IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 11, November 2019

A survey on Sentiment Analysis with Different Classifiers used for Classification

Neha Rai[1], Pooja Meena[2]

M. Tech Scholar, Radharaman Institute of Technology & Science, Bhopal[1]

Assistant Professor, Radharaman Institute of Technology & Science, Bhopal[2]

Abstract: Reviews are considered to be the most important text message with respect to business purpose. Today almost all the people who are working online give their reviews on various issues. The reviews can be related to movie, product, political party, about any organization and many more. Through this reviews one can come to know the performance of that particular issue. Due to this collection of large amount of reviews, there is a technological advancement has also been done. This advancement is rigorously needed because it is not possible to analyze this large amount of reviews without any technological improvement. In technology world this reviews are known as sentiments and it comes under the category of natural language processing. Machine learning algorithms are widely used for sentiment analysis, so this paper will cover the detail explanation of the sentiment analysis with the algorithm which is used to perform analysis on the sentiments.

Keywords: Sentiment, machine learning, classifiers, accuracy, Naïve Bayes, SVM

I. INTRODUCTION

Sentiment Analysis is a progressing field of research in data mining field. It is the computational treatment of assessments, suppositions and subjectivity of content [1].

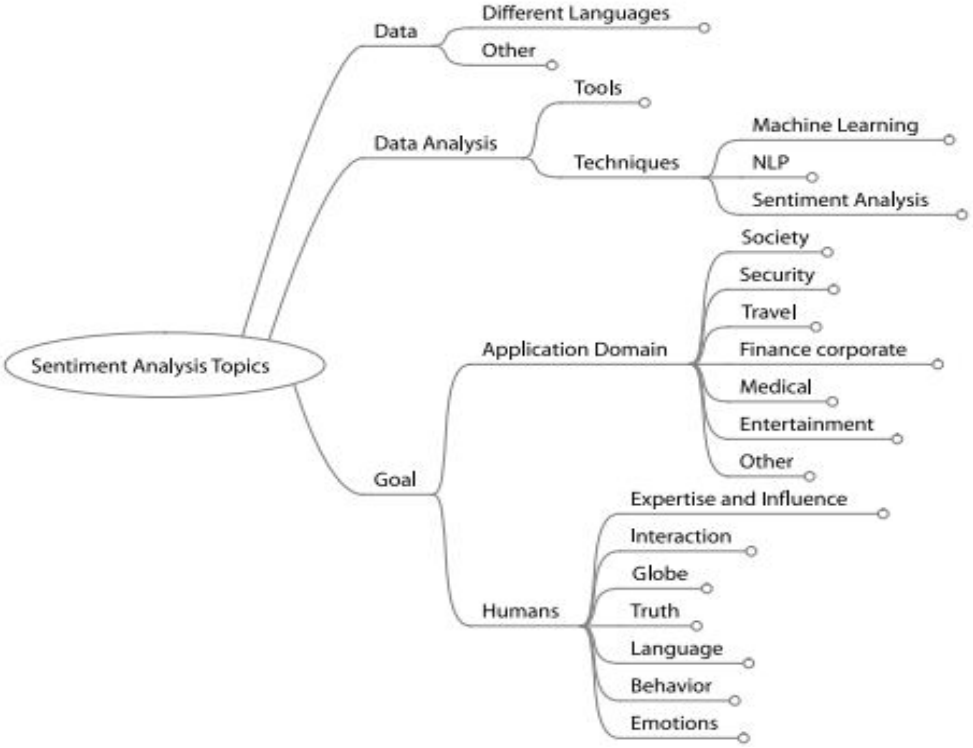

Figure 1 Various phases in sentiment analysis

Number of machine learning algorithms are already been developed in the area of sentiment analysis and different Sentiment based applications are examined and exhibited quickly in this survey. These articles are classified by their commitments in the different Sentiment analysis methods. Fields related to machine learning (transfer learning, emotion detection, and building assets) has pulled the researcher in this area. The fundamental objective of this

overview is to give almost full picture of Sentiment analysis systems and the related fields with brief analysis. The primary commitments of this paper incorporate the complex orders of an enormous number of ongoing articles and the outline of the ongoing pattern of research in the opinion examination and its related regions. Sentiment analysis is one of the new difficulties showed up in automatic language processing with the advent of social networks. Exploiting the measure of data is presently accessible, research and industry have looked for approaches to naturally break down Nowadays, social networks have changed the manner in which individuals express their feelings and purposes of view. This facility is given through literary distributions, online discourse locales, item assessment sites and so on. Individuals depend vigorously on this user produced content. Social networks provide extensive measure of substance produced by the client, it is significant substance for investigation and offer more administrations adjusted to the necessities of users. In the recent years there is lot of improvements in the field of data and opinion exchange has launched the research for the sentiments collected through social network. The analyses of sentiments utilizes, in addition to other things, the recognition of assessments on interpersonal organizations, explaining customer conduct, prescribing items and clarifying the result of the decisions. It comprises of scanning for evaluative messages on the Internet, for example, reactions, proposals and examining the sentiments communicated in that in a programmed or manual way, so as to comprehend open opinion.

A large number of individuals are utilizing Twitter and this has made the most visited websites with an average of 58 million tweets for each day. Although social networks like Facebook, Twitter and Google+ are increasingly associated with many social phenomena such as hate speech, harassment, intimidation, depression or even suicide. That is it is very essential that we must detect this type of potential victims as fast as we can so that we can reinforce the prevention of this type of phenomena using social networks. Sentiment Analysis can be viewed as a classification process, mainly there are three are three classification level in sentiment analysis i.e.: document-level, sentence-level, and aspect-level. In document level classification document is classified on the basis of sentiments that means a whole document that contains information related to any topic are classified. In case of sentence level as the name suggest classification is perform on the single sentence related to any single topic. Last one is related aspect level in which sentiments are classified on the basis of various aspects of entities. Considering the sentence level classification in which first we determine whether the sentence is subjective or objective on the basis of this later on subjective sentences are expressed as positives.

Other than these three levels classification sentiment analysis has to face challenges during analysis like semantic analysis that is evaluated and implement anew semantic similarity to recognize the real aspect of a work that is in different directions. The purpose of this paper is to give detail about the sentiment analysis with different machine learning classification algorithm. The paper is organized as follows: In section 2 there will be literature survey, the section 3 will cover the classification algorithm with the parameters represented during sentiment analysis and the last section will cover the conclusion and future work.

Analyzing the challenges

The goal of the paper indicates what are the different types of challenges arises in sentiment analysis. We identified two classes based on types of goals: first one is Human behavior oriented and the second one is Application domain oriented. Application domain-oriented goals focus on areas that can also be called as "business" domain of sentiment analysis. It was divided into six sections.

- There are some likely words were like: policy, school, elections, tobacco, debate, City, Planning, Citizen, Partnership
- Security words: Words like terrorism, attack, danger, crisis, disaster, Emergency, crime.
- The trip had four topics where the most likely words were
  Words like: airline, travel, tourism, destination, learner, Restaurant, Food, Hotel, Tip.
  There were six topics in finance and corporate where the most likely words where words such as: advertising, brand, sales, firm, Banks, financial forecasting, software projects, share price,
- Medical had three subjects where the most likely words were
  Where words like: disease, health, patient, health care, Drugs, suicide, depression.
- There were five topics of entertainment where most likely
  Words where words like: books, imdb (international film Data base), television programs, sports, players, newspapers, Football, fan, box office.
  Others had six subjects, each of which specified other applications. Domains, such as citation analysis, education, traffic, crowd sourcing. Human and behavior-oriented goals focused on areas that can be used in many application domains. Research here is still goal-oriented rather than data or data analysis methods.
- There were five topics and one sub-topic recommendations / questions in expertise and influence. Most likely and illustrative

The words were like: expert, reputation, leader, follower, questions, ratings, recommendations.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 11, November 2019

II. LITERATURE SURVEY

Sentiment analysis is treated as an errand of regular language handling at a few degrees of granularity. There has been a lot of research on feeling investigation, rule-based approaches, from bag of-words to machine learning algorithms. From a document level in Turney [2] sentence level classification in Hu and Liu [3] and recent sentence level in Wilson et. al. [4]. One of the famous social networking websites is Twitter, through which clients distribute tweets on current scenario and sentiments on any theme. The mining approach can be done at the document level or at the sentence level.

Tsytsarau and Palpanas [5] have stated that Opinion retrieval has built up itself as a significant part of web indexes. Ratings, sentiment patterns and agent feelings advance the hunt understanding of clients when joined with conventional archive recovery, by uncovering more bits of knowledge about a subject. Supposition accumulation over item surveys can be extremely valuable for item showcasing and situating, uncovering the clients' frame of mind towards an item and its highlights along various measurements, for example, time, geological area, what's more, understanding. Following how assessments or exchanges develop after some time can support us distinguish fascinating patterns and examples and better comprehend the manners in which that data is spread in the Internet.

Yu et al. has [6] shown the presence and intensity of emotion words as features to classify the sentiment of stock market news articles. To recognize such words and their power, a logical entropy model is created to grow a lot of seed words produced from a little corpus of securities exchange news stories with feeling explanation. The relevant entropy model estimates the closeness between two words by looking at their logical conveyances utilizing an entropy measure, taking into consideration the disclosure of words like the seed words. Exploratory results show that the proposed strategy can find increasingly helpful feeling words and their comparing force, consequently improving arrangement execution. Execution was additionally improved by the fuse of power into the grouping, and the proposed strategy beat the already proposed pointwise mutual information (PMI)- based development strategies.

Tao et al. has presented a technique that adopts a classification strategy depends on a novel semantic direction portrayal model called S-HAL (Sentiment Hyperspace Analog to Language). S-HAL fundamentally creates a lot of weighted features dependent on encompassing words, and describes the semantic direction data of words by means of a particular component space. Since the technique fuses the thought basic HAL and the speculation checked by the strategy for semantic direction derivation from pointwise shared data (SO-PMI), it can rapidly and precisely distinguish the semantic direction of terms without the utilization of an Internet web index. The aftereffects of an experimental assessment show that our technique beats other known strategies.

Masks and Vossen [7] has developed a vocabulary model for the portrayal of verbs, nouns and adjective words which is to be utilized in applications like opinion feeling and sentiment mining. The model means to depict the point by point subjectivity relations that exist between the entertainers in a sentence communicating separate frames of mind for every on-screen character. Subjectivity relations that exist between the various entertainers are marked with data concerning both the character of the frame of mind holder and the direction (positive versus negative) of the frame of mind. The model incorporates a classification into semantic classes applicable to feeling mining and opinion investigation and gives intends to the distinguishing proof of the mentality holder and the extremity of the frame of mind and for the depiction of the feelings and assumptions of the various entertainers engaged with the content. Unique consideration is paid to the job of the speaker/essayist of the content whose point of view is communicated and whose perspectives on what's going on are passed on in the content. At long last, approval is given by a comment study that shows that these unobtrusive subjectivity relations are dependably recognizable by human annotators.

In paper [8], authors have described the multi-view ensemble approach to SemEval-2017 Task 4 on Sentiment Analysis in Twitter, specifically, the Message Polarity Classification subtask for English. The article is based on voting ensemble where each base classifier is trained in a different feature space. The first space is a bag-of-words model and has a Linear SVM as base classifier. The second and third spaces are two different strategies of combining word embeddings to represent sentences and use a Linear SVM and a Logistic Regressor as base classifiers. The proposed system was ranked 18th out of 38 systems considering F1 score and 20th considering recall.

In paper [9], authors have considered dataset of Twitter (total 1000 comments) and applied various machine learning approach and ensemble approach (majority voting) to classify the comments. They have used twitter specific features as an input to classifier for classification.

Desai and Radhi states that [10] has stated that sentiment refers to the feelings or opinion of person towards some particular domain. Analysis of sentiment (opinions) and its classification based on polarity is a challenging task. Other challenges are overwhelming amounts of information on one topic with all having different representation. Classification and clustering are two major methods applied to perform sentiment analysis of twitter data. They have used Possibilistic Fuzzy C-Means with SVM to improve accuracy on movie tweets and worked on upto 3-grams.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 11, November 2019

III. APPROACHES USED IN SENTIMENT ANALYSIS

Today machine learning and deep learning has made an impact in the world of analysis. Through these two approaches one can perform analysis on the different types of datasets related to various fields. In the same way sentiment analysis is also done through machine learning approaches though which can perform analysis on the sentiments. In the below given below figure it is shown that how sentiment data set go through different steps during analysis. The steps involve different text preprocessing techniques, dimensionality reduction approaches, classification algorithms and different computing parameters.

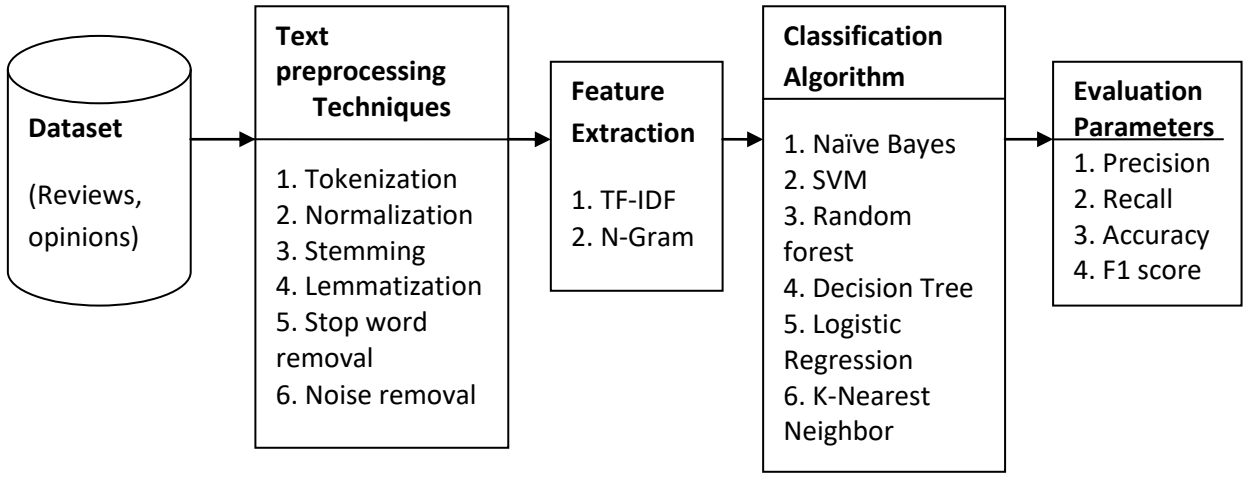

Figure 2. Steps during classification of sentiments

Let us discuss these steps in details:

Dataset – It's a collection of reviews or sentiments or opinions related to any product, movie, organization, political party and many more. These records are mainly stored to perform analysis. In today's world most of the dataset comes under the category of big data [11].

Text preprocessing – The preprocessing of dataset is very much needed during analysis because this is the step which can perform removal of noise, removal of redundancy, handling of missing values etc. Some of the preprocessing techniques which applied during sentiment analysis are as follows:

- Tokenization: This step breaks the large paragraphs called chunks of text is broken into tokens which are actually sentences. These sentences can further be broken into words. For example, suppose there is sentence like, "I am doing M. Tech. degree from Bhopal", but after tokenization this sentence will form the tokens and split into individual words like {"I", "am", "doing", "M.Tech.", "degree", "from", "Bhopal"}.
- Stemming: The stemming process is used to change different tenses of words to its base form this process is thus helpful to remove unwanted computation of words. For example: likes, liked, likely, liking are replaced by like during stemming process.
- Lemmatization – It's the process of merging of two or more words into single word, this process actually analyze the morphology of the word and later on eliminates the ending of the word like caught is replaced by catch, blunders are replaced by mistakes, etc.
- Noise removal – It is observed that almost all the datasets are in raw form, so it is very much needed that there should be some cleaning process along with the help of regular expression of NLP used to remove noises. Sometimes it happens that removal of noise process also remove eliminates a few numbers of rows of the dataset which leads to decreased accuracy.
- Removing Stop words – Stop words are considered as the common words in English language that does not make any contribution during sentiment analysis like is, an, are, that, of, etc.
- Normalization – There are many ways through which normalization can be performed like converting a complete text either into lower case or upper case, removing punctuation and transforming the numbers to equivalent words. Through normalization uniformity of the text can be increased.

Dimensionality Reduction

In case of dimensionality reduction there are two main approaches one is feature selection [12] and the other one is feature extraction [13]. Through this feature selection and feature extraction, classification of the algorithms can also be

improved [14]. In this paper the discussion will be on feature extraction methods one is Term Frequency – Inverse Document frequency (TF-IDF) and N-Gram.

- TF-IDF – This is a well recognized method to evaluate the importance of the word in the document. Term frequency of a specific word is calculated as the number of occurrence of that particular word in the document with the total number of words in the document. Inverse Document frequency is related with the importance of the particular word in the document. The words like “a”, “is”, “an”, “are” etc. occurs frequently in all documents but they don’t have any importance during sentiment analysis. IDF is calculated as IDF (t) = log(N/DF), where N is the number of documents and DF is the number of document containing term t. TF-IDF is a better way to convert the textual representation of information into a Vector Space Model (VSM). Let’s assume a word ‘good’ appears in any document 15 times and the total number of words be 300 in that case term frequency will be 15/300=0.05 and assume that there is total number of 60000 documents and 800 document contains the term happy, then IDF will be 60000/800=75. Computing both these terms then TF-IDF (happy) will be 0.05*75= 3.75.

- N-Gram – This term is defined as formation of features of text in case of supervised machine learning algorithms. There can be sequence of n tokens for the given text. The values of the n can be 1,2,3 and so on. Suppose we are having a sentence “Engineering is a better option for higher secondary students”, in this case when n=1 it is called as unigram, for n=2 called as bigram and n=3 will be trigram. So for the above sentence if we take n=2 then it will produce “Engineering is”, “a better”, “option for”, “higher secondary”, “students”.

Classifiers used in Sentiment Analysis

Classifiers are used for classification of sentiments.

Naive Bayes - This is powerful algorithm for characterization utilized for arranging information on premise of probabilities. With a huge number of records additionally this algorithm works marvelously. It essentially chips away at Bayes hypothesis and utilizations different probabilities to order information. In Naïve Bayes class with most extreme likelihood is viewed as the anticipated class. Naive Bayes is otherwise called Maximum a Posterior Naïve Bayes has different focal points and inconveniences crosswise over various spaces. It is a quick and profoundly adaptable algorithm and It is utilized on both Multiclass and Binary Classification. It can likewise be utilized on little datasets and therefore gives great outcomes [15].

K-Nearest Neighbor - This algorithm is basic and has applications primarily in design acknowledgment, interruption recognition and a lot more are likewise there. In this separation between information purpose of which we need to recognize class is determined utilizing Euclidean separation (different estimates like Manhattan separation and so on.) is determined with the current information focuses and the k closest neighbor (estimation of k is at first chosen can be 3, 4 and so forth.) will decide in favor of the class of new information point. Majority voting will choose the class [16].

Support Vector Machine - This is an effective algorithm for regression as well as classification purpose. It attracts a hyperplane to isolate classes. This algorithm works amazingly well with relapse, the impact of SVM increments as we increment dimensional space. SVM likewise perform well when the measurement number is bigger than the example number. There exists a disadvantage too it doesn't perform well on enormous datasets. SVM broadly utilizes cross-approval to expand its computational efficiency [17].

Random Forest - It is a gathering of choice tree calculations which can be utilized for both classification and regression. In this algorithm for the most part, more trees compare to better execution and productivity. In a given preparing set, remove an example set of information focuses by utilizing bootstrap strategy. After this build a choice tree dependent on the yield of past step. Apply past two stages and we will get number of trees. Each tree built will decide in favor of information point [18].

Decision Tree - This algorithm can be utilized for both classification and regression. The core idea is to partition the dataset into littler subsets and simultaneously tree related is steadily made. This can handle both categorical as well numerical data. We can utilize Gini index too data gain parameter to choose which property will be utilized for further division of dataset. If we use use Gini index than decision tree is called CART (classification and regression tree) and if we use information gain than it is called ID3. This algorithm can be effectively utilized for a sentiment analysis [19].

IV. CONCLUSION

This survey paper is related to information about the sentiment analysis, dimensionality reduction and the different types of classifier used during sentiment analysis. Paper has shown how the sentiment analysis is done and how it pass through various phases during classification. The paper also gives a brief idea about the classification algorithm. In the future we will work on the hate speech analysis in which we classify the hate sentiments through different classification algorithm.

REFERENCES

[1]. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, 2008.
[2]. P. D. Turney, "Thumbs up or thumbs down?," 2001.
[3]. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.
[4]. T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Comput. Linguist.*, 2009.
[5]. M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Min. Knowl. Discov.*, 2012.
[6]. L. C. Yu, J. L. Wu, P. C. Chang, and H. S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Syst.*, 2013.
[7]. I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," in *Decision Support Systems*, 2012.
[8]. E. A. Corrêa Júnior, V. Q. Marinho, and L. B. dos Santos, "NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis," 2018.
[9]. R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," in *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, 2018.
[10]. R. D. Desai, "Sentiment Analysis of Twitter Data," in *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, 2019.
[11]. R. Nair and A. Bhagat, "A Life Cycle on Processing Large Dataset - LCPL," *Int. J. Comput. Appl.*, 2018.
[12]. J. Li *et al.*, "Feature selection: A data perspective," *ACM Computing Surveys*. 2017.
[13]. I. Guyon and A. Elisseefl, "An introduction to feature extraction," *Stud. Fuzziness Soft Comput.*, 2006.
[14]. R. Nair & A. Bhagat, "Feature selection method to improve the accuracy of classification algorithm," *Int. J. Innov. Technol. Explor. Eng.*, 2019.
[15]. D. Jurafsky and J. Martin, "Naive Bayes and Sentiment Classification," *Speech Lang. Process.*, 2017.
[16]. Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, 2018.
[17]. P. W. Wang and C. J. Lin, "Support vector machines," in *Data Classification: Algorithms and Applications*, 2014.
[18]. A. Liaw and M. Wiener, "Classification and Regression with Random Forest," *R News*, 2002.
[19]. C. Bulac and A. Bulac, "Decision Trees," in *Advanced Solutions in Power Systems: HVDC, FACTS, and AI Techniques*, 2016.