments of the tweets based on analysing the aspect and sentiment of the opinion.

For e.g., "yes…. I love #rainbow". In this tweet, #rainbow is the entity, love is the sentiment. The opinion on this general aspect is positive.

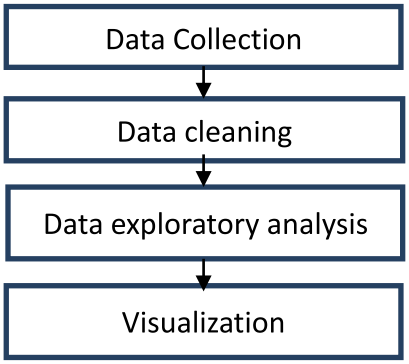

Fig.1. Modules of sentiment analysis

Sentiment Analysis Process

Data extraction: Twitter contains huge amount of data. Therefore, we need to extract the tweets on a particular topic from the twitter API.

Data pre-processing: This technique involves in cleaning of data by removing the punctuations, stem words, spell correction etc.

Applying algorithms: Algorithms are applied to categorize the tweets based on the polarity and emotion of the tweets.

Visualization: The result of the sentiment classification is represented in the graphs.

9. FUTURE ENHANCEMENTS

In future, this study can be extended by classifying the tweets in aspect level analysis. Aspect level sentiment analysis can be helpful to classify even a single tweet with mixture of emotions. Considering the aspect of a tweet, the accuracy of sentiment analysis will be more precise.

10. CONCLUSION

This study shows how the research on sentiment analysis on social media data has evolved in years. Analysis is carried out based on word-level for a specific television program data. In future, this study clearly demonstrates how the aspect-level analysis will be helpful to achieve accuracy and precision.

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes twitterusers: Real-time event detection by social sensors*, in Proc. 19thint.Conf. WWW, Raleigh, NC, USA, 2010.

[2] D. Chakrabarti and K. Punera, *Event summarization using tweets*,in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona,Spain, 2011.

[3] Y. Hu, A. John, F. Wang, and D. D. Seligmann, *Et-lda: Joint topicmodeling for aligning events and their twitter feedback*, inProc.26th AAAI Conf. Artif. Intell. Vancouver, BC, Canada, 2012.

[4] Asmaa Mountassir , Houda Benbrahim, Ilham Berrada, *Anempirical study to address the problem of unbalanced data sets insentiment classification*,IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX,Seoul, Korea.

[5] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan*Interpreting the Public Sentiment Variations on Twitter*, IEEETransactions on Knowledge and Data Engineering, VOL. 26, NO.5, MAY 2014.

[6] Alexandre Trilla, Francesc Alias, *Sentence-Based SentimentAnalysis for Expressive Text-to-Speech*, IEEE TRANSACTIONSON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL.21, NO. 2, FEBRUARY 2013.

[7] Rui Xia, Feng X, Chengqing Zong, Qianmu Li, Yong Qi, Tao Li,*Dual Sentiment Analysis: Considering Two Sides of One Review*,IEEE TRANSACTIONS ON KNOWLEDGE AND DATAENGINEERING, VOL. 27, NO. 8, AUGUST 2015.

[8] Rasheed M. Elawady, Sherif Barakat, Nora M.Elrashidy,"*DifferentFeature Selection for Sentiment Classification*, “InternationalJournal of Information Science and Intelligent System, 3(1): 137-150, 2014.

[9] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, Ming Zhou,“*A Joint Segmentation and Classification Framework for SentenceLevel Sentiment Classification*,"IEEE/ACM TRANSACTIONSON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL.23, NO. 11, NOVEMBER 2015.

[10] X.Wang, F.Wei, X. Liu, M. Zhou, M. Zhang,"*Topic sentimentanalysis in twitter: A graph-based hashtag sentiment classificationapproach*,” in Proc.20th ACM CIKM, Glasgow, Scotland, 2011.

[11] Brendan O’Connor, Ramnath Balasubramanyan,"*From Tweets toPolls: Linking Text Sentiment to Public Opinion Time Series*,"Proceedings of the International AAAI Conference on Weblogsand Social Media, Washington, DC, May 2010.

[12] M. Hu, B. Liu,"*Mining and summarizing customer reviews*,"Proc.10th ACM SIGKDD, Washington, DC, USA, (2004).