Journal of Advances in Science and Technology
Vol. 14, Issue No. 2, September-2017, ISSN 2230-9659

Review on Sentiment Analysis Techniques in Data Mining Domain

Anup Haribhau Raut[1]* Dr. Rahul K. Pandey[2]

[1] PhD Student, Maharishi University of Information Technology, Lucknow

[2] PhD Guide, Maharishi University of Information Technology, Lucknow

Abstract – We know that the internet is a collection of networks, and the use of the internet has changed the way people express their thoughts and feelings. People are connecting with each other with the help of the internet through blog posts, online conversation forums, and many more. Sentiment analysis is mainly concerned with the identification and classification of opinions or emotions of each post. Sentiment analysis is broadly classified in the two types: first one is a feature or aspect based sentiment analysis and the other is objectivity based sentiment analysis. To correctly classify the tweets, machine learning technique uses the training data. So, this technique does not require the database of words like used in knowledge-based approach and therefore, machine learning technique is better and faster. Several methods are used to extract the feature from the source text. Feature extraction is done in two phases: In the first phase extraction of data related to twitter is done i.e. twitters specific data is extracted. Now by doing this, the tweet is transformed into normal text. In the next phase, more features are extracted and added to feature vector. Each tweet in the training data is associated with class label. This training data is passed to different classifiers and classifiers are trained. Then test tweets are given to the model and classification is done with the help of these trained classifiers. So finally, we get the tweets which are classified into the positive, negative and neutral. In this paper, our goal is to present the study on different sentiment analysis methods and feature extraction methods designed by various researchers. Additionally, we are presenting the current research gap based on analysis of all recent methods of sentiment analysis. The outcome of this paper is the current research problems and motivation for sentiment analysis in data mining.

Key Words- Tweet, Classifier, Extract, Sentiment Analysis, Feature.

---------------------------♦-----------------------------

1. INTRODUCTION

Now days, many people are utilizing social network locales like Facebook, Twitter, Google In addition, and so forth to express their feelings, conclusion and offer perspectives about their day by day lives. Through the online groups, we get an intelligent media where shoppers illuminate and impact others through gatherings. Social media is producing a substantial volume of sentiment rich information as tweets, status updates, blog entries, remarks, surveys, and so forth. Besides, social media gives a chance to businesses by giving a stage to interface with their clients for promoting. People for the most part rely on client created content over online, all things considered, for basic leadership. For e.g. on the off chance that somebody needs to purchase an item or needs to utilize any administration, at that point they initially look into its surveys on the web, talk about it on social media previously taking a choice. The measure of substance produced by clients is excessively tremendous for an ordinary client, making it impossible to break down. So there is a need to mechanize this, different sentiment analysis techniques are generally utilized. Sentiment analysis (SA) educates client whether the data regarding the item is tasteful or not before they get it. Advertisers and firms utilize this analysis information to comprehend about.

Textual Data recovery techniques predominantly center on preparing, searching or dissecting the real information exhibit. Actualities have a target segment in any case; there are some other textual substances which express subjective attributes. These substances are principally assessments, sentiments, examinations, dispositions, and feelings, which frame the center of Sentiment Analysis (SA). It offers numerous testing chances to grow new applications, for the most part because of the gigantic development of accessible data on online sources like sites and social networks. For instance, recommendations of things proposed by a recommendation framework can be anticipated by considering contemplations, for example, positive or negative suppositions about those things by making utilization of SA.

With the increase in the popularity of social networking, micro-blogging and blogging websites, a huge quantity of data is generated. We know that the internet is a collection of networks, and the age of the internet has changed the way people express their thoughts and feelings. People are connecting with each other with the help of the internet through blog posts, online conversation forums, and many more. The quantity of information is unreasonable for a normal person to analyze using naive technique. Sentiment analysis is mainly concerned with the classification and identification of opinions or emotions of each post. Sentiment analysis is broadly classified in the two types: first one is a feature or aspect based sentiment analysis and the other is objectivity based sentiment analysis. The posts related to movie reviews come under the category of the feature based sentiment analysis. Objectivity based sentiment analysis does the exploration of the posts which are related to the emotions like hate, miss, love etc. In general, various symbolic techniques and machine learning methods are used to analyze the sentiment from the twitter data. So in another way we can say that a sentiment analysis is a system or model that takes the documents that analyzed the input, and generates a detailed document summarizing the opinions of the given input document. In the first step pre-processing is done. In pre-processing we are removing the stop words, white spaces, repeating words, emoticons and #hash tags. To correctly classify the tweets, machine learning technique uses the training data. So, this technique does not require the database of words like used in knowledge-based approach and therefore, machine learning technique is better and faster. Several methods are used to extract the feature from the source text. Feature extraction is done in two phases: In the first phase extraction of data related to twitter is done i.e. twitters specific data is extracted. Now by doing this, the tweet is transformed into normal text. In the next phase, more features are extracted an added to feature vector. Each tweet in the training data is associated with class label. This training data is passed to different classifiers and classifiers are trained. Then test tweets are given to the model and classification is done with the help of these trained classifiers. So finally, we get the tweets which are classified into the positive, negative and neutral.

In data mining Sentiment analysis has been very important topic, while the prevalence of social networking, more and more research is being done in the area of tweet analysis. The profitable data recovered from person to person communication destinations can be used from numerous points of view one of which can be to examine, comprehend and anticipate the market for particular items which is exceptionally basic to enhance characteristics of the separate item. Because of examination, certain interpersonal interaction destinations are constantly refreshing their client subordinate protection strategies for their clients, which thusly are turning into somewhat of a test for mining them. Subsequent to gathering the coveted data the most imperative part comprehends the substance of this data. NLP is a field of software engineering and phonetics worried about the co-operations amongst PCs and human (regular) languages (Alexander and Paroubek, 2010). One particular application in NLP that can be utilized for this design is slant investigation. It can be utilized to distinguish and extract subjective data from the data source gathered. With every one of these procedures and strategies, it is conceivable to manufacture a framework which can extract application subordinate data, process it and create information which can be utilized for considering and conclusions in view of the data recovered.

2. RELATED WORKS

Richa Bhayani (2009)

In they present a novel approach for consequently classifying the sentiment of Twitter messages. These messages are named either positive or negative as for a question term. This is helpful for buyers who need to re-look through the sentiment of products before buy, or organizations that need to screen the public sentiment of their brands. They introduce the consequences of machine learning algorithms for classifying the sentiment of Twitter messages utilizing far off supervision. Our preparation data comprises of Twitter messages with emoticons, which are utilized as loud labels. This kind of preparing data is liberally accessible and can be acquired through mechanized means. We demonstrate that machine learning algorithms (Guileless Bayes, Most extreme Entropy, and SVM) have accuracy over 80% when prepared with emoji data. This paper likewise describes the preprocessing steps required to accomplish high accuracy. The primary contribution of this paper is utilizing tweets with emoticons for far off supervised learning.

Alexander Pak (2010)

In [2] this paper, they center on utilizing Twitter, the most prominent micro blogging stage, for the undertaking of sentiment analysis. We demonstrate to naturally gather a corpus for sentiment analysis and opinion mining purposes. They perform phonetic analysis of the gathered corpus and clarify found wonders. Utilizing the corpus, we manufacture a sentiment classifier that can decide positive, negative and neutral sentiments for a report. Experimental evaluations demonstrate that our proposed techniques are effective and perform superior to already proposed strategies. In our exploration, we worked with English; be that as it may, the proposed system can be utilized with some other dialect.

James Spencer (2012)

In [3] detest Sentiment or, a device for sentiment analysis of Twitter data. Sentiment or uses the naive Bayes Classifier to arrange Tweets into positive, negative or objective sets. They display experimental evaluation of our dataset and arrangement comes about, our discoveries are not contradictory with existing work.

Kouloumpis, Efthymios (2011)

In [4], we explore the utility of semantic features for identifying the sentiment of Twitter messages. We assess the handiness of existing lexical assets and in addition features that capture information about the casual and imaginative dialect utilized as a part of micro blogging. They adopt a supervised strategy to the issue, yet use existing hash labels in the Twitter data for building preparing data.

Sascha Narr, (2012)

In [5], they analyze the characteristics and attainability of a dialect autonomous, semi supervised sentiment grouping approach for tweets. We utilize emoticons as loud labels to produce preparing data from a totally crude set of tweets. They prepare a Naive Bayes classifier on our data and assess it on more than 10000 tweets in 4 dialects that were human explained utilizing the Mechanical Turk stage. As a feature of our contribution, we make the sentiment evaluation dataset publicly accessible. We show an evaluation of the performance of classifiers for every one of the 4 dialects and of the effects of utilizing multilingual classifiers on tweets of blended dialects. Our ex-pediments demonstrate that the order approach can be connected successfully for numerous dialects without requiring additional exertion per extra dialect.

Celikyilmaz et al (2010)

In [6], author proposed pronunciation based word grouping technique for normalizing loud tweets. In pronunciation based word grouping, words having comparable pronunciation are bunched and doled out normal tokens. They likewise utilized content processing techniques like doling out comparable tokens for numbers, html joins, client identifiers, and target organization names for standardization. Subsequent to doing standardization, they utilized probabilistic models to distinguish extremity lexicons. They performed order utilizing the BoosTexter classifier with these extremity lexicons as features and acquired a lessened mistake rate.

Wu et al (2011)

In [7], impact likelihood model for twitter sentiment analysis is presented. On the off chance that @username is found in the body of a tweet, it is impacting activity and it adds to affecting likelihood. Any tweet that starts with @username is a retweet that speaks to an affected activity and it adds to impact likelihood. A solid correlation between these probabilities is watched.

Pak et al. (2010)

In [8], to automatically gather tweets utilizing Twitter Programming interface and automatically explain those utilizing emoticons, author outlined a twitter corpus. They manufactured a sentiment classifier in view of the multinomial Naive Bayes classifier that utilizations N-gram and POS-labels as features utilizing that corpus. In that strategy, there is a shot of mistake since feelings of tweets in preparing set are labeled exclusively in view of the extremity of emoticons. The preparation set contains just tweets having emoticons so it is likewise less effective.

Xia et al. (2011)

In [9], author used the ensemble system for sentiment characterization. Ensemble system is acquired by joining different feature sets and grouping techniques. In that work, they utilized two kinds of feature sets and three base classifiers to shape the ensemble system. Two sorts of feature sets are made utilizing grammatical form information and Word-relations. Naive Bayes, Most extreme Entropy and Bolster Vector Machines are chosen as base classifiers. They connected diverse ensemble techniques like settled blend, weighted mix and Meta-classifier mix for sentiment grouping and got better accuracy.

Neethu M S et.al (2013)

In [10], author proposed the way to deal with analyze the twitter posts about electronic products like mobiles, PCs and so on utilizing Machine Learning approach. By doing sentiment analysis in a particular area, it is conceivable to distinguish the impact of space information in sentiment order. They outlined novel feature vector for classifying the tweets as positive, negative and extract people groups' opinion about product. They exhibited the relative investigation among various classifiers against their proposed classifier regarding precision, recall and accuracy performances.

3. COMPARATIVE ANALYSIS

Table1 is showing the comparative study among different methods discussed in this paper with their drawbacks.

Ref. No.	Classifier	Evaluated Metrics	Limitations
1	Naive Bayes, Maximum Entropy, and SVM	Accuracy, precision, recall	Accuracy not good.
2	multinomial Naive Bayes classifier	Accuracy, fmeasure with unigrams, bigrams, and trigrams	It not works on multilingual corpus.
3	Naïve Bayes classifier	True positive, False positive	
4	SVM	Fmeasure	Two fold techniques is use so time consuming.
5	3-way sentiment classification	Accuracy, precision, recall	Not a deep learning technique.
6	Knowledge-based approach	Sentiment score	It is difficult task if constantly encounter new words
8	KNN, Entropy, and SVM	Negative influenced probabilities, Positive influenced Probabilities.	Not limit to static influence computing,
9	Naive Bayes, Maximum Entropy, and SVM	Accuracy with different classifier	Feature selection for syntactic relations is issue.

4. SENTIMENT ANALYSIS

The A process that done automatic mining of attitudes, views, opinions and emotions from speech, text, tweets and database sources through Natural Language Processing (NLP) is called Sentiment analysis. Sentiment analysis involves classifying opinions in text into categories like ""negative" or “positive" or "neutral". It's also referred as subjectivity analysis, appraisal extraction and Sentiment Analysis. The words opinion, view, sentiment and belief are utilized interchangeably but there are various between them.

- Opinion: A conclusion opens to dispute (because different experts have different opinions)
- View: subjective opinion
- Belief: deliberate acceptance and intellectual assent
- Sentiment: opinion representing ones feelings

Sentiment Analysis is a concept that may contains many tasks such as sentiment extraction, subjectivity classification and sentiment classification, summarization of opinions or opinion spam detection, among others. It aims to analyze people's attitudes, opinions emotions, sentiments etc. towards elements such as, products, individuals, organizations, topics, and services.

Sentiment Analysis extracts opinionated text datasets summarizing them in an understandable form for end users. It extracts “positive”, “negative” or “neutral opinions” from unstructured data. It involves computational management of opinion and text subjectivity. Natural Language Processing (NLP) handles text element processing which is transformed to machine format by NLP. Artificial Intelligence (AI) uses NLP provided information applying math’s to determine whether an opinion is positive or negative. Various methods exist to determine a user’s view on topics from natural language textual information. Sentiment Analysis tracks the mood of the people concerning a specific product or topic. This provides automatic extraction of opinions, emotions and sentiments in text and tracks attitudes and feelings on the web. Tracking products or brands and determining whether they are positive or negative can be done using the web. Sentiment Analysis is referred by many names like “opinion extraction”, “sentiment analysis”, “subjectivity analysis”, “sentiment mining”, “affect analysis”, “emotion analysis” and “review mining”. But, all come under Sentiment Analysis.

Many approaches are used in Sentiment Analysis, the most common being “lexicon based and machine learning”. In lexicon, simple text representation is a “bag-of-words” approach where documents are considered as a collection of all words without help of relations between single words. Sentiment orientation and words are resources associating to Opinion lexicons. The method's drawback is that a word considered positive and negative depends on a situation.

Most of the existing methods have used the terms Sentiment Analysis and sentiment analysis interchangeably. Below defined quintuple of the Sentiment Analysis.

Sentiment Analysis=(t,s,h,T) ………………….. (1.1)

Where‘t’ is the opinion target, ‘h’ is the opinion holder, ,‘s’ is the sentiment about the target and ‘T’ is the time when the opinion or twitted was posted.

From the above definition we can say that sentiment mining is part of the Sentiment Analysis. Sentiment mining can be a binary type (Ex: subjective vs. objective) or topic detection. In this thesis, when comes to the classification task the term sentiment analysis (SA) will be used. The four broad research dimensions of the sentiment analysis are shown in Figure 1.

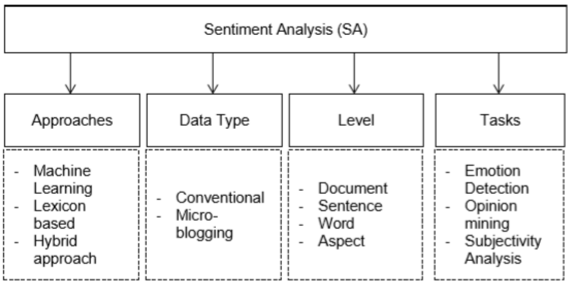

Fig 1 Research dimensions in sentiment analysis

5. CONCLUSION

In this study we presented the important of sentiment analysis, and then discussed the different techniques of sentiment analysis proposed since in last decade. The detailed comparative analysis of such techniques is presented in this project report. After the study of existing methods, their problems have been openly discussed. After the study of all methods, we presented their comparative study in terms of performance metrics, limitations. The research gap presents the research problems associated with existing solutions for sentiment analysis. For future work, we suggest to design model which perform sentiment analysis on different language.

REFERENCES

A. Celikyilmaz, D. Hakkani-Tur, and J. Feng (2010). “Probabilistic model-based sentiment analysis of twitter messages,” in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010.

A. Pak and P. Paroubek (2010). “Twitter as a corpus for sentiment analysis and opinion mining,” in Proceedings of LREC, vol. 2010.

Go, Alec, Richa Bhayani, and Lei Huang (2009). "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1: 12.

Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore (2011). "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11: pp. 538-541.

Narr, Sascha, Michael Hulfenhaus, and Sahin Albayrak (2012). "Language-independent twitter sentiment analysis." Knowledge Discovery and Machine Learning (KDML), LWA: pp. 12-14.

Neethu M. S. & Rajasree R. (2013). "Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India, IEEE.and Economics, Vol. 5 (3).

Pak, Alexander, and Patrick Paroubek (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010

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

Spencer, James, and Gulden Uchyigit (2012). "Sentimentor: Sentiment analysis of twitter data." Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2012.

Y. Wu and F. Ren (2011). “Learning sentimental influence in twitter,” in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

Corresponding Author

Anup Haribhau Raut*

PhD Student, Maharishi University of Information Technology, Lucknow