Sentiment Classification using Machine Learning: A Survey

Kushall Pal Singh[1]
PG Scholar, Department of computer Engineering and Application,
National Institute of Technical Teacher's Training & Research, Bhopal, India
E-mail: kpal090@gmail.com

Sanjay Agrawal[2]
Professor, Department of computer Engineering and Application,
National Institute of Technical Teacher's Training & Research, Bhopal, India
E-mail: sagrawal@nitttrbpl.ac.in

Abstract—The World Wide Web has brought a new way of expressing the reactions of people about any product, things, and topics, etc. The sentiment Analysis of written textual content on the web is one of the text mining areas used to find out sentiments in a given text. The process of sentiment analysis is a task of detecting, extracting and classifying critiques and sentiments expressed in texts. Twitter is also a medium with the huge amount of information wherein users can view the opinion of other users that labeled into different sentiment classes such as positive, negative, and neutral and are increasingly more developing as a key element in decision making. "Till now, there are few extraordinary problems predominating in this research community, namely, sentiment classification, feature-based classification and dealing with negations. This paper presents a survey covering the strategies and techniques of sentiment classification and demanding situations appear within the area."

Keywords-*Machine learning; Sentiment Classification; Naive Bayes; Max Entropy; SVM; Boosted Tree;*

*****

I. INTRODUCTION

This research paper covers the analysis of the contents on the web covering allots of areas which are growing exponentially in numbers as well as in volumes as sites are dedicated to particular varieties of products and that they specify in gathering customers reviews from numerous websites inclusive of Amazon, and many others.

In recent times, people purchase products online and giving reviews about that product. Analysis of reviews is known as sentiment analysis. Sentiment analysis is also referred as opinion mining. Opinion mining is the assignment of judging whether document expresses a positive or a negative opinion (or no opinion) regarding a particular product. The scope of expressing person thoughts is often restrained when humans have recognized to given evaluations about a product in form of score/ megastar score. But when a person is authorized to explicit evaluation in form of open textual content he can be very precise about what aspects of the product are good and what aren't. Sentiment analysis engines parse through these textual reviews and generate output in the form of polarities e.g. positive, negative or neutral. This helps in finding the reasons at the back of vital fluctuations in sales of products and they can be rectified accordingly. Sentiment analysis is a task that involves facts extraction from customer feedback and other authentic sources like survey groups. Because the phrase inspiration it includes detecting sentiments of any people from the textual content that is written in digital format. There are wide arrays of applications of this concept. This concept becomes the center of attention since industry got re-evolution with the change in the paradigm of seller's market to buyer's markets in order as a way to seize market share.

There are numerous demanding situations in Sentiment analysis. The first is an opinion word that is considered to be positive in one situation may be considered negative in some other situation. A second challenge is that humans don't usually express opinions in the same manner. Most traditional text processing relies on the reality that small variations between two pieces of textual content do not alternate the means vary an awful lot. In Sentiment analysis, however, "this phone is great could be very one of a kind from "this phone is not great". Human beings may be contradictory in their statements. Most reviews could have both positive and negative remarks. That's extremely conceivable by using analyzing sentences separately. But within the more casual medium like twitter or blogs, the more probably humans are to mix distinct evaluations inside the identical sentence which is easy for a human to recognize, but greater tough for a computer to parse. Sometimes even other people have problem expertise what someone notion primarily based on a short piece of text because it lacks context. In the sentiments negation phrases identification and classifying in accordance that polarity may be a very regular hassle due to this trouble accuracy of classifier could be decreased. In the reviews straight forward sentences are not used such as, the movie was great instead of straight forward sentences some contradictory sentences are used such as the movie was not good, this movie is not good than previous or this movie is not good then X(here X is the name of a particular movie).

There are following types of sentiment classifications [1].

- Document level
- Sentence level
- Attributes level

For sentimental classification purpose, basically, two approaches[1] used such as machine learning techniques and semantics orientation. In the machine learning, two sets of document required one is the training set another one is the test set or we can say input set whose predict the output from the training set. Languages that have been studied in preferred are English and in Chinese language, currently, there are only a few researches finished on sentiment classification for other languages like Arabic, Italian and Thai. This survey focuses on machine learning techniques used for sentiment classification.

For the sake of comfort, the rest of this paper is organized as follows: Section 2 presents the related work done by the authors. Section 3 describes algorithms that are used in sentiment classification. Section 4 represents some applications of sentiment analysis and the last section describes the conclusion and future directions for research.

II. RELATED WORK

G. Gautam[2]describes the process of sentiment analysis and comparison among different classifiers. Author(s) used twitter API's for collecting dataset then pre-processed data feature extract from pre-processed dataset further pre processed dataset and feature list will be passed for training using machine learning techniques. By using WordNet accuracy will be improved 89.9% from 88.2%.

S. Shah[3] proposed an algorithm for sentiment analysis following are main points in the algorithm.

- Topic classification
- Polarity classification
- Emotion analysis for data that's neither positive nor negative

Topic modelling or thing categorization is a big issue in sentiment analysis. Author resolved this by using hash tag classifier.

Author's B R Jadhav et. al[4] used an Adaboost algorithm for training and classification. Adaboost algorithm is a version of boosted tree classifier. Its first profound a data enlargement approach in which each training and test review are reversed, after data expansion authors used dual training algorithm to make use of original and reversed training reviews in sets for analyzing sentiment classifier after this classification algorithm named as Adaboost algorithm, and a dual prediction algorithm to classify the test reviews on test review again Adaboost algorithm is used.

In this paper[5] [6]author(s) used three stages named as Polarity Shifter, Detection, and Elimination, Ensemble for Sentiment Analysis at the document level. First, split every document into a set of sub-sentences and build a hybrid model that employs rules and statistical methods to detect explicit and implicit polarity shifts, respectively. Secondly, proposing a polarity shift removal method, to remove polarity shift in negations. Later, author train base classifiers on training subsets divided by different types of polarity shifts and use a weighted aggregate of the issue classifiers for sentiment classification.

In this paper[7] authors used syntactic parser is an iterative parser, which uses Penn Tree Bank parser to assign Parts of Speech (POS) tags to each word in the sentence. The name entities and idioms worried in a sentence are also recognized in syntactic parsing.

In this paper,[8] Authors proposed a new framework has been proposed to normalize the text and pick out the polarity of textual data as positive, negative or neutral using an ETL (Extract, Transform, and Load) large data tool referred to as Talend. The algorithm advanced focuses on parallelism for overall performance speed and contributes closer to the end result via comparing the accuracy of trendy data set.

III. SENTIMENT CLASSIFICATION

A lot of research exists on sentiment analysis of user opinion data, which particularly judges the polarities of user opinions. The nature language processing strategies (NLP) is used in this area, mainly in the document sentiment detection.

The machine learning methods are applicable to sentiment analysis ordinarily belongs to supervised learning in trendy and textual classification strategies in particular. Some of machine learning techniques have been followed to classify the reviews. Machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), Logistic Regression, Random Forest, and Boosted Tree Classifiers have achieved great success in text categorization.

A. Naive Bayes Classifier

The Naïve Bayes classifier is based totally on Bayes theorem. It's a baseline type algorithm. Naïve Bayes classifier assumes that the classes for classification are impartial. Although that is rarely authentic Bayesian classification has shown that there are a few theoretical reasons for this obviously unreasonable efficiency. The fundamental advantage of Naïve Bayes is it requires low processing memory and less time for execution. It's suggested that this classifier used while training time is a crucial factor in the system. Naïve Bayes simply overestimates the class that sure object belongs too. Assuming that we are the use of it best for making decisions (which is genuine in the case of sentiment analysis problem) the decision making is accurate and the model is beneficial. Naïve Bayes is used as a classifier in diverse real global problems like Sentiment analysis, e-mail junk detection,

e-mail Auto Grouping, e-mail sorting by means of priority, Document Categorization and Sexually specific content detection [13].

$$c^* = \text{avgmax}_c \, PB_{NB}(c|d)$$

In the above equation c^* assign to tweet d.

$$PB_{NB}(c|d) = \frac{(P(c)\sum_{i=1}^{k} P(f|c)^{q_i(d)})}{P(d)} \quad (1)$$

In Equation 1 shows the probabilistic analysis of Naïve Bayes classifier in phrases of sentiment analysis. The feature represents by f, number of feature f_i found in tweet d represents by $q_i(d)$, k is total features in a document, parameter p(c) and p(f|c) are achieve through maximum like hood[2]. This means that in order to locate wherein class we need to classify a new document, we need to estimate the product of the probability of every word of the document given a particular class (likelihood), multiplied by the probability of the particular class (prior). After calculating the above for all the classes of set C, we choose the only with the highest chance.

Binary Multinomial Naïve Bayes is an enhancement version of base Naive Bayes and used whilst frequencies of the words don't pay a key role in our classification. Such as example is Sentiment analysis in which it doesn't depend on how many times someone enters the phrase 'bad' or good' but instead simplest the fact that he does Pre-processed data together with extracted features is provided as input for training the classifier uses of naïve bayes. As soon as training is complete, during classification it provides the polarity of the sentiments. As an example for the review comment "I am happy" it provides Positive polarity as result.

B. M ax Entropy Classifier

Another famous classifier is the Max Entropy Classifier or MaxEnt as few peoples prefer to call it. The concept behind MaxEnt classifiers is that we should decide upon the maximum uniform models that fulfil any given constraint. MaxEnt models are featured primarily models. We use these features to find a distribution over the distinctive classes using logistic regression. The probability of a particular data point belongs to a specific class is calculated as follows [5]:

$$P_{ME}(c|d) = \frac{exp\,[\sum_i w_i, f_i(c,d)\,]}{sum\,(\exp\,[\sum_i w_i, f_i(c,d)])} \quad (2)$$

In equation 2, c is the class (positive or negative), d is the data point we are looking at, and w_i is a weight vector. The weight vectors determine the importance of a feature in classification.

In Max Entropy any number of features will add features like bigrams and phrases to MaxEnt without demanding about feature overlapping. The precept of max entropy is beneficial explicitly simplest when implemented to testable information. Entropy maximization with no testable information takes place below a single constraint: the sum of the probabilities must be one. MaxEnt using Priors are more effective in Natural language processing applications. Maximum entropy maximizes the entropy described on the conditional probability distribution [13]. It even handles overlap feature and is same as logistic regression which finds distribution over classes. It also follows sure feature exception constraints.

C. Support Vector Machine

Support vector machine analyzes the data; define the decision boundaries and makes uses the kernels for computation which are performed in input space. The input data are two sets of vectors of length m each. Then every data represented as a vector is classified in a specific class. Now the challenge is to find a margin between two classes that is some distance from any document. The space defines the margin of the classifier, maximizing the margin reduces indecisive decisions.

SVM also supports classification and regression which are useful for statistical learning theory and it facilities recognizing the factors exactly, that wishes to be taken into account, to recognize it efficiently. Support Vector Machines (SVMs) uses a diverse loss function from Logistic Regression [5]. They're also identified differently (maximum-margin). However, in the traditional way, an SVM with a linear kernel isn't so much different from a Logistic Regression due to the fact the trouble won't be linearly separable. The veracity is that a Logistic Regression can also be used with an exponential kernel, but at that point you might be higher off going for SVMs for practical reasons. Another associated cause to use SVMs is if you are in a relatively dimensional space. The term semantic orientation refers to an actual quantity size of the positive or negative sentiment expressed through a word or phrase. Usually, phrases terms conforming to specific part-of-speech templates representing feasible descriptive composition are used. Unluckily, the higher disadvantage of SVMs is they can be painfully inefficient to train. So, would not commend them for any its problem.

D. Logistic Regression

Logistic regression is a classification set of rules that may be skilled so long as expect the capabilities to be about linear and the problem to be linearly separable. It is able to do some feature engineering to expose great sort of non-linear features into linear an awful lot efficiently. It's also very robust to noise and can avoid over fitting or even do feature choice by using the use of l2 or l1 regularization [5]. Logistic regression can also be used in large statistics eventualities considering the fact that its overall performance is efficient and can be distributed using, for example, ADMM. A final advantage of LR is that the output may be evaluated as a probability. This is something that

comes as a pleasant side effect considering that you could use it, as an instance, for ranking as opposed to class.

Even in a case in which you would not assume Logistic Regression to work 100 %, run an easy l2-regularized LR to come up with a baseline before you pass into the usage of "fancier" processes.

E. Random forest classifier

Random forests are an ensemble learning method for classification that perform with the aid of building a mess of selection trees at training time and outputting the class that is the mode of the training output through individual trees. It produces multi-altitude selection trees at inputting phase and output is generated inside the form of multiple decision trees. The correlation among trees is decreased by using randomly deciding on trees and as a result the prediction strength will increase and leads to boom in performance. The predictions are made by aggregating the predictions of numerous ensemble information units. Studies show that the performance is growing visible constantly [13]. There is no downtrend for the performance of this set of rules in any to be had statistics sets. Applications and actual life of examples Random Forests are huge. There is no single kind for RF data sets. They are able to range from any sort of packages like scientific as well as widespread statistics sets. Decrease is less applicable statistics set to 50% impacts the RF classifications in reducing the result accuracy. RF is a parallelized and multi-middle friendly set of rules. So simultaneous running of different trees is likewise an aid feature. The recognition of this machine increased with practical machine learning research and their related algorithms. Authors came across experimental effects in our observe wherein humans had used Random woodland for Opinion mining and have determined remarkable accuracy in classification in their facts units.

The major blessings of this algorithm can be indexed out as follows:

- Non-parametric so that you don't should worry approximately the linearity of the input statistics set.
- If parameters are there they may be without difficulty entered hence doing away with the want for pruning the trees.
- The classification model is rapid and scalable.
- Significance and relevance of textual content/tokens in a class is routinely generated.
- Robust to inappropriate textual content present in report. A drawback that we came throughout in our studies is that the random wooded area classifier without problems over fits its class.

F. Boosted Tree Classifier

Boosted tree is a classifier that is basically a mixture of Boosting and decision tree. Boosting is a system Meta-learning set of rules for decreasing preconception in supervised learning. In Boosting predictive classifiers are used to develop weighted trees which might be further combined into single prediction models. Boosted trees combine the strengths of algorithms: regression trees (models that relate a reaction to their predictors with the aid of recursive binary splits) and boosting (an adaptive technique for combining many simple models to present improved predictive overall performance). Most boosting algorithms consist of iteratively learning vulnerable classifiers with recognize to a distribution and including them to a final strong classifier [5]. Whilst they're introduced, they're normally weighted in a few ways this is commonly related to the weak learner's accuracy. After a weak learner is added, the statistics is re–rated and new weights are produced and examples which might be incorrectly categorized benefit weight and examples which can be categorized correctly lose weight.

There are numerous variations of Boosting algorithms some of them are – AdaBoost, LP boost, Logit Boost, Gradient Boosted Regression trees and many others. Boosting algorithms such as AdaBoost are recognized to perform nicely worked for classification and are very resistant to over fitting with respect to misclassification errors, even though conditional class probability estimates subsequently diverge to zero and one, implying whole over fit in terms of CCPF (Conditional magnificence probabilities) estimation but not classification [13]. Gradient boosting is a machine learning approach for regression problems, which produces a prediction model in the form of an ensemble of weak prediction fashions, typically decision tree. It builds the model in a level-wise fashion like other boosting techniques do, and it generalizes them by using allowing optimization of an arbitrary differentiable loss function. The gradient boosting approach can additionally be used for classification issues through reducing them to regression with an appropriate loss function. Boosted tree classifier has several advantages like training time required is very much less without compromising on accuracy while training.

Other advantages are it can perform satisfiable for data set with missing values and predictor variable. There some drawbacks such as it cannot compute conditional magnificence probabilities.

IV. APPLICATIONS OF SENTIMENT ANALYSIS

Some of the applications of sentiment analysis consist of online marketing; identify capability risks, Opinion summarization in forums and many others. Online advertising has grown to be one of the most important sales resources of today's internet atmosphere. Sentiment analysis discovers its recent utility in Dissatisfaction orientated online advertising and Blogger-Centric Contextual advertising,

which refers to the task of personal advertisements to any weblog page, selected inconsistent with bloggers hobby.

Whilst faced with monstrous amounts of online records from various online forums, information seekers typically discover it very hard to yield correct information this is beneficial to them. As a way to become aware of capacity risks, it is important for companies to acquire and examine data about their competitors' products and plans. Sentiment analysis finds a major function in competitive knowledge to extract and visualize comparative relations among products from purchaser reviews, with the interdependencies among relations considered, to help companies find out capability risks and further design new products and advertising strategies.

Opinion summarization summarizes opinions about any product by means of telling sentiment polarities, degree and the correlated activities. With opinion summarization, a client can easily see how the present customers feel about a product, and the product producer can collect the region why some people love it or what some complain about that product. The issue of relevant sentence selection is mentioned, after which topic and opinion facts are summarized. Opinion summarizations are visualized by means of representative sentences.

Other applications include online message sentiment filtering-mail sentiment class, net weblog author's attitude evaluation and so on.

V. CONCLUSION AND FUTURE WORK

Sentiment detection has a huge form of applications in information systems, which includes classifying reviews, summarizing evaluation and other real-time programs. There are possible to be many different applications that are not discussed. It is determined that sentiment classifiers are seriously dependent on domains or subjects. From this survey, it is obvious that neither classification model consistently outperforms the different, features is one of the types that has distinct distributions. it's also found that exclusive varieties of features and classification algorithms are mixed in a green way so as to conquer their man or woman drawbacks and advantage from each other deserves, and finally enhances the sentiment classification overall performance.

Future research will be devoted to these challenges present in the area of sentiment classification and sentiment analysis. In future, more work is needed on, in addition, improving the accuracy of sentiment classification using the fusion of machine learning algorithms and semantic orientations. Sentiment analysis can be applied for new applications. Even though the techniques and algorithms used for sentiment analysis are advancing fast, however, a variety of troubles in this area of taking a look at remain unsolved. The principle tough elements exist in use of other languages, dealing with negation expressions; produce a summary of reviews based totally on product functions/attributes, the complexity of sentence/ record, managing of implicit product feature, and so on.

REFERENCES

[1] G. Vinodhini and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 6, pp. 282–292, 2012.

[2] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis.," *2014 Seventh Int. Conf. Contemp. Comput.*, no. September, pp. 437–442, 2014.

[3] S. Shah, K. Kumar, and R. K. Saravanaguru, "Sentimental analysis of twitter data using classifier algorithms," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 1, pp. 357–366, 2016.

[4] B. R. Jadhav and P. M. Mahajan, "Dual Sentiment Analysis Using Adaboost Algorithm Sentiment Analysis," vol. 6, no. 6, pp. 7641–7645, 2016.

[5] P. K. Manna and S. Bodkhe, "pooja 1 IJACEN," no. 12, pp. 58–61, 2015.

[6] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, "Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis," *Inf. Process. Manag.*, vol. 52, no. 1, pp. 36–45, 2016.

[7] A. Asmi and T. Ishaya, "Negation Identification and Calculation in Sentiment Analysis," *IMMM 2012, Second Int. Conf. ...*, no. c, pp. 1–7, 2012.

[8] S. Sharma, D. S. Kumar Yadav, and M. L. Pal, "Opinion Mining Method for Sentiment Analysis," *IOSR J. Comput. Eng.*, vol. 18, no. 5, pp. 54–60, 2016.

[9] D. Bespalov, B. Bai, Y. Qi, A. Shokoufandeh, and Y. Qi, "Sentiment Classification Based on Supervised Latent N-gram Analysis," *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, pp. 375–382, 2011.

[10] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014.

[11] W. Medhat, A. A. . Hassan, and H. H. . Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[12] A. Inamdar, A. Bhansali, S. A. Khan, and H. Mahajan, "Sentiment Analysis : Classification and Searching Techniques," pp. 2796–2798, 2016.

[13] A. Gupte, S. Joshi, P. Gadgul, and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," *...) Int. J. ...*, vol. 5, no. 5, pp. 6261–6264, 2014.

[14] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R."Sentiment Analysis of Twiyyer data" Proceeding of the workshop on Language in Social Media (LSM 2011), pages 30-38, portland, oregon, 2011.

[15] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. "Target-dependent twitter sentiment" Peoceeding of the 49 Annual Meeting of Association for computational Lingustics, pages 151-160, 2011.

[16] Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462, Mar. 2011.

[17] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," IEEE Trans. Knowl.DataEng., vol. 24, no. 6, pp. 1134 –1145, Jun. 2012.

[18] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach toopinion mining," in Proc. Int. Conf. Web Search Data Mining, 2008,pp. 231–240.Comput. Linguistics, vol. 35, no. 3, pp. 399–433, 2009.