# Sentiment Classification for Film Reviews in Gujarati Text Using Machine Learning and Sentiment Lexicons

**Parita Shah[1],* Priya Swaminarayan[2] & Maitri Patel[3]**

[1]Department of Computer Engineering, Sarva Vidyalaya Kelavani Mandal managed Vidush Somany Institute of Technology and Research, Kadi, India
[2]Faculty of Information Technology and Computer Science, Parul University, Vadodara, India
[3]Department of Computer Engineering, Gandhinagar University, India
*E-mail: paritaponkiya@gmail.com

**Abstract.** In this paper, two techniques for sentiment classification are proposed: Gujarati Lexicon Sentiment Analysis (GLSA) and Gujarati Machine Learning Sentiment Analysis (GMLSA) for sentiment classification of Gujarati text film reviews. Five different datasets were produced to validate the machine learning-based and lexicon-based methods' accuracy. The lexicon-based approach employs a sentiment lexicon known as GujSentiWordNet, which identifies sentiments with a sentiment score for feature generation, while in the machine learning-based approach, five classifiers are used: logistic regression (LR), random forest (RF), k-nearest neighbors (KNN), support vector machine (SVM), naive Bayes (NB) with TF-IDF, and count vectorizer for feature selection. Experiments were carried out and the results obtained were compared using accuracy, precision, recall, and F-score as performance evaluation criteria. According to the test results, the machine learning-based technique improved accuracy by 3 to 10% on average when compared to the lexicon-based approach.

**Keywords**: *Gujarati text; lexicon; machine classifier; movie reviews; sentiment analysis.*

## 1 Introduction

Text categorization is a subset of opinion classification for the evaluation of people's views and attitudes toward various subjects. Usage of social media platforms has increased in recent years, which has resulted in the generation of a huge amount of data on the web. Many different online platforms are available, such as retail, entertainment and content communities, messaging, and blogging services, where users can express their opinions. This results in large data that cannot be analyzed manually, necessitating the use of a method such as sentiment analysis, which offers atomization.

Individual users' perspectives differ from one another. Therefore, it is critical to evaluate many different opinions to offer more realistic thoughts about a topic and a large number of opinions must be analyzed. The bulk of the material on the

Received December 13th, 2021, Revised July 26th, 2022, Accepted for publication September 6th, 2022.
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2023.17.1.1

internet is in English, but owing to the increased awareness of online media, information in regional languages is also increasing quickly. So far, not much attention has been paid in the literature to Indian languages, particularly languages like Gujarati [1].

According to our literature survey, lexicon-based and machine learning-based approaches have been utilized for sentiment classification. Machine learning (ML) algorithms are used to predict sentiment in ML-based methods [2]. On the other hand, sentiment lexicons are utilized in lexicon-based sentiment analysis. Lexicons are characterized as either lexicon-based or corpus-based, based on the resources used to determine sentiment polarity. Lexicon-based methods begin with seed emotion words and then proceed with synonyms and antonyms. Corpus-based methods start with sentiment seed words and then build a huge corpus of opinion words in the same context [3].

Due to a lack of resources for Gujarati, the authors propose an algorithm that relies on a machine-learning classifier as well as a sentiment lexicon. Because there was no widely available dataset of movie reviews written in Gujarati, a dataset of movie reviews in Gujarati was generated. Both suggested methods were evaluated on this dataset. It was determined that ML-based classification produced more accurate results than lexicon-based classification. The same approaches were applied to five unique datasets, comprising various real-world data, to test the accuracy of the suggested method.

## 2 Related Work

The relevant literature can be categorized into two types according to the approach applied: lexicon-based and ML-based. Supervised methods are often used in ML-based systems using TF-IDF and count vectorizer as feature generators [4]. Lexicon-based approaches need the development of a lexicon [5], which may be generated from an existing lexicon or derived from opinion words in a corpus. Using the lexicon's polarity values, the general mood of the document is predicted. Table 1 shows the lexicon technique as well as the prediction model used for sentiment analysis in the Indian language.

The point of convergence of this research was to perform sentiment analysis from review data related to movies. A previous study [6], has proposed the Senti-lexical algorithm to find the furthest point of a study as positive, negative, or neutral. The authors furthermore proposed a procedure to manage words that affect studies, while the effect of emoticons can similarly be inspected.

**Table 1** Prediction classifier and lexicon-based sentiment analysis of Indian language.

| Reference | Language | Method Used | Dataset | Accuracy |
|---|---|---|---|---|
| [7] | Hindi | Naive Bayes | Movie Review | 80% (Hindi) |
| [8] | Hindi, Tamil, Bengali | POS using SentiWordnet | SAIL2015 | 34% (Bengali)<br>56.7% (Hindi),<br>39.3% (Tamil) |
| [9] | Konkani | Konkani WordNet | Not Specified | Not Specified |
| [10] | Malayalam | Malayalam WordNet | Movie Reviews | 85% |
| [11] | Punjabi | Punjabi WordNet | Blogs and News Papers | Not Specified |
| [12] | Kannada | Kannada WordNet | General Document | Not Specified |
| [5] | Gujarati | Synset Replacement Algorithm (Guj SentoWordNet), WordNet, Bag-of words | Tweets | 52.72% |
| [13] | | Support vector Machine | Normal total 40 Tweets | 92% |
| [14] | Hindi | Machine Translation<br>Hindisentiwordnet<br>Support vector Machine | Movie Reviews | 65.96%<br>60.31%<br>78.14% |
| [15] | Bengali | NB, SVM, KNN, Decision tree, SVM, RF. | Bengali Horoscope | 98.7% (SVM) |
| [16] | Tamil | Maximum entropy, NB, SVM, decision tree. | Movie | 75.9% (SVM) |
| [17] | Punjabi | Naive bayes | Blogs and News Papers | Not specified |
| [18] | Kannada | Decision Tree (ID3) | Kannada movie reviews | 79% |
| [19] | Hindi | Naive bayes | Hindi Tweets | 56% |
| [20] | | Decision Tree | Hindi tweets (weka) | 45% |
| [21] | Malayalam | Lexicon based | Malayalam text | 87.5% |
| [22] | Hindi | Lexicon based | Movie review | 75% |
| [23] | Hindi and Bengali | Lexicon based | Hindi and Bengali tweets | 50.75% (Hindi)<br>48.82% (Bengali) |
| [24] | Kannada | Lexicon based<br>Maximum entropy<br>Naïve Bayes | Kannada documents | 78% (lexicon-based)<br>90% (NB)<br>93% (ME) |

The goal of Ref. [25] was to extract opinions from movie reviews and to group the opinions communicated at various levels using sentiment lexicons. This paper offers exploratory outcomes utilizing a technique inspired by nature (molecule swarm streamlining) for marking. This improvement strategy repeatedly marks all words in a lexicon and evaluates the accuracy of the assessment

characterization utilizing a lexicon until the ideal names for words in the lexicon are found. A helper approach using an AI strategy is incorporated into the technique presented in [26]. This method could classify over 90% of the texts and accomplished preferable outcomes compared to the lexicon-based method. A crossover model can be utilized for sentiment analysis in human-robot communications.

Opinion mining, a subdiscipline of information mining and computational etymology, refers to computational methods for extracting, characterizing, understanding, and evaluating sentiment. Each article presents an unique imaginative examination structure, computational techniques, and selected results and models. The authors in Ref. [28] explored opinion mining and sentiment classification and concentrated on three non-English dialects to study the characterization techniques and the proficiency of every algorithm utilized. It was discovered that most of the non-English research followed techniques employed for the English language, with only little use of unambiguous linguistic features such as morphological variation. According to all reports, the field of application appears to be confined to specified topics.

In another paper, [29], the author presented a technique for producing an area-explicit lexicon utilizing a probabilistic approach together with monetary-based information, i.e., stock value weight. This work is different from common approaches in that it builds area-explicit dictionaries and the opinion scores are processed by considering the earlier stock value change, achieving an average of 50% of accuracy.

In the overview in [30], several potential computation enhancements and many SA applications are explored and briefly described. These items were categorized according to their function among the various SA processes. Disciplines related to SA that have recently attracted specialists are discussed. The primary goal of this research was to provide a nearly complete picture of SA techniques and related disciplines.

## 3 Materials and Methods

To apply sentiment analysis to Gujarati, a text dataset had to be created by the authors of the present paper due to the unavailability of a standard dataset, which required effort, research, and time from the authors' end. The authors had to perform text processing to generate more accurate results. To apply the lexicon-based and machine learning-based techniques without a valid dataset for result analysis was a challenging task, even though the authors overcame these challenges and managed to produce satisfactory results by applying the lexicon-based and machine learning-based techniques.

### 3.1 Data Extraction

Data extraction is a critical step in big data analysis. The Beautiful Soup library, which is a free and open-source library, was used to collect data from various movie review websites. It is possible to retrieve reviews automatically based on a single seed URL. A seed URL allows us to navigate across all web pages of a domain. After extracting reviews, a sentiment score of 0 or 1 was assigned to each review based on the ratings given on the website for the movie by critics/users; if the rating of the movie was greater than 3, a score of 1 was assigned, meaning 'good movie'; if the rating was less than 3, a score of 0 was assigned, meaning 'terrible film', as stated in Figure 1.

| | text | experience |
|---|---|---|
| 0 | વાર્તા એક સમયે નાની વકીલ મીરા કપૂર પરિણીતી ચો... | 1 |
| 1 | જીવનની વિવિધ પસંદગીઓ ધરાવતા નિષ્ક્રિય પરિવારન... | 1 |
| 2 | બદનામ આઇએએસ અધિકારી ચંચલ ચૌહાણ માટે વાર્તા જી... | 1 |
| 3 | વાર્તાસોલસૌથી અણધારી જગ્યાએ જોવા મળે છે જે સા... | 1 |
| 4 | વાર્તા તેને કબૂતરનો અનાસતો કહે છે પરંતુ મધુ મ... | 0 |
| 5 | વાર્તા જ્યારે પીટીના એક યુવાન શિક્ષકને નવા કો... | 1 |
| 6 | વાર્તા વિવિધ પાત્રો સાથેની અનેક વાર્તાઓ એક સા... | 1 |
| 7 | લક્ષ્મી વાર્તા રશ્મિ કિયારા અડવાણી તેને સંબંધ... | 1 |
| 8 | કોઇ રોમેન્ટિક ભૂતકાળ સાથે લગ્ન કરવા માટે તલપા... | 1 |
| 9 | એક કમનસીબ રાત્રે સ્થાનિક કેબી બ્લેકી ઇશાન ખટ્... | 0 |
| 10 | દલિત પરિવારની વાર્તા આયં મણિ નવાઝુદ્દીન સિદ્... | 0 |
| 11 | વાર્તા કાજલ સ્વતંત્ર જીવન જીવવા ઇચ્છતી એક યુવ... | 1 |
| 12 | હોમો રાક્ષસ ો લા હોમો સેપિયન્સ અને વડા પ્રધાન... | 1 |
| 13 | રહસ્યમય સંજોગોમાં તેની માતા શકુંતલા દેસાઇના મ... | 0 |
| 14 | વાર્તા જ્યારે એક ટોચના પોલીસને પોલીસ તાલીમ એક... | 1 |
| 15 | ખુદા હાફિઝ વાર્તા નવદંપતી નરગિસ શિવાલીકા ઓબેર... | 1 |
| 16 | વાર્તા આ ફિલ્મમાં ભારતીય વાયુસેનાના ભૂતપૂર્વ ... | 1 |
| 17 | વાર્તા જીવન પ્રિન્ટિંગ પ્રેસના કર્મચારી નંદન ... | 1 |
| 18 | વાર્તા રહસ્યમય સંજોગોમાં તેના લગ્નની રાત્રે પ... | 1 |
| 19 | શકુંતલા દેવીના જીવન પર એક જીવનચરિત્રાત્મક નાટ... | 1 |

**Figure 1** Movie review dataset.

Based on reviews from various websites, the datasets were named Gujwebiduniya, Hungama, Times of India, User, and FilmFare. These names are used in this paper to refer to these datasets. Table 2 lists the sources the reviews were retrieved from.

**Table 2** Details of extracted reviews.

| Sr.no. | Website URL | No. of reviews |
|---|---|---|
| 01. | https://gujarati.webdunia.com/movie-review (the source language of this website is Gujarati) | 232 |
| 02. | https://www.bollywoodhungama.com (machine translation was used to convert reviews from English to Gujarati) | 592 |
| 03. | https://www.filmfare.com/reviews (machine translation was used to convert reviews from English to Gujarati) | 396 |
| 04. | https://timesofindia.indiatimes.com/entertainment/movie-reviews (machine translation was used to convert reviews from English to Gujarati) | 572 |
| 05. | Manual collection of reviews from users in the Gujarati language | 293 |

## 3.2 Data Preprocessing

Data preparation entails adequate data fragmentation and data cleaning. In this study, we utilized NLP preparation techniques [6], such as stopword removal, tokenization, and so on. Data preparation results in more robust data with reduced noise.

### 3.2.1 Removing Stop Words, Special Characters, and Hyperlinks

Stop words do not have any effect on the opinion expressed, so they must be removed for normalization.

### 3.2.2 Tokenization into Sentence and Word

Delimiters are used to split textual material into sentences, which are then broken down into words.

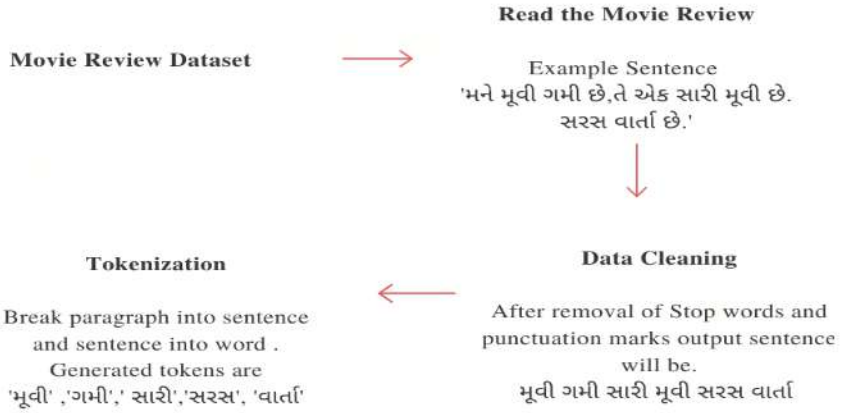


**Figure 2** Gujarati text pre-processing steps.

### 3.3 GLSA (Gujarati Sentiment Lexicon (Lexicon) based approach)

Sentiment classification based on a lexicon approach was used to help to identify sentiments related to film reviews in the Gujarati language, where overall review polarity was graded as positive or negative using Gujarati SentiWordNet. The Synset Replacement Algorithm was used instead of accuracy enhancement for Gujarati words that are not available in Gujarati SentiWordNet. Figure 3 represents the lexicon-based method [31].

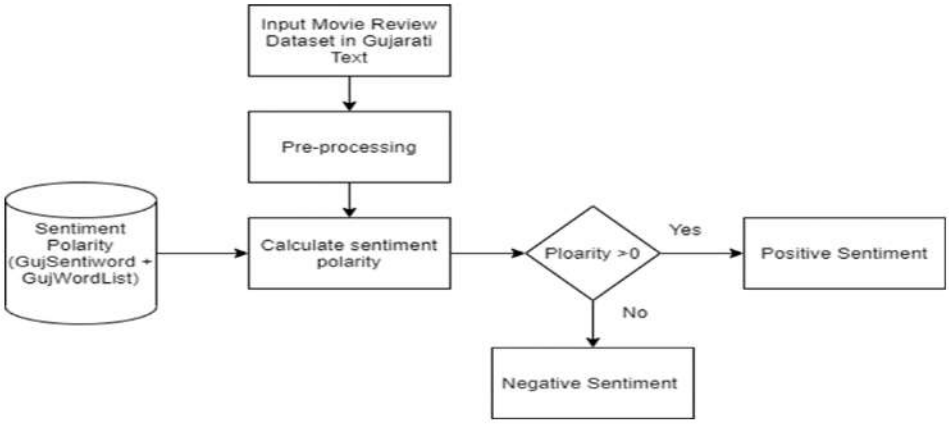


**Figure 3** Lexicon learning-based proposed architecture.

Sentiment resources were generated for Gujarati text, called GujaratiSentiWordNet, as shown in Figure 4. To generate GujaratiSentiWordNet, the Indowordnet interface was used by mapping the synset of HindiWordNet to GujaratiWordNet. The features that are crucial to determining sentiments regarding movie reviews accessible in Gujarati were selected based on sentiment polarity in this technique.

| | POS_TAG | ID | POS | NEG | LIST_OF_WORDS |
|---|---|---|---|---|---|
| 0 | a | 10363 | 0.000 | 0.000 | અનૌપચારિક |
| 1 | a | 2627 | 0.000 | 0.750 | મૃત,મૂએલું,મૃતક,મુડદું,શબ,વૈકુંઠવાસી,પરલોકવાસી... |
| 2 | a | 11476 | 0.125 | 0.000 | પરવર્તી |
| 3 | a | 28106 | 0.250 | 0.375 | સારી,સારું,અચ્છા,યોગ્ય,વાજબી,ઠીક |
| 4 | a | 1156 | 0.875 | 0.000 | ભાગ્યશાળી,નસીબદાર,સદ્‌ભાગી,સુભાગી,ખુશનસીબ,ખુશકિ... |
| 5 | a | 2279 | 0.000 | 1.000 | અભાગી,દુર્ભાગી,કમનસીબ,મનહૂસ,બદકિસ્મત,દુર્ભાગ્ય... |
| 6 | a | 2384 | 0.000 | 0.875 | બેઘર,આવાસહીન,ઘરવિહોણું,અનિકેત,નિવાસ_રહિત |
| 7 | a | 4714 | 0.250 | 0.125 | સુગંધિત,સુવાસિત,સુગંધીદાર,ખુશબોદાર,ખુશ્બૂદાર,સ... |
| 8 | a | 1488 | 0.000 | 0.750 | દુર્ગંધિત,દુર્ગંધવાળું,ગંધાતું |
| 9 | a | 29150 | 0.000 | 0.000 | લાગેલું,લગાવેલું |

**Figure 4** Sentiment resource generation for Gujarati text.

### 3.4 GMLSA (Gujarati Sentiment Analysis with Machine Learning Approach)

A prediction-based approach was applied to movie reviews that were prepared in the Gujarati Language, as shown in Figure 5. An experiment was conducted on a different dataset and showed that language-specific enhanced results were achieved.

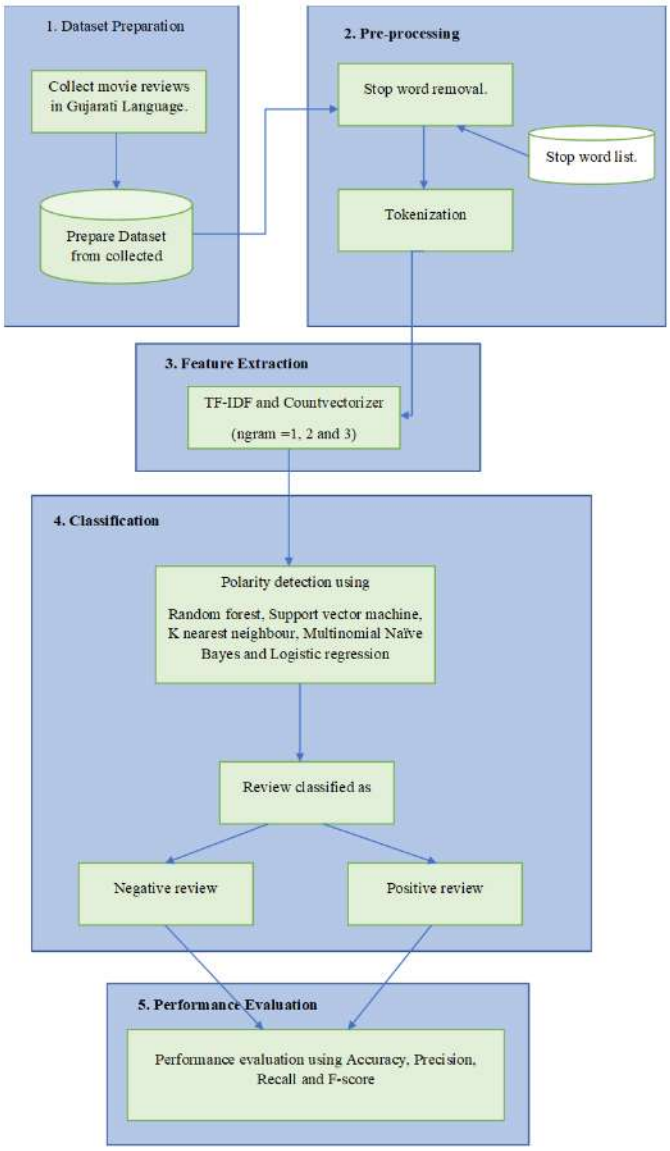


**Figure 5** Proposed machine learning-based architecture.

To achieve the desired results, we performed data preprocessing, which provides a list of tokens that are helpful in the feature selection task. A feature vector generated using TF-IDF and the count vectorizer technique was used as input for different machine learning based classifiers, which generated a confusion matrix, based on which the accuracy of the different classifiers was measured. Minor accuracy variations may occur after applying the same model to a different dataset, however, the model generated adequate results [32-35].

## 4 Results and Discussion

Five different datasets were created to perform these experiments, as shown in Table 2. These datasets were subjected to preprocessing methods to balance them to achieve more accurate results. The assessment metrics used were correctness, precision, recall, and F-measure. GujaratiSentiwordnet was created for the evaluation of the proposed lexicon-based architecture to classify the sentiments of the reviews as positive or negative. Figure 6 shows the resource-based classification technique used with GujaratiSentiWordNet for evaluating the proposed method to identify the sentiments for each movie and classify them as 0 (negative) or 1 (positive).

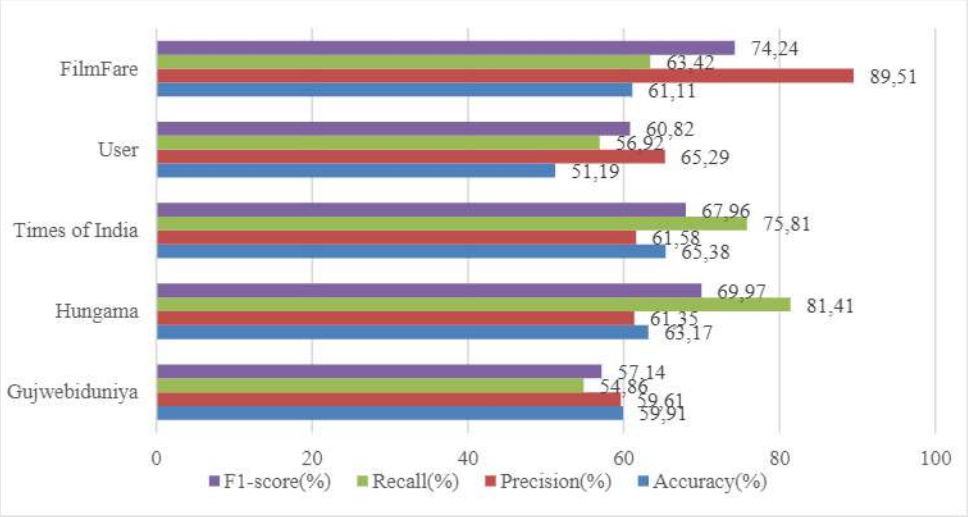


**Figure 6** Dataset-wise performance evaluation based on sentiment score.

Table 3, shows the accuracy, precision, recall, and F-score generated by the lexicon-based method on the movie review dataset. According to Figure 5, the accuracy scores were satisfactory, which means GujaratiSentiWordNet could be used as a benchmark.

**Table 3** Result generated based on sentiment score.

| Dataset Name | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| Gujwebiduniya | 59.91 | 59.61 | 54.86 | 57.14 |
| Hungama | 63.17 | 61.35 | 81.41 | 69.97 |
| Times of India | 65.38 | 61.58 | 75.81 | 67.96 |
| User | 51.19 | 65.29 | 56.92 | 60.82 |
| Filmfare | 61.11 | 89.51 | 63.42 | 74.24 |

The proposed AI-based system was tested for execution utilizing a split rate to decide the preparation and test sets. By and large, the system performed better with 70% preparation information and 30% testing information. The examination was finished utilizing five unique classifiers on the five different datasets. Table 4 shows the exact correlation of the five different classifiers in the various datasets utilizing TF-IDF and CV as element choices.

**Table 4** Dataset-wise accuracy comparison of all classifiers using TF-IDF and CV.

| Dataset Name / Model Name | Gujwebiduniya | | Hungama | | Times of India | | User | | Filmfare | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TFIDF-Accuracy (%) | N-Gram-Accuracy (%) | TFIDF-Accuracy (%) | N-Gram-Accuracy (%) | TFIDF-Accuracy (%) | N-Gram-Accuracy (%) | TFIDF-Accuracy (%) | N-Gram-Accuracy (%) | TFIDF-Accuracy (%) | N-Gram-Accuracy (%) |
| **Multinomial Naive Bayes** | 77.14 | 75.51 | 91.57 | 86.4 | 98.26 | 86.67 | 57.95 | 69.35 | 86.55 | 95.24 |
| **Random Forest** | 88.57 | 75.51 | 96.63 | 92.8 | 91.86 | 92.5 | 57.95 | 72.58 | 96.64 | 95.24 |
| **Logistic regression** | 80 | 75.51 | 93.26 | 86.4 | 97.09 | 89.17 | 57.95 | 70.97 | 86.55 | 95.24 |
| **Support Vector Machine** | 88.57 | 73.47 | 95.51 | 82.4 | 97.09 | 85.83 | 57.95 | 64.52 | 96.64 | 95.24 |
| **KNeighborsClassifier** | 58.57 | 65.31 | 74.16 | 73.6 | 75.58 | 58.33 | 54.55 | 46.77 | 85.71 | 59.52 |

Figure 7 shows the effectiveness-based correlation of the classifiers. All classifiers performed well on the various datasets and achieved a precision of over 75%, except for KNN.

Five different datasets of movie reviews created in the Gujarati language were tested using the two proposed framework architectures. The results are shown in Table 5. All five classifier generated higher accuracy result compared to the lexicon-based approach. Considering the Hungama, Times of India, and Filmfare datasets, the machine learning strategy had 10% greater accuracy when compared to the lexicon method. The ML-based technique outperformed the lexicon-based approach by 10% on the Hungama, Times of India, and Filmfare datasets. On the datasets with reviews from users and Gujwebiduniya, both approaches got nearly identical results. Based on the generated and compared findings, it was determined that the proposed ML-based framework produced more accurate results than the proposed lexicon-based approach.

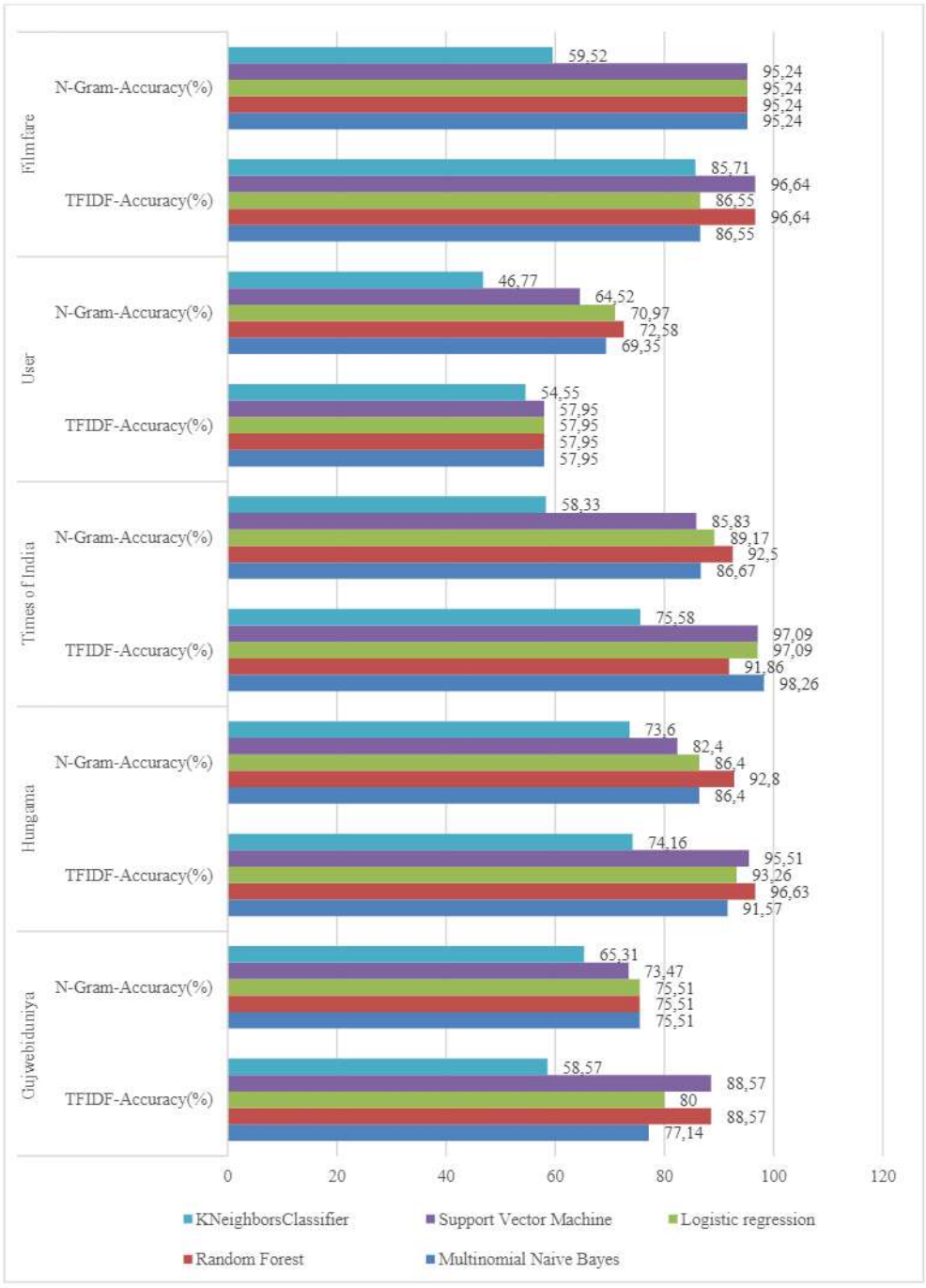


**Figure 7** Dataset-wise accuracy comparison of all classifiers using TF-IDF and CV.

**Table 5** Overall performance-based comparison of the ML-based and lexicon-based methods.

| Dataset Name | Classifier | Feature Selection Method | TFIDF-Accuracy (%) | CV-Accuracy (%) | TF-IDF-Precision (%) | CV-Precision (%) | TF-IDF Recall (%) | CV-Recall (%) | TF-IDF F-Score (%) | CV-F-Score (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| Gujwebiduniya | Multinomial Naive Bayes | TF-IDF and CV | 77.14 | 75.51 | 100 | 66.67 | 58.97 | 80 | 74.19 | 72.73 |
| | Random Forest | | 88.57 | 75.51 | 94.29 | 72.22 | 84.62 | 65 | 89.19 | 68.42 |
| | Logistic regression | | 80 | 75.51 | 100 | 75.51 | 64.1 | 75.51 | 78.13 | 75.51 |
| | Support Vector | | 88.57 | 73.47 | 100 | 62.96 | 79.49 | 85 | 88.57 | 72.34 |
| | KNeighborsClassifier | | 58.57 | 65.31 | 65.63 | 54.29 | 53.85 | 95 | 59.15 | 69.09 |
| | Lexicon Based | | 59.91 | | 59.61 | | 54.86 | | 57.14 | |
| Hungama | Multinomial Naive Bayes | TF-IDF and CV | 91.57 | 86.4 | 87.25 | 90.32 | 97.8 | 83.58 | 92.23 | 86.82 |
| | Random Forest | | 96.63 | 92.8 | 95.7 | 93.94 | 97.8 | 92.54 | 96.74 | 93.23 |
| | Logistic regression | | 93.26 | 86.4 | 89.9 | 86.4 | 97.8 | 86.4 | 93.68 | 86.4 |
| | Support Vector | | 95.51 | 82.4 | 93.68 | 89.47 | 97.8 | 76.12 | 95.7 | 82.26 |
| | KNeighborsClassifier | | 74.16 | 73.6 | 70.27 | 80.36 | 85.71 | 67.16 | 77.23 | 73.17 |
| | Lexicon Based | | 63.17 | | 61.35 | | 81.41 | | 69.97 | |
| Times of India | Multinomial Naive Bayes | TF-IDF and CV | 91.57 | 86.4 | 87.25 | 90.32 | 97.8 | 83.58 | 92.23 | 86.82 |
| | Random Forest | | 96.63 | 92.8 | 95.7 | 93.94 | 97.8 | 92.54 | 96.74 | 93.23 |
| | Logistic regression | | 93.26 | 86.4 | 89.9 | 86.4 | 97.8 | 86.4 | 93.68 | 86.4 |
| | Support Vector | | 95.51 | 82.4 | 93.68 | 89.47 | 97.8 | 76.12 | 95.7 | 82.26 |
| | KNeighborsClassifier | | 74.16 | 73.6 | 70.27 | 80.36 | 85.71 | 67.16 | 77.23 | 73.17 |
| | Lexicon Based | | 65.38 | | 61.58 | | 75.81 | | 67.96 | |
| User | Multinomial Naive Bayes | TF-IDF and CV | 57.95 | 69.35 | 57.95 | 75 | 100 | 86.67 | 73.38 | 80.41 |
| | Random Forest | | 57.95 | 72.58 | 57.95 | 72.58 | 100 | 100 | 73.38 | 84.11 |
| | Logistic regression | | 57.95 | 70.97 | 57.95 | 70.97 | 100 | 70.97 | 73.38 | 70.97 |
| | Support Vector | | 57.95 | 64.52 | 57.95 | 73.47 | 100 | 80 | 73.38 | 76.6 |
| | KNeighborsClassifier | | 54.55 | 46.77 | 56.79 | 92.86 | 90.2 | 28.89 | 69.7 | 44.07 |
| | Lexicon Based | | 51.19 | | 65.29 | | 56.92 | | 60.82 | |
| FilmFare | Multinomial Naive Bayes | TF-IDF and CV | 86.55 | 95.24 | 86.55 | 95.06 | 100 | 100 | 92.79 | 97.47 |
| | Random Forest | | 96.64 | 95.24 | 96.26 | 95.06 | 100 | 100 | 98.1 | 97.47 |
| | Logistic regression | | 86.55 | 95.24 | 86.55 | 95.24 | 100 | 95.24 | 92.79 | 95.24 |
| | Support Vector | | 96.64 | 95.24 | 96.26 | 95.06 | 100 | 100 | 98.1 | 97.47 |
| | KNeighborsClassifier | | 85.71 | 59.52 | 86.44 | 93.88 | 99.03 | 59.74 | 92.31 | 73.02 |
| | Lexicon Based | | 61.11 | | 89.51 | | 63.42 | | 74.24 | |

## 5 Conclusion

The techniques proposed in this work attempted to improve the quality of sentiment analysis on textual movie reviews in Gujarati. For sentiment analysis on Gujarati text, lexicon and machine learning-based techniques were presented in this paper. The authors created GujSentiWordNet, a sentiment lexicon enhanced with a synonym lexicon, and deployed the two proposed classification methods to five different datasets to assess the accuracy of Gujarati text lexicon-based sentiment analysis (GLSA). On the ML side, five classifiers were employed, MNB, SVM, RF, KNN, and LR, on five different datasets to ensure reliable functioning Gujarati text machine learning-based sentiment analysis (GMLSA).

Based on the results of both proposed methods it can be concluded that the ML-based strategy provides greater accuracy than the lexicon-based approach. When compared to lexicon-based techniques, each of the five classifiers achieved good precision results. When compared to the lexicon-based strategy, the ML-based method had a 10% higher precision when using the Hungama, Times of India, and Filmfare datasets. With the Hungama, Times of India, and Filmfare datasets, the ML-based technique outperformed the lexicon-based methods by 10%. For the datasets with reviews from users and Gujwebiduniya, the two approaches got nearly identical results. Given the different findings, it is not certain that the

suggested ML-based approach produces more accurate results than the proposed lexicon-based methodology.

In our technological age, everyone expresses their opinions on social media platforms on a regular basis and because these perspectives are generally voiced in regional languages, most of the information created is in regional languages. As a result, to make reliable predictions, it is also important to target those audiences. We can assist the audience in using technology effectively for their benefit by generating evaluations, suggestions, or comments in their native language. The result of this research can be used as a baseline for applications such as agriculture helplines for farmers, tourism, voice assistance, question-answer systems, etc. In the future, lexicon-based algorithms can be applied to enormous amounts of data to provide more accurate results by developing domain-specific lexicons. Similarly, a machine learning-based method may be used on a huge dataset to get more accurate results.

## References

[1] Kaur, J. & Saini, J.R., *A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families*, International Journal of Data Mining and Emerging Technologies, Diva Enterprises Private Limited, **4**(2), 53, 2014. DOI: 10.5958/2249-3220.2014.00002.0

[2] Gukanesh, A.V. & Kumar, G.K., *Saranya KKRK| N. Twitter Data Analytics – Sentiment Analysis of an Election*, International Journal of Trend in Scientific Research and Development, South Asia Management Association, **2**(3), pp. 1600-1603, 2018. DOI: 10.31142/ijtsrd11457.

[3] Fouad, M.M. & Gharib, T.F. & Mashat A.S., *Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble*, Advances in Intelligent Systems and Computing, Springer International Publishing, pp. 516-527, 2018. DOI: 10.1007/978-3-319-74690-6_51.

[4] Ahuja, R., Chug, A., Kohli, S., Gupta, S. & Ahuja, P., *The Impact of Features Extraction on the Sentiment Analysis*, Procedia Computer Science, Elsevier BV, **152**, pp. 341-348, 2019. DOI: 10.1016/j.procs.2019.05.008.

[5] Gohil, L. & Patel, D., *A Sentiment Analysis of Gujarati Text using Gujarati Senti Word Net*, Regular Issue, Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, **8**(9), pp. 2290-2292, 2019. DOI: 10.35940/ijitee.i8443.078919.

[6] Mumtaz, D. & Ahuja, B., *Sentiment Analysis of Movie Review Data Using Senti-Lexicon Algorithm*, 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), IEEE, 2016. DOI: 10.1109/icatcct.2016.7912069.

[7] Jha, V., Manjunath, N., Shenoy, P.D., Venugopal, K.R. & Patnaik, L.M., *HOMS: Hindi Opinion Mining System,* 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), IEEE; Jul 2015. DOI: 10.1109/retis.2015.7232906.

[8] Liu, B., *Opinion Mining and Sentiment Analysis*, Web Data Mining, Springer Berlin Heidelberg, pp. 459-526, 2011. DOI: 10.1007/978-3-642-19460-3_11.

[9] Pandey, P. & Govilkar, S., *A Framework for Sentiment Analysis in Hindi using HSWN. International Journal of Computer Applications*, Foundation of Computer Science, **119**(19), pp. 23-26, 2015. DOI: 10.5120/21176-4185.

[10] Rehman, Z.U. & Bajwa, I.S., *Lexicon-based Sentiment Analysis For Urdu Language,* 2016 Sixth International Conference on Innovative Computing Technology (INTECH), IEEE, 2016. DOI: 10.1109/intech.2016.7845095.

[11] Popale, L. & Bhattacharyya, P., *Creating Marathi WordNet*, The WordNet in Indian Languages. Springer Singapore, pp. 147-66, 2016. DOI: 10.1007/978-981-10-1909-8_8.

[12] Rohini, V., Thomas, M. & Latha, C.A., *Domain Based Sentiment Analysis in Regional Language-Kannada Using Machine Learning Algorithm*, 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, May 2016. DOI: 10.1109/rteict.2016.7807872.

[13] Joshi, V.C. & Vekariya, V.M., *An Approach to Sentiment Analysis on Gujarati Tweets. Advances in Computational Sciences and Technology*, **10**(5), pp. 1487-1493, 2017.

[14] Mishra, A., Joshi, A. & Bhattacharyya, P., *A Cognitive Study of Subjectivity Extraction in Sentiment Annotation*. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2014. DOI: 10.3115/v1/w14-2623.

[15] Ghosal, T., Das, S.K. & Bhattacharjee, S., *Sentiment Analysis on (Bengali horoscope) Corpus,* 2015 Annual IEEE India Conference (INDICON), IEEE, 2015. DOI: 10.1109/indicon.2015.7443551.

[16] Se, S., Vinayakumar, R., Anand Kumar, M. & Soman, K.P., *Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms*, Indian Journal of Science and Technology, Indian Society for Education and Environment, **9**(45), pp. 1-5, 2016. DOI: 10.17485/ijst/2016/v9i45/106482.

[17] Kaur, A. & Gupta, V., *N-gram Based Approach for Opinion Mining of Punjabi Text*, Multi-disciplinary Trends in Artificial Intelligence, Springer International Publishing, pp. 81-88, 2014. DOI: 10.1007/978-3-319-13365-2_8.

[18] Deepamala, N. & Ramakanth Kumar, P., *Polarity Detection of Kannada Documents,* 2015 IEEE International Advance Computing Conference (IACC), IEEE, 2015, DOI: 10.1109/iadcc.2015.7154810.

[19] Venugopalan, M. & Gupta, D., *Sentiment Classification for Hindi Tweets in a Constrained Environment Augmented Using Tweet Specific Features, Lecture Notes in Computer Science,* Springer International Publishing, pp. 664-670, 2015. DOI: 10.1007/978-3-319-26832-3_63.

[20] Prasad, S.S., Kumar, J., Prabhakar, D.K. & Pal, S., *Sentiment Classification: An Approach for Indian Language Tweets Using Decision Tree,* Lecture Notes in Computer Science, Springer International Publishing, pp.656-663, 2015. DOI: 10.1007/978-3-319-26832-3_62.

[21] Ashna, M.P. & Sunny, AK., *Lexicon Based Sentiment Analysis System for Malayalam Language*, 2017 International Conference on Computing Methodologies and Communication (ICCMC), IEEE, Jul 2017. DOI: 10.1109/iccmc.2017.8282571.

[22] Mittal, N., Agarwal, B., Chouhan, G., Pareek, P. & Bania, N., *Discourse Based Sentiment Analysis for Hindi Reviews,* Pattern Recognition and Machine Intelligence, Springer Berlin Heidelberg, pp. 720-725, 2013. DOI: 10.1007/978-3-642-45062-4_102.

[23] Sarkar, K. & Chakraborty, S., *A Sentiment Analysis System for Indian Language Tweets. Lecture Notes in Computer Science,* Springer International Publishing, pp. 694-702, 2015. DOI: 10.1007/978-3-319-26832-3_66.

[24] Deepamala, N. & Ramakanth Kumar, P., *Polarity detection of Kannada Documents*, 2015 IEEE International Advance Computing Conference (IACC), IEEE, 2015. DOI: 10.1109/iadcc.2015.7154810.

[25] Kour, K., Kour, J. & Singh, P., *Lexicon-Based Sentiment Analysis. Advances in Communication and Computational Technology,* Springer Singapore, pp. 1421–30, 2020. DOI: 10.1007/978-981-15-5341-7_108.

[26] Machová, K., Mikula, M., Gao, X. & Mach, M., *Lexicon-based Sentiment Analysis Using the Particle Swarm Optimization. Electronics,* MDPI AG, **9**(8), 1317, 2020. DOI: 10.3390/electronics9081317.

[27] Esuli, A., Sebastiani, F. & Abasi, A., *AI and Opinion Mining, Part 2. IEEE Intelligent Systems*, Institute of Electrical and Electronics Engineers (IEEE), **25**(4), pp, 72-79, 2010. DOI: 10.1109/mis.2010.94.

[28] Medagoda, N., Shanmuganathan, S. & Whalley, J., *A Comparative Analysis of Opinion Mining and Sentiment Classification in Non-English Languages,* 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE, 2013. DOI: 10.1109/icter.2013.6761169.

[29] Turner, Z., Labille, K. & Gauch, S., *Lexicon-Based Sentiment Analysis for Stock Movement Prediction*, Journal of Construction Materials, Institute of Construction Materials, **2**(3), 2021. DOI: 10.36756/jcm.v2.3.5.

[30] Feldman, R., *Techniques and Applications for Sentiment Analysis,* Communications of the ACM, Association for Computing Machinery (ACM), **56**(4), pp. 82-9, 2013. DOI: 10.1145/2436256.2436274.

[31] Shah, P. & Swaminarayan, P., *Lexicon-Based Sentiment Analysis on Movie Review in the Gujarati Language*, International Journal of Information Technology, Communications and Convergence, Inderscience Publishers, **4**(1), pp. 63-72, 2021. DOI: 10.1504/ijitcc.2021.10042767.

[32] Shah, P.V. & Swaminarayan, P., *Sentiment Analysis – An Evaluation of the Sentiment of the People: A Survey*, Data Science and Intelligent Applications. Springer Singapore, pp. 53-61, 2020. DOI: 10.1007/978-981-15-4474-3_6.

[33] Shah, P., Swaminarayan, P. & Patel, M., *Sentiment Analysis on Film Review in Gujarati Language Using Machine Learning*, International Journal of Electrical and Computer Engineering (IJECE). Institute of Advanced Engineering and Science, **12**(1), pp. 1030-1039, 2022. DOI: 10.11591/ijece.v12i1.pp1030-1039.

[34] Shah, P.V. & Swaminarayan, P., *Sentiment Analysis on Gujarati Text: A Survey,* Journal of Computational and Theoretical Nanoscience, American Scientific Publishers, **17**(9), pp. 4075-4082, 2020. DOI: 10.1166/jctn.2020.9022.

[35] Shah, P., Swaminarayan, P., Patel, M. & Patel, N., *Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm*, International Journal of Engineering Trends and Technology. Seventh Sense Research Group, **70**(1), pp. 313-326, 2022. DOI: 10.14445/22315381/IJETT-V70I1P236.