International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Sentiment Analysis Using Twitter Dataset

Kanimozhi P

Department of Computer Science and Engineering, Mount Zion College of Engineering and Technology, Pudukottai, Tamilnadu, India-622507

Abstract: ***Sentiment analysis is an upcoming field of text mining area. Sentiment analysis / opinion mining is the process of tracing opinions, views or suggestions of a particular twitter dataset. Today internet plays a vital role in the world. People are used online applications in their day-to-day life. By means of these online applications huge number of opinions is given by the user. The reviews of twitter dataset which gives the success level of the twitter. There are many algorithms have been used to find the opinion in sentiment analysis. Retrieving documents by subject is goal of information retrieval. There are some aspects of textual content, which form equally valid selection criteria. This paper presents the sentence-level sentiment classification using k-means clustering, CART, C4.5 algorithm.***

Keywords: Sentiment analysis, Sentence level, Opinion mining

1. Introduction

1.1 Data mining

Data mining is a mining knowledge from large amount of data. Data mining uses sophisticated mathematical algorithms to fragment the data and evaluate the probability of future events. Data mining is also called as Knowledge Discovery in Data (KDD).

1.2 Sentiment Analysis

Sentiment analysis is learning of people's emotions, views, attitude, and opinions. It is also called an opinion mining. Sentiment analysis identifies the sentiment articulated in a text then analyzes it. So sentiment analysis to find the opinions, show the sentiment and classify the polarity. Opinion mining used to study the sentiments spoken by people on the internet through reviews. Opinion mining is a type of Natural Language Processing (NLP) for tracking the mood of the public through a particular product. There is a huge literature on sentiment analysis (Pang and Lee, 2008; Liu, 2012), with particular interest in determining the overall sentiment polarity of a document. For example, movie reviews help new users to decide whether the movie is watch or not. Though, the huge numbers of review become information overload absence of automated methods for computing their sentiment polarities. There are three levels in sentiment analysis. Document-level, Sentence-level, and aspect-level.

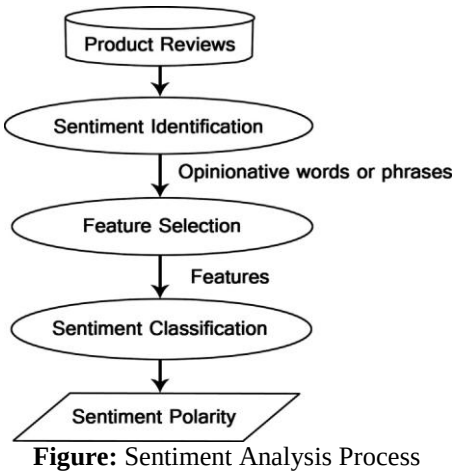

Figure: Sentiment Analysis Process

1) Document-level: It classify the document as positive, negative or neutral. It is known as document-level sentiment classification.
2) Sentence-level: It classify the sentences as positive, negative or neutral. It is known as sentence-level sentiment classification.
3) Aspect-level: It classify the sentiment to the specific aspects of entities. It is known as aspect-level sentiment classification.

1.3 Document-level Sentiment Analysis

Document-level sentiment analysis aims to organize the view text as expressing an optimistic or pessimistic opinion or sentiment. Document-level sentiment classification to classify a textual analysis which is specified on a particular topic. The task is also commonly known as the document-level sentiment classification for the reason that it considers the document as the basic information unit.

1.3.1 *K*-Means Clustering

***k*-means clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum.

1.3.2 CART

Globally-optimal classification tree analysis (GO-CTA) (also called hierarchical optimal discriminant analysis) is a generalization of optimal discriminant analysis that may be used to identify the statistical model that has maximum accuracy for predicting the value of a categorical dependent variable for a dataset consisting of categorical and continuous variables.

1.3.3 C4.5 Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason,

C4.5 is often referred to as a statistical classifier. In 2011, authors of the R-Tool machine learning software described the C4.5 algorithm as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date".

2. Literature Survey

The most prominent work done by Walaa Medhat et al [1] the authors have discussed the feature selection techniques and sentiment classification techniques during information beside through their connected articles referring to several originating references. These fields include Emotion Detection, Building Resources and Transfer Learning. The accuracy percentage of context based SA, which is called as domain dependent data is more than that of domain independent data.

Research in the opinion mining of movie reviews at document level proposed by Richa Sharma et al [2] has mentioned the planned work is directly connected to the Minqing Hu and Bing Liu work on mining and shortening purchaser reviews. The proposed system divided into phases such as (i) Data Collection (ii) POS Tagging (iii) Extracting opinion words and seed list preparation and polarity detection and classification. The conduct experiment outcome shows that the Document based Emotion Orientation System at hand well between outlays to the picture meadow as compared to 'AIRC Sentiment Analyzer'. Upcoming structure makes the truth of 63%.

In the paper Document-level sentiment classification: An empirical comparison between SVM and ANN proposed by Rodrigo Moraes et al[3] focused on comparing SVM and ANN in terms of requirements to achieve better classification in accuracies. Their experiments evaluated both methods as a function of selected terms in a bag of words approach. Although the accuracy between them has never exceeded by 3%, ANN have achieved the greatest organization precision in every datasets. However their results indicated that SVM technique is less affected by noisy terms than ANN, when the data imbalance increases.

Researches in Better document level sentiment study from RST Discourse Parsing planned by Parminder Bhatia et al[4] has presented two different ways of combining RST discourse parse with sentiment analysis. The methods are simple and can use in combination with an "off the shelf" dissertation parses. They consider the following two architectures (i) Reweighting the contribution of each discourse unit and (ii) Recursively propagating sentiment up through the RST parse. Both the construction can be used in grouping with also a lexicon-based sentiment analyzer or a educated classifier. They evaluated on the Pang and Le data and consider only lexicon-based sentiment analysis, obtaining document level inaccuracies between 65% (for baseline) and 72% (for their best discourse-augmented system).

Identifying high impact substructures for convolution kernel in document level sentiment classification proposed by Zhaopeng Tu et al[5] evaluated diverse linguistic structures determined as difficulty kernels for the document level sentiment classification trouble, to use syntactic structures without defining explicit linguistic rules. They explored Subset Tree (SST) and Partial Tree (PT) kernels for component and reliance parse tress correspondingly. The best performance had achieved by combining VK and DW kernels, gaining a significant improvement of 1.45 point in accuracy.

The paper, "Retrieving topical sentiment from online document collections" planned by Matthew Hurst et al[6] has offered a lightweight but robust move toward to combining topic and polarity. The method they used for analyzing a paper was individual language by responsibility a fine-grained NLP based textual study and machine learning classification based approach. This paper strikes away a central point view connecting these two approaches and argues for a union of polarity and topically. The evaluated three aspects of their move toward (i)the presentation of the subject classifier on sentences (ii)the routine of the polarity detection system and (iii)the hypothesis that polar sentences are on area include polar language.

Wei-Hao Lin et al[7] a new problem of learning to name the outlook from which a text written at the document and sentence levels. A large amount of the document's perspective articulated in word procedure, and arithmetical knowledge algorithms such as SVM and Naive Bayes duplicate are clever to profitably determine missing the word patterns that articulate author point of view with high accuracy.

Ainur Yessenalina et al[8] covert variable structured model used for the document sentiment classification assignment. These models do not rely on sentence-level interpretation, and educated jointly to directly optimize document-level precision.

In the paper, "computing sentiment polarity of texts at document and aspect levels proposed by Vivek Kumar Singh et al[9] proposed two methods such as lexicon based and heuristic based scheme. The authors evaluate presentation of four different sentiment analysis schemes on six diverse datasets absent of the four implementations, two be machine learning classifies since the left over two are lexicon-based methods.

In the paper[10] objective of this paper shows that it is to determine the polarity of the movie reviews or criticisms at the document level. The results produced by the system are shortened and supportive for the client in decision making. Experimental outcomes state that the Document-based Sentiment placement system implement healthy in this domain. Opinion mining is very substantial these days from the general man to a business man , everyone is needy on the web. The opinions communicated on the web benefits the users to limit which creation or movie is good for them and it helps the businessman to regulate what the clients thinks about their products . So, it is compulsory to mine this large amount of criticisms and organize them, so it is helpful for them to read and yield conclusions.

This paper [11] covers our participation in the ABSA (Abstract based –sentiment Analysis task of semi level the ABSA task involves of 4 subtasks. For every subtask we suggest both embarrassed and unhindered attitude. The controlled descriptions of our classification are established innocently on machine learning procedures. The proposed approaches accomplish very good results. The constrained varieties were always above average, habitually by the large boundary the unconstrained versions were ranked midst the greatest systems.

The objective of this paper [12] to determine the polarity of the customer reviews of mobile phones at aspect level. The system performs the aspect-based opinion mining on the given reviews and the feature wise brief results created by the system will be supportive for the user in enchanting the decision. Cautious technique results or outcomes indicate that the Aspect based Sentiment orientation system that is capable well and has completed the accurateness of 67%.

In this paper [13] design a deep learning model to analyze the aspect based sentiments and demonstrate competitive or better performance comparing to the results of SemEval'15 in all subtasks. Propose a novel approach to connecting sentiments with the corresponding aspects based on the constituency parse tree. This model also shows promising performance in an unseen domain. In the future work, we are interested in testing the model on other datasets and evaluating the performance of transfer learning. Would also like to explore more sophisticated models in aspect prediction by using adaptive thresholds.

In the [15] a novel method to deal with the problems. An augmented lexicon- based method specific to the twitter data was first functional to accomplish sentiment analysis. Complete Chi-square test on its output, further blinkered tweets could be acknowledged. A binary sentiment classifier is then skilled to disperse sentiment polarizations to the freshly- identified blinkered tweets, whose preparation data is delivered by the lexicon-based method. Realistic experiments display the proposed method is high – effective and hopeful.

3. Proposed System

A sentence based opinion mining classify the document as positive, negative and neutral. It is also handled Naive Bayes(NB). Navie bayes require a high amount organize review as expressing optimistic or pessimistic opinions at sentence level. The running time of NB is higher than that of twitter dataset. Classification model gives the best accuracy among three models. But it requires more training time than Navie bayes.

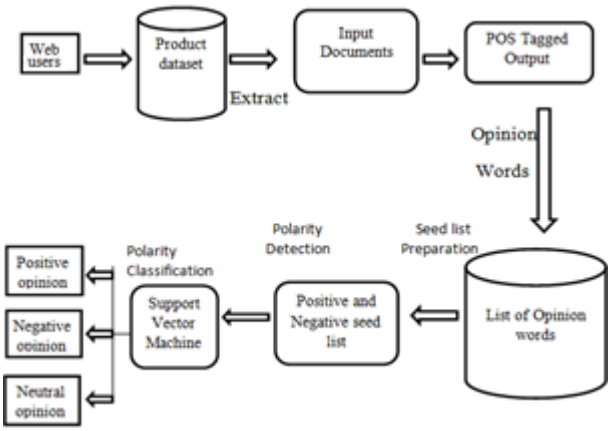

Figure 3.1: Proposed Works

The main goal is to retrieving documents by subject and other content access system. The two standard sentiment analysis datasets shows improvement in performance. The classification task is well modeled by jointly solving an extraction subtask. The experiment uses Sentiment analysis twitter dataset obtained from UCI machine learning repository. The data set consists of total 1038 instance 2 attributes; In this experiment 2 attributes are used.

3.1 R-Programming Tool

This is written in C and FORTRAN, and allows the data miners to write scripts just like a programming language/platform. Hence, it is used to make statistical and analytical software for data mining. It supports graphical analysis, both linear and nonlinear modeling, classification, clustering and time-based data analysis. The R language is widely used among data miners for developing statistical software and data analysis.

3.2 POS Tagging

Once gathering the reviews, they are shown to the POS tagging element where POS taggers that tag all the words of the sentences to their suitable part of speech tag. POS tagging is an essential segment of opinion mining, it is compulsory to fix the structures and opinion words from the reviews. POS tagger is used to tag all the words of reviews or criticisms.

3.3 Extracting Opinion Words

Basically few of the general opinion words along with their polarity that is stored in the seed list . All the opinion words are mined from the tagged (pos) output, The extracted opinion words that are matched with the words stored in the seed list. If the word is not found in the seed list then the synonyms are determined with the help of word net. Every synonym is harmonized with the words in the seed list if any accorded synonym then the mined opinion word is stored with the similar polarity in the seed list.

3.4 K Means Clustering Algorithm

The algorithm has a loose relationship to the *k*-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means due to the

name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by *k*-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

3.5 Cart Algorithm

Classification and regression trees (CART) are a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. Decision trees are formed by a collection of rules based on variables in the modeling data set.

3.6 C4.5 Algorithm

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set f already classified samples. Each sample S_i consists of a p-dimensional vector where the x_j represent attribute values or features of the sample, as well as the class in which s_i falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then re-curses on the partitioned sub-lists.

4. Conclusion

In this paper about the sentence-level sentiment analysis. Users are getting information from the online whether it may be a positive polarity or negative polarity or neutral. Finally they come under specific polarity from the reviews/views/opinions from the online like social media. Sentence may text or emotions or emoticons or expressions. Even though emotions look like a diagram, it represents some textual information. This Naïve Bayes model is easy to build and it is used in large data sets. This Bayes Algorithm is used in Real Time Prediction, Multiclass Prediction, Recommendation system and Text Classification or Spam Filtering or Sentiment Analysis. Hence the NB is performed in Sentence Level sentiment analysis. Nowadays, the document level opinion mining is widely used to perform twitter dataset. It is good at its performance and accuracy is better. Hence the dataset and other documents can be processed and can be classified by their polarity as positive, negative and neutral using NB based sentence level sentiment analysis.

References

[1] Walaa Medhat et al, “Sentiment analysis algorithms and applications: A survey”, Ain Shams Engineering Journal (2014) 5, 1093-1113.
[2] Richa Sharma et al, “Opinion Mining of Movie Reviews at Document Level”, International Journal on Information Theory (IJIT), Vol.3, No.3, July 2014.
[3] Rodrigo Moraes et al, “Document-level sentiment classification: An empirical comparison between SVM and ANN”, Expert Systems with Applications 2012.
[4] Parminder Bhatia et al, “Better document-level sentiment analysis from RST Discourse Parsing”.
[5] Zhaopeng Tu et al, “Identifying High-Impact Sub-Structures for Convolution Kernels in Document-level Sentiment Classification”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. July 2012.
[6] Matthew Hurst and Kamal Nigam, “Retrieving Topical Sentiments from Online Document Collections”.
[7] Wei-Hai Lin et al, “Which Side are You on? Identifying Perspectives at the Document and Sentence Levels”, In Proceedings of the Tenth Conference on Natural Language Learning (CoNLL’06).
[8] Ainur Yessenalina et al, “Multi-level Structured Models for Document-level Sentiment Classification”, proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.
[9] Vivek Kumar Singh et al, “Computing Sentiment Polarity of Text at Document and Aspect Levels”, Ecti Transaction on Computer and Information Technology Vol.8, No.1 May 2014.
[10] B. Pang, L. Lee, and S. Vaithyanathan, (2002), “Thumbs up? Sentiment classification using machine learning techniques” In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[11] B. Liu, “Sentiment analysis and opinion mining,” Proceedings of 5th Text Analytics Summit, Boston, June 2009.
[12] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other Kernel-Based learning methods 1st ed. Cambridge University Press.
[13] Pang, B & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2, 1-135.
[14] Bing Liu, (2012), “Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers”.
[15] Ronen Feldman. 2013. Techniques and applications for sentiment analysis. Communications of the ACM, 56(4): 82-89.