MAT JOURNALS

Journal of Control System and Control Instrumentation
Volume 5 Issue 1

Survey on Sentiment Analysis Using Machine Learning

Parth Deshmukh[1,*], Adesh Gadge[1], Aniket Ganbote[1], Swapnali Garud[1], Prof. D. S. Kulkarni[2]

[1]Student, Department Of Computer Engineering, Dr. D. Y. Patil College Of Engineering, Savitribai Phule Pune University, Ambi, Pune, India

[2]Professor, Department Of Computer Engineering, Dr. D. Y. Patil College Of Engineering, Savitribai Phule Pune University, Ambi, Pune, India

***Email:** parthdeshmukh725@gmail.com*

DOI: http://doi.org/10.5281/zenodo.2614704

Abstract

Sentiment Analysis is the mining of opinions, sentiments & subjectivity of the context. It is the process of computationally identifying & categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Sentiment Analysis (SA) or Opinion Mining (OM) is the study of people's opinions, monitoring social media & other online resources for customer reviews to understand customer understanding of significant in business analysis. SA or OM is used over a large area so that peoples can make an effective type of decision from other people's reviews. SA plays important role in our day to day life while taking decisions about buying online products & for movie reviews. It is field of study that identifies & extracts subjective information from structured & unstructured texts.

Keywords *Natural Language Processing (NLP); Text Mining; Information Retrieval (IR); Data Mining; Sentiment Analysis; Opinion Mining; Machine Learning (ML); SentiWordNet (SWN); Lexicon; Natural Language Processing Toolkit (NLTK).*

INTRODUCTION

With the growth of internet, it becomes very easy to post our opinions on the web. We can post our opinions after using any product, visiting some place or after watching movies. Other peoples can access these reviews later for decision making. Also, firms or organizations can access these opinions to improve their products or services. Number of documents containing opinions is available on the web, which can provide very meaningful information to the user, to improve products and services or for the betterment of decision making. [1]

A **sentiment** is an attitude that people generally have which is based on their thoughts & feelings. It's an idea or feeling that someone expresses in words. It's a complex combination of feelings & opinions as a basis for action or judgment. Sentiment Analysis is the process of computationally identifying & categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

With the evolution of internet & web Technology, tremendous amount of data is present on the web for the internet users. These users can not only use the available resources on the web, but also give their feedbacks, thus it generates supplemental useful information. Due to this, user's ideas, thoughts, feedbacks & suggestions are now available through the web resources. It's very much essential to identify, analyse & categorize their feedbacks or reviews for the better decision making. [9]

Opinion Mining or Sentiment Analysis is a Natural Language Processing (NLP) &

Information Extraction (IE) process that identifies the user's reviews explained in the form of positive, negative or neutral comments. [9]

Generally, SA tries to determine the sentiment of a writer about some aspect or the overall contextual polarity of a document.

Text mining offers a way to exploit the very large amount of information available on the Internet.[9] Many websites provides user rating & commenting services, & these reviews could reflect users' opinions about a product. In current time people search for other people's opinions from the net before taking purchase decision over a product or before seeing a movie because practically nobody would want to waste his/her money & most importantly time over a wrong decision. Also, user's feedbacks & suggestions could provide essential information for businesses not only to market their products but also to manage their reputations.

A key problem in this area is sentiment classification, where a document is labelled as a positive or negative evaluation of a target object (film, book, product etc.).In recent years, the problem of SA has seen increasing attention. Classification of sentiments is a recent sub field of text classification which is concerned not with the topic a document is about, but with the opinion it expresses. Sentiment classification also goes under different names, among which opinion mining, sentiment analysis, sentiment extraction, or affective rating. [9]

Current-day Opinion Mining & Sentiment Analysis is a field of study at the crossroad of Information Retrieval (IR) & Natural Language Processing (NLP) & share some characteristics with other disciplines such as text mining & Information Extraction.[9] Opinion mining is a process to identify, detect & extract the subjective information in text documents. The sentiment may be his or her idea, judgment, mood or just a thought.

RELATED/PREVIOUS WORKS

In Today's technology era, users' feedbacks, comments or opinions are emerged as sentiment analysis. Sentiment Analysis started in late 2000s, but it is been in effect since 2004 in product reviews area. There are a lot of people who have done research on sentiment analysis, its process & its various techniques. There are some authors who also have worked on the languages other than English. The ones who have performed Sentiment analysis in Indian languages, few of them are listed below.

This is about movie review data in Hindi [1]. It performs opinion mining at document level. It classifies documents in three categories viz., positive, negative & neutral. It used ML & POS tagging. In the POS tagging only adjectives are concerned. Authors have performed both the methods in Machine Learning i.e. supervised & unsupervised. They have taken care of the negation as well.

In [2], authors have done sentiment analysis on mixed language sentences. Here they have considered only two languages i.e. Hindi & English. The analysis has been done in phrase as well as sub-phrase level of the sentence. Grammatical transitions are taken into consideration while predicting the overall sentiment of the sentence.

This is mainly a survey paper [3]. It gives various methods used in opinion mining. It is the summary of what work has been done related to opinion mining in Hindi language. It describes different challenges that need to be overcome while performing opinion mining in Hindi language. Authors have accepted that it is not easy to perform sentiment analysis in Hindi as it is a native language & due to lack of resources.

In this paper sentiment analysis is done on movie reviews in Malayalam language. They have used rule based approach for sentiment analysis. Negation h&ling has done in order to extract sentiment from the review[4]. The corpus is made from Malayalam websites. They have got 85% accuracy.

In the second paper [5] by the same authors as [4], they have taken the research work further ahead. They have used machine learning with rule based approach. They have used two techniques Support vector machine (SVM) & Conditional r&om fields (CRF). In this it is concluded that SVM is more better than CRF.

MACHINE LEARNING APPROACHES

It has mainly two approaches based on Supervised & Unsupervised learning:

Classification based on Supervised Learning

Sentiment classification obviously can be evaluated by a supervised learning problem with three classes - positive, negative & neutral. Reviews can be used as a training dataset. Since every review already has a user comment about product, training dataset are readily available. As an example, a review with positive opinion is considered a positive review; a review with negative statement is considered as a negative review.

Naive Bayes classification & SVM are supervised learning methods that can be applied to SA. This approach can be taken to classify product reviews into two classes positive & negative. It was shown that using unigrams (a bag of individual words) after that as more sufficient data in classification it run skilfully with either SVM or naive Bayes. Some of the supervised ML approaches are as follows:

Terms & their frequency: The number of words & their frequency counts is considered for the classification. Features are individual words or word n-grams & their frequency counts.

Parts-Of-Speech: Grammatical terms in sentence as a verb, adjectives & adverb usage for classification.

Negations: Negations are those words which affect the sentiment orientation of other words in a sentence. Examples include not, no, never, cannot, shouldn't, wouldn't, etc.

Classification based on Unsupervised Learning

Unsupervised learning is based on sentence level sentiment analysis. Lexicon based sentiment analysis. It extracts phrases containing adjectives or adverbs as adjectives & adverbs are good indicators of opinions.

It estimates the semantic orientation of the extracted phrases using the point wise mutual information (PMI)

Support Vector Machine

SVM is a supervised ML algorithm used for classification where the dataset teaches SVM about the classes so that SVM can classify any new data. In this we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Afterwards we perform classification by finding the hyper-plane that differentiates the two classes very well. This works by separating the data into different classes by finding a line which separates the training data set into classes.

IT offers best classification performance on the training dataset. The best thing about it is that it doesn't make any harsh assumptions on data. It renders more effectiveness for correct classification of the future data. It doesn't over-fit the data. It is memory efficient. It is effective in high dimensional spaces [6]. Though, it doesn't perform well,

when we've large dataset because the required training time is higher.

There are ample applications of SVM such as Classification of images, H&-written characters recognition etc. It is commonly used for stock market forecasting by various financial institutions. Also It has been widely applied in the biological & other sciences. As an example, it can be used to compare the relative performance of the stocks when compared to performance of other stocks in the same sector. The comparative study of stocks helps manage investment making decisions based on the classifications made by the SVM learning algorithm.

Naïve Bayes

Naïve Bayes is not only easy but also very fast as it needs less training data. It can be used for making predictions in real time. Naïve Bayes Classifier is one of the most famous ML methods grouped by a similarity that works on the popular Bayes Theorem of Probability- to build ML models particularly for disease prediction & document classification. It's a classification method based on Bayes theorem.

It assumes that a particular feature is independent of other features. This model is easy to build & particularly useful for very large datasets. This model is popularly known to outperform many classification methods. Also it is well known for multi class prediction feature [6].

Its applications include Sentiment analysis, Spam filtering, Classify documents based on topics, Image classifications, Information retrieval, Medical field, Document Categorization etc. Also used for classifying news articles about Sports, Entertainment, Technology, Politics, etc.

Maximum Entropy

It is a probabilistic classifier. It doesn't assume that the features are independent of each other. This is based on the principle of maximum entropy from all the models that fit the training data, selects the one which has the largest entropy.

The principle of ME states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy. This classifier can be used to solve a large variety of text classification problems such as language detection, sentiment analysis, topic classification & more.

Due to the minimum assumptions that the ME classifier makes, we regularly use it when it is unsafe to make any such assumptions. ME requires more time to train datasets as compared to Naive Bayes [6]**.**

ISSUES RELATED TO SA

Issues related to SA are as follows:

Named entity recognition: - NER is a sub-task of IE that seeks to locate & classify named entities into pre-defined groups such as the names of organizations, locations, persons, expressions of times, monetary values, quantities, percentages etc. Problem is we don't know what a person actually wants to convey. [7]

Polarity: Generally the polarity of some particular word or sentence is +ve or -ve or neutral. But how much +ve or -ve is another question. "Good" & "Best" both are positive but second one conveys stronger feelings than the first one. [7]

Sarcasm: Sarcasm is nothing but use of acerbic language to mock. In simple word sarcasm is nothing but using irony & a sneering tone of voice. [7]

Negation handling: Only writing "NOT" in the statement doesn't make it a -ve sentence e.g. "The song is not bad" has +ve opinion, but it may considered as -ve b'cause of the "not" word. [7]

Aspect based Sentiment Analysis: - We should know on which aspect the opinion is expressed so as to evaluate the sentiment. Rather than evaluating sentiment of whole sentence finding the aspect is more important. [7]

Word Ambiguity: Word ambiguity is another pitfall you'll face working on a sentiment analysis problem. Problem with word ambiguity is the impossibility to define polarity in advance as the polarity for some words is strongly dependent on the sentence context. [7]

CONCLUSION

Sentiment analysis (SA) or opinion mining (OM) plays a significant role in business decision making. Many of the organizations & enterprises will take their business decision only based on their customer reviews. There are several techniques for performing sentiment analysis. This paper specifies the sentiment analysis on Machine learning based & gives the clear knowledge about various approaches. This survey gives the knowledge about the sentiment analysis issues such as Polarity shift problem, data scarcity briefly & how they are handled in different domains.

REFERENCES

1. Jha, V&ana, N. Manjunath, P. Deepa Shenoy, K. R. Venugopal, & Lalit M. Patnaik. "Homs: Hindi opinion mining system." In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*, pp. 366-371. IEEE, 2015.
2. Sitaram, Dinkar, Savitha Murthy, Debraj Ray, Devansh Sharma, & Kashyap Dhar. "Sentiment analysis of mixed language employing Hindi-English code switching." In *Machine Learning & Cybernetics (ICMLC), 2015 International Conference on*, vol. 1, pp. 271-276. IEEE, 2015.
3. Sharma, Richa, Shweta Nigam, & Rekha Jain. "Opinion mining in Hindi language: a survey." *arXiv preprint arXiv: 1404.4935* (2014).
4. Nair, Deepu S., Jisha P. Jayan, & Elizabeth Sherly. "SentiMa-sentiment extraction for Malayalam." In *Advances in Computing, Communications & Informatics (ICACCI), 2014 International Conference on*, pp. 1719-1723. IEEE, 2014.
5. Nair, Deepu S., Jisha P. Jayan, R. R. Rajeev, & Elizabeth Sherly. "Sentiment Analysis of Malayalam film review using machine learning techniques." In *Advances in Computing, Communications & Informatics (ICACCI), 2015 International Conference on*, pp. 2381-2384. IEEE, 2015.
6. Ms. A. M. Abirami Dept. of Information Technol, Ms.V.Gayathri, "A SURVEY ON SENTIMENT ANALYSIS METHODS & APPROACH", 2016 IEEE Eighth International Conference on Advanced Computing (ICoAC)
7. Snehal V. Pawar, Prof. Swati Mali," Sentiment Analysis in Marathi Language" International Journal on Recent & Innovation Trends in Computing & Communication.
8. Rachana. Baldania, "Sentiment analysis approaches for movie reviews forecasting: A survey," 2017 International Conference on Innovations in Information, Embedded & Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-6.
9. Manoj Kumar Das, Binayak Padhy & Brojo Kishore Mishra "Opinion Mining And Sentiment Classification: A Review" International Conference on Inventive Systems and Control (ICISC-2017)
10. Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, & Ting Liu," Sentence Compression for spect-Based Sentiment Analysis" IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, & LANGUAGE

PROCESSING, VOL. 23, NO. 12, DECEMBER 2015
11. Fangzhao Wu, Yongfeng Huang, Yangqiu Song, Shixia Liu," Towards building a high quality micro blog-specific Chinese sentiment lexicon", Decision Support Systems-2016.
12. V.K. Singh, R. Piryani, A. Uddin, P. Waila," Sentiment Analysis of Movie Reviews", conference on IEEE-2013.
13. Isidro Peñalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia-Garcia," Feature-based opinion mining through ontologies", Expert Systems with Applications-2014.
14. Efstratios Kontopoulo, Christos Berberidis, Theologos Dergiades, Nick Bassiliades," Ontology- based sentiment analysis of twitter posts", Expert Systems with Applications 40(2013)4065-4074.
15. Ziang Lia, Wei Xu, Likuan Zhang, Raymond Y.K. Lau," An Ontology-based Web Mining Method for nemployment Rate Prediction", Decision Support Systems -2014.