Analysis Implementation and Evaluation of Classification Approach for Customer Sentiments using Machine Learning

Dr.ECCLESTON

Asst. Professor, Dept. of Computer Science & Engineering, Chouksey Engineering College Bilaspur (CG), India

Dr.JULIE

Master of Technology Scholar, Dept. of Computer Science & Engineering, Chouksey Engineering College Bilaspur(CG), Inida

ABSTRACT

Sentiments are the attitude, opinions, thoughts, beliefs or feelings of the writer towards something, such as people, artifacts, company or location. Sentiment analysis intends to conclude the judgment of a presenter or an author apropos to some subject matter or on the whole relative polarity of the manuscript. These constraints find out existing relationship between them. It is possible to combine all textual features obtained from inner and outer media with a machine learning algorithm for sentiment analysis. In this research work, various machine learning algorithms like naïve bayes, SVM, KNN are implemented for the sentiment analysis. The hybrid algorithm is proposed in this research work for the sentiment analysis. The proposed algorithm give high performance in terms of accuracy, precision and recall as compared to existing algorithms.

Keywords: Sentiment Analysis, SVM,KNN, Naïve Bayes.

1. INTRODUCTION

Sentiments are the attitude, opinions, thoughts, beliefs or feelings of the writer towards something, such as people, artifacts, company or location. Sentiment analysis intends to conclude the judgment of a presenter or an author apropos to some subject matter or on the whole relative polarity of the manuscript [1]. The outlook could be the perception or assessment, emotional condition, or the projected poignant message of the person behind. Opinions are decisive influencer of our behavior. Our views and insights of veracity are conditioned on how others perceive the world. The rudimentary job in opinion mining deals with deducing the inclusive polarity of the document on some specific subject matter. Sentiment analysis is a 'suitcase' field of research that contains numerous diverse disciplines, not just associated to computer science but also to social disciplines, such as psychology, philosophy, and ethics [2].Mining of opinions is an artistry of trailing the frame of mind of the community regarding a certain creation or matter from a massive set of judgments or reviews openly obtainable in web. Opinion mining is useful as when we require making decision, we habitually hunt out for other opinions. For example: we could buy a camera or a mobile phone only after checking reviews or comments or by taking opinions of others. Opinion mining is a progression for examining people's frame of mind regarding some merchandise, event or issue [3]. It may also be termed as a sub-branch of computational linguistics. This approach emphasizes on extricating public opinions out of web. A lot of steps are included in the whole process. These steps include online cleaning of text, removing white spaces, amplifying acronym, stemming, stop word elimination, refusal managing and lastly feature selection [4].Later, opinions are classified as positive, negative and neutral using classification approaches.

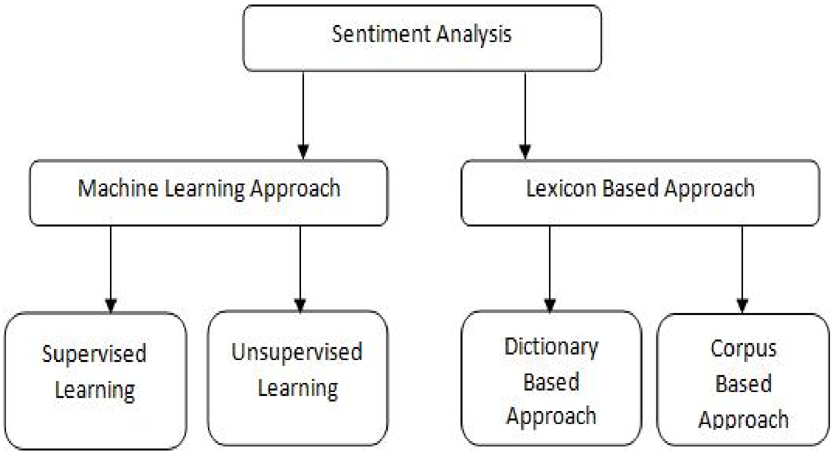

Figure 1 Techniques of Sentiment Analysis

Machine learning is completely based on machine learning approaches. These approaches provide solution of sentence level classification issue. Also, these approaches make the decree of syntactic features. Machine learning approaches are of two types namely supervised learning and unsupervised learning. Machine learning is expected to allow machines to adjust their interior configuration in such a way that they can predict the upcoming performance boost [4].

1. **Supervised Learning:** Supervised learning considers classification issues. The general purpose is to obtain the workstation to discover a classification scheme that we have formed. Digit recognition, once again, is a well-known example of classification learning. More widely, classification learning is appropriate for any problem where classification learning is valuable and classification detection is easy. In some cases, it might not be compulsory to give programmed classifications to every occurrence of a problem if the method can itself perform classification.
2. **Unsupervised Learning:**Without referring any labeled results, the patterns of any dataset are assumed through these types of algorithms [4]. In contrast to supervised machine learning, it is not possible to apply unsupervised machine learning techniques to a regression or a classification problem. This makes the training of algorithm complicated in normal way. In its place, unsupervised learning can be utilized for discovering the underlying data structure. Unsupervised machine learning is used to reveal earlier unknown data patterns.

2. LITERATURE REVIEW

Saumya Chaturvedi, et al**.** (2017) [5] presented a survey of sentiment analysis related research and its application in the business applications. The decision points of any domain could be impacted by sentiment analysis. Also, there is no general solution found highly appropriate for sentiment analysis even though several machine learning approaches were applied in the business domains. One specific kind of approach was applied by people in various organizations and then it was used repeatedly. The business prediction could follow no particular approach. Thus, for various business domains, it was important to study more about the generalized standard machine learning approaches.

N SrivatsAthindran, et al**.** (2018) [6] used the methodology of sentiment analysis to infer singular client recognitions about the various features of the new arrivals of two driving Smartphone brands in India- Vivo and Oppo. This work made use of a hybrid classifier for classifying all tweets. This classifier combined Lexicon Based Sentiment analysis and the Naive Bayes algorithms for improving accuracy level. At that point, the general sentiments for both cell phones are analyzed which gave an elevated perspective on the client observations about the two mobiles. Finally, in order to look at the client sentiments of the individual features in each of the two cell phones by doing more work. This work acted as incredible feedback scheme for organizations, which could be used for making prompt remedies or for improving the plan of their ensuing models.

Pankaj, et al. (2019) [7] presented a client input surveys on item, which included opinion mining, text mining and sentiments. All these factors influenced the encircled world by changing their sentiment on a particular item. Information utilized in this examination were online item audits gathered from http://Amazon.com. A relative sentiment analysis of recovered audits was carried out in this work. This examination work furnished companies with wistful investigation of different advanced smartphone opinions by classifying them into Positive, Negative and Neutral Behavior.

Ashwin Perti, et al. (2020) [8] stated that the vast majority of the work done in the region of Sentiment Analysis and opining mining had been carried out utilizing different kinds of Machine Learning Algorithms. This work was one step forward in the direction of Sentiment Analysis as it was focused on Twitter Sentiment Analysis. The numerous difficulties looked by Twitter Sentiment Analysis were considered as the objective and attempted to discover the solution. The client user had consistently assumed a significant function in recognizing the client sentiments and the sentiment of the client whether it was related to online voting or the client item survey. The work was performed by considering the Sentiments of an individual and afterward socio-Economic data was used for mapping.

R. Meena, et al. (2019) [9] revealed that individuals were more depended via online media for their health-related inquiries and the twitter analysis showed that there was a huge raise in the level of positive sentiments in the tweets shared by the associations and people on cancer. It was possible to use sentiment analysis and text-based mining as an extreme instrument for finding the client observation and general health intercession. The lexical and statistical analysis could be used as a monitoring instrument for analyzing medical care data. Explicit strategies could be applied to the information to acquire ease wanted outcomes.

Vasundhara, et al. (2020) [10] studied that data mining was the most well-known domain of computer science for data analysis. Different methodologies had been explained and furthermore talked about along with the various standards and procedures that were utilized for analyzing sentiments with the help of Weka software. This dataset was utilized to enhance data classification by analyzing performance of different classification algorithms (Naïve Bayes, Support Vector Machine, Decision tree and so on). The main aim here was to obtain optimal and most accurate outcomes. It was concluded that SVM (Support Vector Machine) outperformed the all other classification algorithms.

Minu Choudhary, et al. (2018) [11] analyzed that more than 5000 surveys, of various mobile brands had been retrieved from one of the basic online media stage "twitter" and afterward sentiment analysis had been performed with the help of Lexicon based methodology. The aftereffect of sentiment analysis had been introduced in the graph form. which can be used by the customers in dynamic while buying another cell phone. This can be additionally used by merchants to improve their business. The users could use these results in decision making while buying a new mobile phone. The merchants could also use these results to expand their business.

Anurag P. Jain, et al. (2015) [12] introduced a new scheme that made use of data mining classifier models to analyze the customers' sentiments. In this work, the efficiency level of single classifiers had been compared over ensemble classifiers for analyzing sentiments. The tested outcomes revealed that that k-nearest neighbour classification model generated highly accurate predictive results. It was also depicted in the results that single classifier models performed better than ensemble classifier models.

Reeta Rani, et al. (2018) [13] introduced a new ontology-based approach for generating summarized input data. The implementation of new and earlier methodologies had been carried out in python tool while results were analyzed with respect to accuracy.In the obtained results, the introduced approach showed higher accuracy in comparison to SVM classifier model. This work also presented a new approach that was capable of generating summary of the text data. This work applied weight based algorithmic approach for generating the summary.

3. PROPOSED METHODOLOGY

In this work, sentiment analysis is performed on twitter data. The important steps followed in the novel methodology are mentioned below:

1. Extraction of Microblogs data and its pre-processing: Different clients post information in different forms in the form of tweets to express their sentiments on variety of topics. The pessimistic and affirmative are the two categorizations among which the Twitter data sample is applied. Tweeter data is generally collected using Twitter API. Twitter API stands for Application Programming Interface. Twitter API facilitates software developers to access and interrelate with openly available Twitter data. In order to interact with this API, Developers may write their own scripts or may use one of the public libraries accessible in various programming languages.

2. Pre-processing: After capturing tweets required for sentiment analysis, the next step is to prepare the data. The data on social media exist in raw form. It implies that this data is noisy, rough and required cleaning. This is a vital step as the quality of the data will bring about more consistent outcomes. There are several tasks involved in preprocessing a Twitter dataset. For example, eliminating all sorts of inappropriate information such as emojis, special characters, and additional blank spaces. It may also perform more tasks such as improving format; deleting duplicate tweets, or tweets smaller than three characters.

3. Feature Extraction: There are several properties included in the preprocessed data sample. The features of developed data sample are extracted using the characteristic extraction method. Further, in a phrase, the optimistic and pessimistic polarity is calculated such that the individuals using replicas can be formatted. To perform dispensation, there are few machine learning methods that require representation of key features of contents. The characteristic vectors used for performing categorization are used for measuring input characteristics. This work makes use of N-grams for feature extraction. A brief description of this approach is provided below:

4. Training: For providing solutions to categorization issues, managed learning is known to be an important technique. To perform prospect forecasting of unidentified information, it is easier to perform training of classifier. To extract the dataset features, KNN classifier method is applied. To define the centroid points, k-mean approach is applied by KNN classifier. From these points, the Euclidian distance is calculated. In one class, the similar points are categorized. K-Nearest Neighbour is a very machine learning algorithm. This algorithm depends on supervised learning approach. This approach makes assumption about the similarity amid the novel case/data and accessible cases. This approach place the novel case into the category most analogous to the existing categories. This algorithm stores all the existing data and performs the classification of a new data point on the basis of similarity. It implies that new data can be effortlessly classified into an appropriate category with the help of this approach. This algorithm can be used for both Regression as well as Classification. However, it is mainly employed for the classification issues. It is a non-parametric algorithm. It means that this algorithm does not assume any underlying data. It is also known as a lazy learner algorithm. This algorithm does not learn from the training set instantaneously rather than it stores the dataset. During classification, this algorithm works on the dataset. At the training stage, this approach merely stores the dataset. After getting novel data, this algorithm performs the classification of this data into a category much analogous to the novel data.

Below are the steps of K-NN algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category to which the number of the neighbor is maximum.

5. Classification: In this research, the Random Forest classifier is used since dual categorization exists in the emption study and for execution, several data samples are available. To train the classifier, the manually generated training set is applied. This method uses a physically generated training sample. Within the training sample, the X:Y relation is implied where, the score of an estimation test is denoted by x and optimistic or pessimistic words are represented by y. It is a supervised classification and regression algorithm. As per the name, this algorithm generates a forest with numerous trees in random manner.

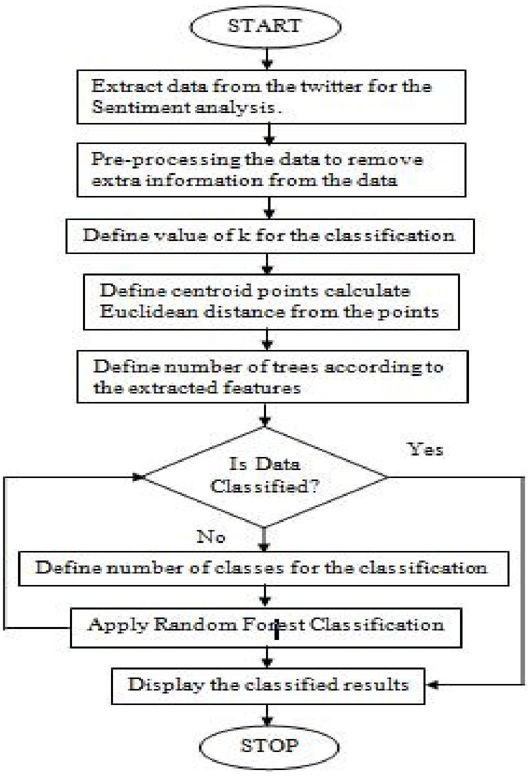
START
Extract data from the twitter for the Sentiment analysis.
Pre-processing the data to remove extra information from the data
Define value of k for the classification
Define centroid points calculate Euclidean distance from the points
Define number of trees according to the extracted features
Is Data Classified?
Yes
No
Define number of classes for the classification
Apply Random Forest Classification
Display the classified results
STOP

Figure 1 Proposed Work

4. RESULT AND DISCUSSION

To predict the accuracy of machine-learning based algorithms there are several classifier available in previous work. Some important measures are:

1. Precision,
2. Recall
3. F-measure
4. Accuracy

1. Precision: The term precision represents two or more closely related values of the measurements. The value of precision varies due to the observational error. The steadiness or reproducibility of the measurement can be identified using this parameter. Equation 3 denotes this parameter as:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (1)$$

2. Recall: Recall can be described as the ratio of accurately predicted positive observations to the all observations in actual class.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (2)$$

3. F1 score- F1 Score can be described the weighted average of Precision and Recall. Hence, this score consider both false positives as well as false negatives.

$$F1\ Score = \frac{2*\mathrm{Recall}*\mathrm{Precision}}{\mathrm{Recall+Precision}} \quad (3)$$

4. Accuracy- Accuracy is perhaps the most primal performance parameter. It can be described as a ratio of suitably predicted observation to the total observations.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (4)$$

Where,

TP= True Positive; TN= True Negative; FP= False Positive; FN= False Negative

To predict the accuracy of machine-learning based algorithms there are several classifier available in previous work.

Precision- Precision is a part of applicable extracted examples. In case of class, the precision is the ratio of number of accurate results (i.e., true positives) and number of all returned results (i.e., the total of true positives and false positives) in classification.

Table 1: Precision Analysis

Classifier	Precision
Naïve Bayes	0.82
Logistic Regression	0.81
SVM	0.35
Basepaper(SVM, LR, NB, RF)	0.84
Proposed	0.87

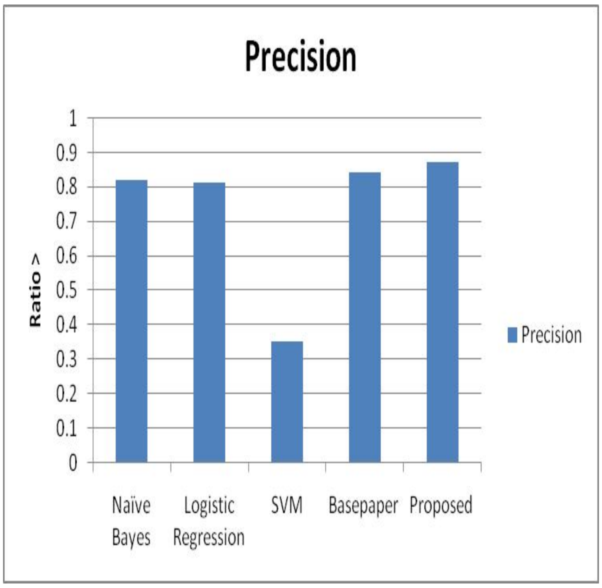

Figure 2. Precision Analysis.

As illustrated in figure 2, the precision value of the existing algorithms likes naïve bayes, logistic regression, SVM, random forest are compared with the proposed model. The precision value of the proposed model is high as compared to other classifiers

Recall- Recall is the part of applicable extracted examples. The recall in this context is the opposite measure. It is described as the ratio of number of accurate outcomes to the number of outcomes that should have been retrieved.

Table 2: Recall Analysis

Classifier	Recall
Naïve Bayes	0.77
Logistic Regression	0.72
SVM	0.59
Base paper (SVM, LR, NB, RF)	0.79
Proposed	0.85

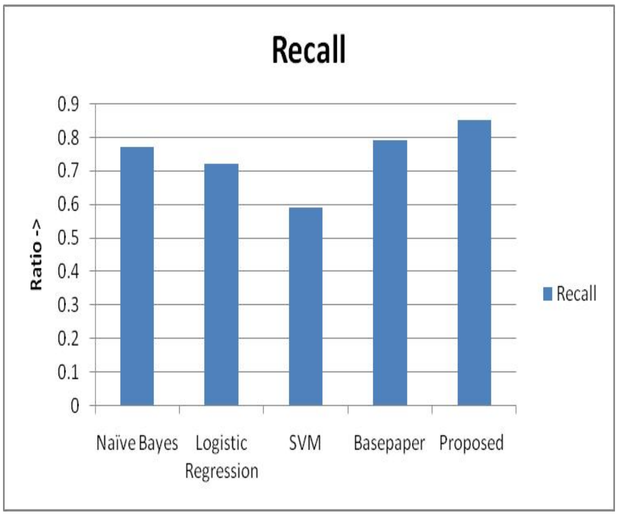

Figure 3. Recall Analysis.

As shown in Figure 3, the recall value of the existing algorithms likes naïve bayes, logistic regression, SVM, random forest are compared with the proposed model. The recall value of the proposed model is high as compared to other classifiers

Accuracy- The capability of a specified classifier to rightly predict the class label of novel or earlier hidden data is known as classification accuracy.

Table 3: Accuracy Analysis

Classifier	**Accuracy**
Naïve Bayes	77.48
Logistic Regression	72.2
SVM	59.01
Base paper (SVM, LR, NB, RF)	79.24
Proposed	84.87

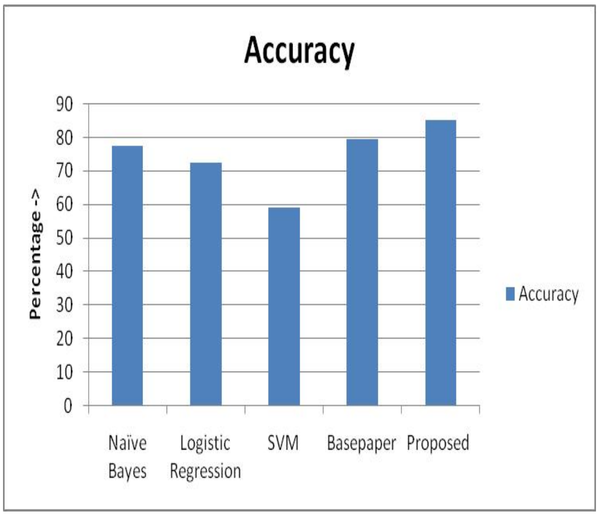

Figure 4. Accuracy Analysis.

As shown in figure 4, the recall value of the existing algorithms likes naïve bayes, logistic regression, SVM, random forest are compared with the proposed model. The accuracy value of the proposed model is high as compared to other classifiers.

5. CONCLUSION

The sentiment analysis methods which are proposed so far have various steps. In the pre-processing stage, the missing and redundant values are removed from the dataset. The feature extraction method established relationship between attribute and target set. In the last step of classification, the classification method is enforced which can categorize data into certain classes like positive, negative and neutral. In the previous method, the hybrid classification method is applied to evaluate the sentiments of the twitter data, but still there's some room to improve accuracy and precision. In this study, a hybrid classification method is designed which is the mixture of KNN and random forest classifier for the sentiment analysis. The various classifiers like naïve bayes, logistic regression, SVM, random forest and proposed model are evaluated in terms of precision, recall and accuracy. It is examined that outcomes for the sentiment analysis of the proposed model is optimized up to 3 to 5 percent approximately.

REFERENCES

1. Hu and Liu, "Mining and Summarizing Customer Reviews," in International Conference on Knowledge Discovery and Data Mining, Seattle, USA, pp. 168-177, 2004.
2. .A. M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Conference on Human Language Technology and Empirical Methods in Natural Language Processing, British Columbia, pp. 339-346, 2005
3. Liu B., "Opinion Mining and Summarization," World Wide Web Conference Beijing, China, 2008, Downloaded from: https://www.cs.uic.edu/~liub/FBS/opinion-mining-sentiment-analysis.pdf [21st June 2016]
4. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, USA, pp. 417-424, 2008.
5. Saumya Chaturvedi, Vimal Mishra, Nitin Mishra, "Sentiment analysis using machine learning for business intelligence", 2017, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)
6. N SrivatsAthindran, S. Manikandaraj, R. Kamaleshwar, "Comparative Analysis of Customer Sentiments on Competing Brands using Hybrid Model Approach", 2018, 3rd International Conference on Inventive Computation Technologies (ICICT)
7. Pankaj, Prashant Pandey, Muskan, NitashaSoni, "Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews", 2019, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)
8. Ashwin Perti, Munesh Chandra Trivedi, Amit Sinha, "Development of intelligent model for twitter sentiment analysis", 2020, Materials Today: Proceedings, In press, corrected proof, Available online
9. R. Meena, V. ThulasiBai , "Study on Machine learning based Social Media and Sentiment analysis for medical data applications", 2019, Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)

10. Vasundhara, Suraiya Parveen, “Towards Sentiment Analysis: A powerful technique for Data Analytics”, 2020, International Conference on Electronics and Sustainable Communication Systems (ICESC)
11. Minu Choudhary, Prashant Kumar Choudhary, “Sentiment Analysis of Text Reviewing Algorithm using Data Mining”, 2018, International Conference on Smart Systems and Inventive Technology (ICSSIT)
12. Anurag P. Jain, Vijay D. Katkar, “Sentiments analysis of Twitter data using data mining”, 2015, International Conference on Information Processing (ICIP)
13. Reeta Rani, Sawal Tandon, “Chat Summarization and Sentiment Analysis Techniques in Data Mining”, 2018, 4th International Conference on Computing Sciences (ICCS)