IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Special Issue 1, March 2024

1st International Conference on Machine Learning, Optimization and Data Science

Organized by

Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, India

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || Monthly Peer Reviewed & Refereed Journal |

| "1st International Conference on Machine Learning, Optimization and Data Science"|

Organized by

| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|

|| Volume 12, Special Issue 1, March 2024 ||

A Machine Learning Framework for Sentiment Analysis of Hotel Reviews

Aishwary Kumar Yadav[1], Prof. Preeti Rai[2]

Student, Gyan Ganga Institute of Technology & Science, Jabalpur, M.P., India [1]

Professor, Gyan Ganga Institute of Technology & Science, Jabalpur, M.P., India [2]

ABSTRACT: Text mining research now faces both enormous opportunities and challenges due to the explosive growth of unstructured textual data mountains and the proliferating technologies to analyze it. The challenge of automatically labeling text data comes from the fact that people frequently express their thoughts in intricate ways that can be challenging to understand. Large amounts of work go into labeling datasets, and mislabeled datasets frequently result in poor decision-making. In this study, we develop an opinion mining system for sentiment analysis applied to hotel customer comments. The majority of hotel review datasets that are currently accessible lack labeling, which poses a significant challenge for researchers in terms of text data pre-processing tasks. Furthermore, because sentiments are feelings like emotions, attitudes, and opinions that are frequently teeming with idioms, onomatopoeias, homophones, phonemes, alliterations, and acronyms, sentiment databases are frequently very domain-sensitive and difficult to construct. The suggested approach, known as sentiment polarity, automatically generates a sentiment dataset for testing and training in order to derive objective assessments of hotel services from customer feedback. To find an appropriate machine learning algorithm for the framework's classification component, a comparative analysis was conducted using Naïve Bayes multinomial, sequential minimal optimization, complement Naïve Bayes, and Composite hypercubes on iterated random projections.

KEYWORDS: opinion mining, sentiment analysis, machine learning algorithm, natural language processing, sentiment polarity, dataset labeling

I. INTRODUCTION

The amount of textual data in the globe has increased dramatically in recent years, particularly the unstructured data produced by people expressing their ideas on various websites and social media platforms for a variety of reasons. Massive amounts of textual material could be compared to garbage that needs to be disposed of periodically. But as storage capacity has increased and data mining techniques have become more sophisticated, there are now more opportunities and challenges to analyze and extract meaningful insights from this massive amounts of data.

In this research, we have selected textual data for sentiment analysis using opinion mining from consumer perspectives, specifically hotel reviews. Sentiment analysis automates the sentiment categorization process from reviews by utilizing computational linguistics and natural language processing techniques. Hotels offer comfort, safety, and security. comfort, luxury, and accommodation services for tourists and travelers. In order to better understand customer expectations and enable efficient customer relationship management, it is desirable to mine hotel reviews. The hotel managers would be able to better understand the needs of their customers, identify areas for improvement, and enhance the quality of their services. Customers who have booked reservations at a specific hotel are the only ones who can give the hotel reviews. Customer service quality, cleanliness, cuisine quality, location, and the warmth shown by hotel employees are all mentioned in reviews left by guests. Additionally, sentiment analysis of hotel reviews is essential to uncovering patterns in the data that could be used to significantly boost performance [1].

II. LITERATURE REVIEW

Sentiment analysis [2] and opinion mining [3] are phrases used to describe the field of research that examines people's opinions, assessments, appraisals, attitudes, and feelings regarding various entities, including persons, organizations, individuals, events, products, and themes, as well as their characteristics [4]. Positive and negative sentiments were defined by these phrases, which were used interchangeably [4, 5]. The polarity of the opinions expressed in a particular review is determined by sentiment. By categorizing sentiment analysis as concentrating on emotion recognition and

opinion mining as polarity detection, Camilla et al. [6] dispelled the confusion surrounding the interchange of these ideas. It is sufficient for the opinion mining system to comprehend polarity, which can be positive, negative, or neutral based on the type of language expressed in a review [7]. Analyzing attitudes on a given topic is closely related to the process of polarity detection.

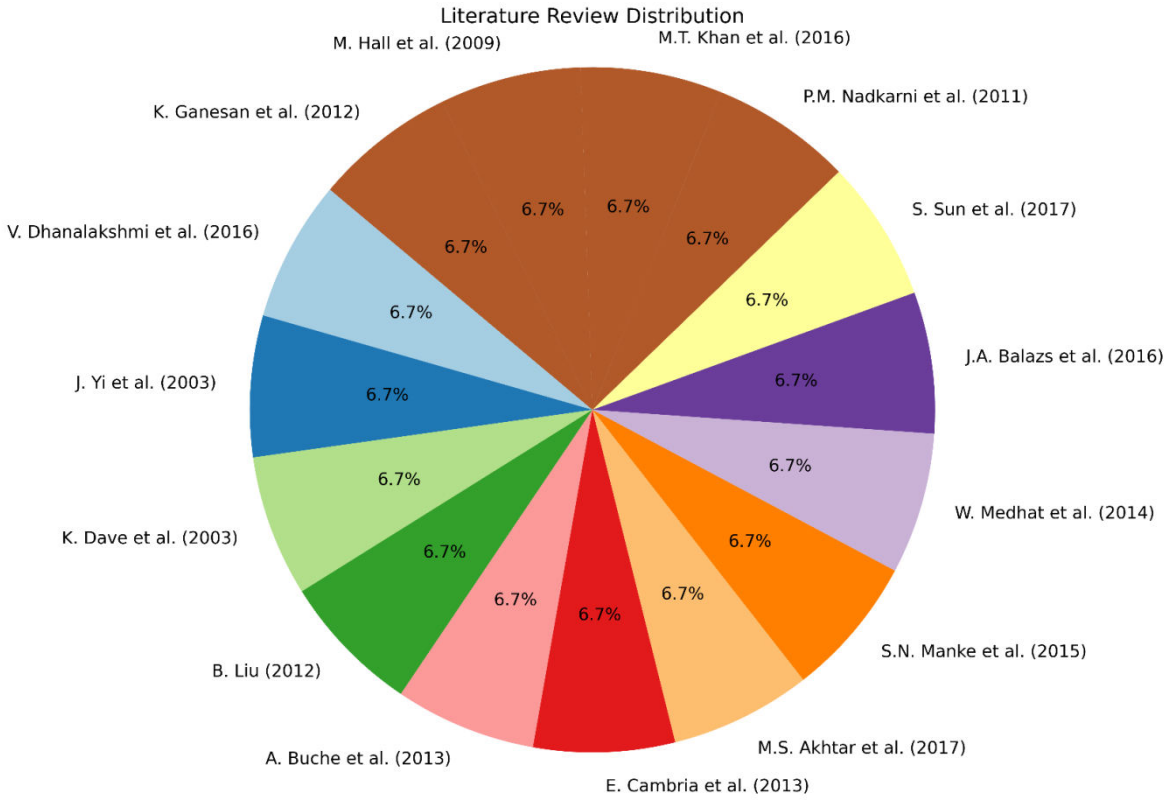
The majority of sentiment analysis studies concentrate on descriptive data. Manke and Shivale [8] emphasized the value of social networks as the favored settings for sentiment analysis and opinion mining. They presented the first opinion categorization technique and evaluated their algorithm using real social network data sets. Based on their research, they came to the conclusion that social networks have characteristics that make them appropriate for opinion mining. Comprehensive surveys on opinion mining techniques have been given [9–11], with a scant emphasis on aspect-oriented analysis.

Most sentiment analysis techniques available today aim to identify a review's polarity in relation to the organizations like lodgings and facilities with their own features, like meals and internet access, for example. In contrast, the aim of this study is to identify the features of target entities and the sentiment expressed towards each aspect by aspect-based sentiment analysis. The sentiment analysis summarizes the positive and negative aspects of customer reviews of goods and services. This task has always been challenging [7] because it requires a number of subtasks, including feature extraction, feature grouping, polarity classification, and evaluation methods, to obtain an objective opinion. These tasks are typically completed under the unrealistic assumption that there are no grammar errors.

The processing of textual data is the focus of the discipline of natural language processing (NLP), which includes hotel reviews as a major theme [12]. The point where artificial intelligence and linguistics converge [6] is what makes natural language processing (NLP) approaches suitable for sentiment analysis. Two fundamental methods for sentiment analysis are machine learning and lexicon-based approaches [9, 10]. The machine learning methodology that this study is based on makes use of both supervised and unsupervised learning approaches. Sentiment analysis has made use of supervised learning algorithms including convolutional neural network deep learning, Naïve Bayes, K-nearest neighbor, and support vector machine.Labeled data sets must be used to train machines using supervised learning algorithms. On the other hand, because unsupervised learning algorithms like K-means and fuzzy C-means clustering learn by observation, they do not need training datasets.

Supervised learning algorithms, like those used in this study, can be applied to large amounts of labelled training data on hotel evaluations to provide insightful information that could assist hotels outperform rivals in terms of performance and overall ratings. The accuracy of various supervised learning algorithms is assessed using benchmark metrics, which include true positive rate, false positive rate, prediction, recall, and F-measure rate, receiver operating characteristicistic (ROC) area, and precision recall curve (PRC) area. But in order to obtain reliable findings from these measurements, a framework that supports the automatic creation of appropriately labeled datasets with real customer sentiment expressed must be designed.

The usage of tools for carrying out various NLP tasks is crucial to the success of sentiment analysis and opinion mining. Red Opal, a tool that helps consumers identify products based on attributes, is one that can be utilized for NLP activities. Tools like Sentic Net, Luminoso, Factiva, Attensity, and Convers eon are used by businesses to extract and analyze client comments about their products from blogs. Commonly used NLP toolkits for implementing fundamental NLP tasks including POS tagging, natural entity recognition, and parsing are NLTK, OpenNLP, and Stanford Core NLP [11]. A popular tool with several applications for data analysis and predictive modeling is the W EKA system. It facilitates a number of common data mining activities, such as data pre-processing, clustering, regression, feature selection, association rule mining, and visualization [13]. An external source can provide data in forms like comma-separated values (CSV) files for importation. Since the raw data might not be compatible with the necessary processing, it must be cleaned. Preprocessing is the process of converting unprocessed data into a format that can be worked with by the right tools. Additionally, the performance of the learning algorithms is intrinsically quantified in terms of the standard evaluation metrics, and the more accurate the data, the more accurate the algorithms work.



III. METHODOLOGY

A. The Intuition Model

The feedback collection is where the intuition model's conceptual view opens in Figure 1. Clients answer surveys regarding their experiences with the services they obtained from the chosen hotels. There are several ways to accomplish this, such as setting up a web portal where clients can leave feedback. Labeling the comments using intuition will be the next step in the process. Human agents will carry out this task; they will presumably read the comments and apply labels based on their impressions. The next step will be to utilize filters to turn the labelled text into feature vectors after the data has been turned into the desired format. This will simplify the implementation of a classification algorithm for data testing and training. The following stage is to choose a suitable classification algorithm, and the final step is to train and test the chosen algorithm on a dataset and record the findings.

B. The Model Based on Sentiment Polarity

The sentimental property based model (SPBM), as illustrated in Figure 2, was used for the research reported in this paper. The SPBM mode l starts with the elic itation of opinions, which is the step omitted because we employed the raw OpinRank dataset [14, 15], just like the intuition based model (IBM). Using the proper user interface, customers answer questions on the services they received from the chosen hotels. The next step will be to use a sentiment polarity algorithm to label the comments according to their sentiment polarity score. Whether a comment is positive, negative, or neutral will depend on the score that was obtained. After the data have been processed, the ne xt s tep uses filters to turn the labelled te xt into feature vectors. Choosing an appropriate classification algorithm is the next step. The chosen classification method is trained, tested, and the results are recorded in the final phase. One of SPBM's unique selling points is its automatic labeling system, which labels sentiments without the need for human participation. However, IBM mainly relies on human participation to categorize attitudes, which can occasionally be inconsistent and requires a time-consuming and arduous process.

C. Mathematical Formulation

Problem Definition

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ be the set of hotel reviews, where d_i represents the i-th review. The goal is to assign a sentiment

label $y_i \in$ { positive, negative, neutral } to each review d_i.

1 Data Preprocessing
The preprocessing step transforms the raw text data into a structured format suitable for analysis.
Tokenization:
For each review d_i, we define a set of tokens (words) $T_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$, where t_{ij} represents the j-th token in review d_i.

$$T_i = \text{Tokenize}(d_i)$$

Stop Words Removal:
Let S be the set of stop words. We remove all tokens $t_{ij} \in S$ from T_i :

$$T_i' = T_i \setminus S$$

Stemming/Lemmatization:
We apply a stemming or lemmatization function ϕ to each token $t_{ij} \in T_i'$:

$$T_i'' = \{\phi(t_{ij}) \mid t_{ij} \in T_i'\}$$

Feature Extraction
We convert the processed tokens into numerical features.
Bag of Words (BoW):

Define a vocabulary $V = \{v_1, v_2, \ldots, v_k\}$ as the set of unique tokens across all reviews. The BoW vector $\mathbf{x}_i$ for review d_i is a k-dimensional vector where each element x_{ij} represents the frequency of token v_j in T_i'' :

$$x_{ij} = \text{count}(v_j, T_i'')$$

Term Frequency-Inverse Document Frequency (TF-IDF):
Let $\text{tf}(v_j, d_i)$ be the term frequency of v_j in d_i and $\text{df}(v_j)$ be the document frequency of v_j in $\mathcal{D}$. The TF-IDF weight w_{ij} is given by:

$$w_{ij} = \text{tf}(v_j, d_i) \cdot \log\left(\frac{n}{\text{df}(v_j)}\right)$$

The TF-IDF vector for review d_i is $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ik})$.

Sentiment Classification
We employ a machine learning classifier $f: \mathbb{R}^k \rightarrow$ { positive, negative, neutral } to assign a sentiment label to each review. The classifier is trained on a labeled dataset $\mathcal{L} = \{(\mathbf{x}_i, y_i) \mid i = 1,2, \ldots, n\}$.

Logistic Regression:
The probability of a review d_i being classified as positive, negative, or neutral is modeled using a logistic function: where $\mathbf{w}_c$ is the weight vector for class c.

Support Vector Machine (SVM):
An SVM classifier finds the optimal hyperplane that separates the classes in the feature space:

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

where C is the regularization parameter.

IV. EXPERIM ENTAL RESULTS

Information from the Opin Rank opinion-based ranking dataset was obtained in order to experimentally assess the effectiveness of the studied learning algorithms in order to identify an appropriate algorithm for the classification part of the SPBM framework. The OpinRan k dataset was chosen due to its unlabeled reviews, which provided the opportunity for customized experimentation. About 259000 unlabeled reviews of automobiles and lodging from 80 to 100 hotels spread across ten different cities worldwide are included in the Opin Rank dataset. Dubai, Beijing, London, New York City, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, and Chicago are some of these cities. To create a sub-dataset for our experiment, we arbitrarily chose hotel reviews from Beijing, Montreal, and London. Using a Python sentiment analysis tool called TextBlob, a library for handling textual data, we labeled the hotel attributes in the dataset. The data was subsequently automatically labeled using the sentiment polarity scores. The dataset was divided into two subsets to establish training and testing datasets after all required processing steps, such as labeling and filtering, were completed. We trained and tested the dataset using four classification algorithms: complement naïve bayes (CNB), sequential min ima l optimization (SMO), hypercubes on iterated random projections (CHIRP), and naïve bayes mu ltino mia l (NBM). The comparative findings of the experiment are displayed in Figure 3. With a precision of 80.9%, the Naïve Bayes multinomial algorithm had the highest accuracy. The complement Naïve Bayes method, with 80.5%, came in close second. The worst-performing algorithm, CHIRP, with a precision score of 75.6%.

Algorithm for Sentiment Analysis Precision Evaluation

Input: Unlabeled reviews R from the Opin Rank dataset, where $R = \{r_1, r_2, \ldots, r_n\}$.
Output: Precision scores P for each classification algorithm $A = \{A_1, A_2, A_3, A_4\}$.

Step 1: Data Preparation

1. Select a subset of reviews $R' \subset R$ from cities {Beijing, Montreal, London}.
2. Use TextBlob to label sentiment polarity for each review $r_i \in R'$.
3. Split R' into training set T and testing set S.

Step 2: Classification Algorithms

1. Define classification algorithms:

- A_1 : Complement Naive Bayes (CNB)
- A_2 : Sequential Minimal Optimization (SMO)
- A_3 : Hypercubes on Iterated Random Projections (CHIRP)
- A_4 : Naive Bayes Multinomial (NBM)

Step 3: Training and Testing

1. For each algorithm $A_i \in A$:

- Train A_i on the training set T.
- Test A_i on the testing set S.
- Compute precision p_i for A_i.

Step 4: Precision Calculation

1. Calculate precision p_i as:

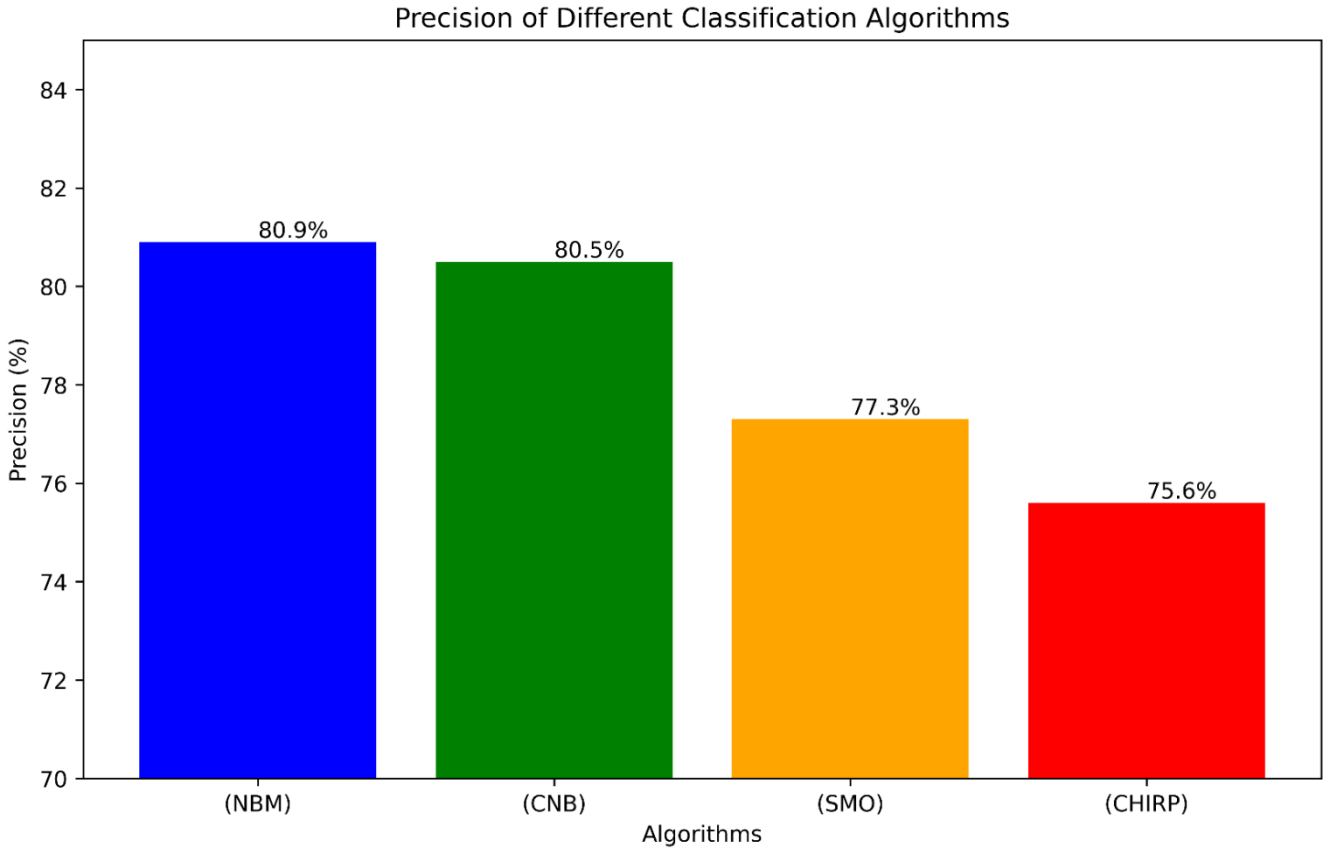
$$p_i = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Step 5: Comparative Analysis

1. Compare the precision scores $P = \{p_1, p_2, p_3, p_4\}$:

- p_4(NBM): 80.9%
- p_1(CNB): 80.5%
- p_2 (SMO): [Precision value]
- p_3 (CHIRP): 75.6%

Algorithm:
Input: R, { Beijing, Montreal, London}, TextBlob, T, S Output: P1. Data Preparation1
This algorithm captures the essence of the data processing and classification procedure described in the provided text. Each step is defined mathematically, with clear input, output, and computational steps for training and evaluating sentiment analysis algorithms.



Comparative Precision of Sentiment Analysis Algorithms for Hotel Reviews

The bar chart titled "Comparative Precision of Sentiment Analysis Algorithms for Hotel Reviews" illustrates the precision scores of four different classification algorithms—Naïve Bayes Multinomial, Complement Naïve Bayes, Sequential Minimal Optimization, and Hypercubes on Iterated Random Projections—demonstrating their effectiveness in analyzing sentiment in hotel review data.

V. CONCLUSION

The study presents an opinion mining framework for sentiment analysis that might be integrated into a hotel technology system to enhance customer relations and management. If a system that predicts sentiment polarity operates on incorrectly labeled data, what useful is it? Based on our sentiment polarity exercise, we discovered that certain comments could be mistakenly seen as neutral while in fact they are either positive or negative. The e-mail below was considered a neutral comment. "Wow, that hotel is really amazing!" Although the word "HELLT EL" is not in the English language, this caustic and legitimately unpleasant phrase was placed in the neutral class. But the majority of comments had considerably more accurate labels. We think there is still a lot of study to be done in this area, particularly in terms of optimizing the feature extraction algorithm of the frame work to minimize classification error.
Labeled datasets are typically utilized to let the system to learn automatically, and it is expected that the system will ascertain feelings in a manner similar to that of humans. In order to prevent false information from entering the system, the suggested framework attempts to ensure that phrases are appropriately labeled. To put it briefly, the structure

suggested in this research aids in automated sentiment dataset labeling. When compared to other classification algorithms, the experimental results of this study showed that the Naïve Bayes multinomial algorithm performed well in terms of the evaluation metrics used. Our goals for future work include employing deep learning algorithms to classify client replies based on emotions, enhance automatic labeling, and feature extraction.

REFERENCES

1. V. Dhanalakshmi, B. Dhivya and A.M. Saravanan, " Opinion mining from st udent feedback data using supervised learning algorithms". IEEE 3rd MEC International Conference on Big Data and Smart City., pp. 1-5, 2016.
2. J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, " Sentiment analyzer: Extract ing sent iment s about a given topic using natural language processing techniques". In Data Mining, 2003. ICDM 2003. Third IEEE International Conference, pp. 427-434, 2003.
3. K. Dave, S. Lawrence D.M. and Pennock, "Mining the peanut gallery: opinion extract ion and semant ic classificat ion of product reviews". In ACM Proceedings of the 12th international Conference on World Wide Web, pp. 519-528, 2003.
4. B. Liu, " Sentiment analysis and opinion mining". Synthesis Lectures on Hum an Language Technologies, vol. 5, no. 1, pp. 1-167, 2012.
5. Buche, D. Chandak and A. Zadgaonkar, " Opinion mining and analysis: a survey". International Journal on Natural Language Computing (IJNLC), vol. 2, no. 3, pp. 39-48, 2013.
6. E. Cambria, B. Schuller, Y. Xia and C. Havasi, "New avenues in opinion mining and sentiment analysis". IEEE Intelligent Systems, vol. 28, no. 2, pp. 15-21, 2013.
7. M.S. Akht ar, D.k Gupt a, A. Ekbal and P. Bhattacharyya. "Feat ure select ion and ensemble const ruct ion: a t wo-st ep method for aspect based sentiment analysis." Knowledge-Based Systems, vol. 125, pp. 116-135, 2017.
8. S.N. Manke and N. Shivale, " A review on: opinion mining and sentiment analysis based on natural language processing". International Journal of Computer Applications, vol. 109, no. 4, pp. 29-32, 2015.
9. W. Medhat , A. Hassan and H. Korashy, " Sentiment analysis algorithms and applicat ions: a survey". Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, 2014.
10. J.A. Balazs and J.D. Velásquez, "Opinion mining and informat ion fusion: a survey". Information Fusion, vol. 27, pp. 95-110, 2016.
11. S. Sun, C. Luo and J. Chen, " A review of natural language processing t echniques for opinion mining syst ems". Information Fusion, vol. 36, pp.10-25, 2017.
12. P.M. Nadkarni, L. Ohno-Machado and W.W. Chapman, "Nat ural language processing: an int roduct ion". Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 544-551, 2011.
13. M.T. Khan, M. Durrani, A. Ali, I. Inayat , S. Khalid and K.H. Khan," Sentiment analysis and the complex natural language". Complex Adaptive Systems Modeling, vol. 4, no. 1, pp. 1-19, 2016.
14. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reut emann and I.H. Witten, "The WEKA data mining software: an updat e". ACM SIGKDDexplorations newsletter, vol. 11, no. 1, pp. 10-18, 2009.
15. K. Ganesan and C. Zhai, " Opinion -based entity ranking". Information retrieval, vol. 15, no. 2, pp. 116-150, 2012.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || Monthly Peer Reviewed & Refereed Journal |

| "1st International Conference on Machine Learning, Optimization and Data Science"|

Organized by

| Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur|

|| Volume 12, Special Issue 1, March 2024 ||

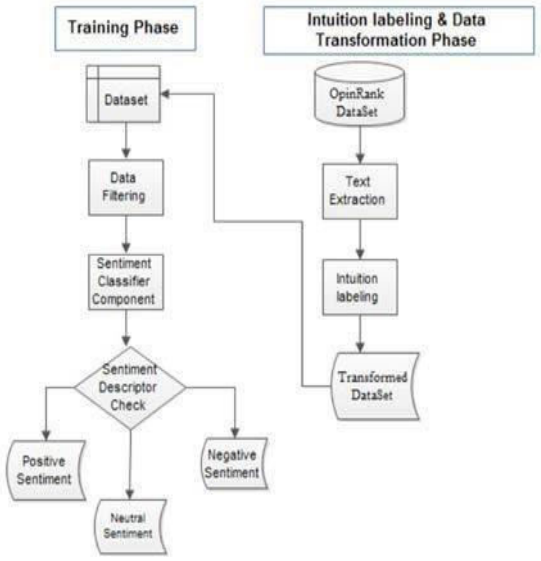


Figure 1. Intuitive sentiment analysis framework

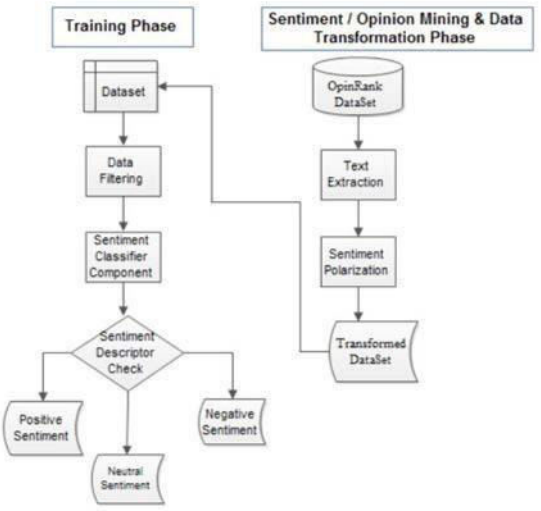


Figure 2. Sentiment polarity analysis framework

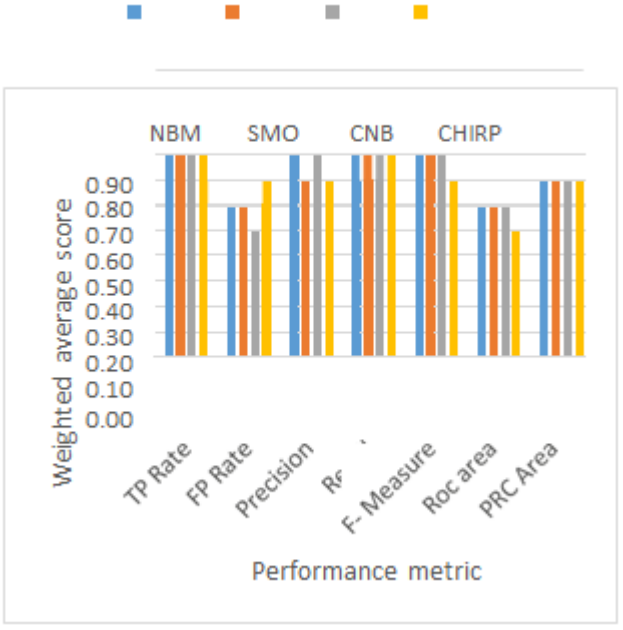


Figure 3. Weighted average score against performance metrics

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details