Sentiment analysis on film review in Gujarati language using machine learning

Parita Shah[1,2], Priya Swaminarayan[3], Maitri Patel[2]
[1]Faculty of Engineering and Technology, Parul University, Vadodara, India
[2]Department of Computer Engineering, Gandhinagar Institute of Technology, Gandhinagar, India
[3]Faculty of Information Technology and Computer Science, Parul University, Vadodara, India

Article Info

Article history:

Received Jan 21, 2021
Revised Aug 3, 2021
Accepted Aug 20, 2021

Keywords:

Classifier
Features selection
Film
Gujarati
Precision
Sentimentality

ABSTRACT

Opinion analysis is by a long shot most basic zone of characteristic language handling. It manages the portrayal of information to choose the motivation behind the wellspring of the content. The reason might be of a type of gratefulness (positive) or study (negative). This paper offers a correlation between the outcomes accomplished by applying the calculation arrangement using various classifiers for instance K-nearest neighbor and multinomial naive Bayes. These techniques are utilized to assess a significant assessment with either a positive remark or negative remark. The gathered information considered on the grounds of the extremity film datasets and an association with the results accessible proof has been created for a careful assessment. This paper investigates the word level count vectorizer and term frequency inverse document frequency (TF-IDF) influence on film sentiment analysis. We concluded that multinomial naive Bayes (MNB) classier generate more accurate result using TF-IDF vectorizer compared to CountVectorizer, K-nearest-neighbors (KNN) classifier has the same accuracy result in case of TF-IDF and CountVectorizer.

Corresponding Author:

Parita Shah
Faculty of Engineering and Technology, Parul University
Waghodia, Vadodara - 391760, Gujarat, India
Email: paritaponkiya@gmail.com

1. INTRODUCTION

Regular language measure is utilized in the region of examination of human conclusion and feeling, for the most part, centers around feeling, disposition, and assessment because of increment utilization of the web [1], [2]. In Indian dialects like Gujarati, language autonomy is basic because of helpless assets, Robustness, and versatility. In this paper, feeling investigation is done on film evaluations available in the Gujarati language to accomplish best in class execution utilizing different AI rehearses. In today's world where everyone relies on web. Sentiment of user becomes important entity because, if anyone wants to buy a new product, enroll for a course, or want to watch a movie they will first find out the review and based on that review they are making their decisions. Due to increased use of web, automation is required so here role of natural language processing (NLP) become crucial. Number of NLP based applications are available used for text translation, retrieval and summarization that is helpful in identifying opinion or feedback people, spam detection, fake news identification and providing digital medical assistant. A broad research is currently going on towards developing NLP based application for Indian language [3].

Usage of web/internet has grown faster which gives growth and opportunities for Indian market. Information is generated with high volume and velocity therefore users in India access web in their regional languages, almost all the information's from review of product/movie, news to advertisement's is also available in regional language on a web. Which gives a scope for many researchers to explore field of NLP and sentiment analysis is one of the important aspects of NLP, so it becomes essential to develop resources to analyze sentiment in Indian languages.

There are two major families are available Arya, Dravida to which Indian languages belongs to and 22 authority dialects spoken in India. Indo-Aryan family comprise of dialects as Hindi, Urdu, Bengali, Oriya, Punjabi, Konkani, Marathi, Nepali, Gujarati, Sindhi, Dogri, Assemese, Sanskrit and Kashmiri. Dravidian family comprises of dialects like Telugu, Tamil, Malayalam, and Kannada [4].

Sentiment analysis for movie review is relatively recent and movie review in Gujarati languages is still area of research. So far there is no dataset available which contains movie review in Guajarati language and no research available for identifying sentiment of movies from the review that is written in Guajarati language. Lexicon based and machine translation approach is so far most used technique for identifying sentiment in Guajarati Language till now. In lexicon-based approach a dictionary of words is created which contains polarity of word and sentiment is identify by sum of polarity (each word in sentence) if it is greater than 1 then its positive sentiment else its negative sentiment. In latest paper they have used lexicon-based sentiment analysis to classify tweets available in Gujarati language by creating SentiWord dictionary for Gujarati words using IndoWordNet interface [2]. Enough resources are available for English Language therefore so many research available for sentiment analysis in English language. As of now information on web available in regional language also it motivates us to perform sentiment analysis on Gujarati language which is 6th highest speaking language in India. We have prepared dataset for movie review in Gujarati language to perform sentiment analysis, but it is a challenging task. Sufficient resources are not available such as corpus and language tagger, makes assessment a troublesome endeavour. We must create our own dataset as no standard dataset available this also requires efforts, research, and time from our end. We must perform text processing for generation more accurate result. To apply machine learning based technique without any valid dataset for result analysis was a challenging task though we have overcome these challenges and mange to produce satisfactory results by applying machine leaning based technique with term frequency inverse document frequency (TF-IDF) and CountVectorizer (CV) as feature selection technique [4].

There are different components determination methods like N-Gram, TF-IDF, count vector and word incorporating that are accessible for AI-based more tasteful, and execution of this classifier can be estimated with various execution boundaries, for example, recall, F-score, accuracy and precision [5]. Preprocessing the initial huge phase in text order which incorporates tasks like sentence/word tokenization, stop word expulsion, evacuation of unique characters, and numbers. Next advance is feature determination [6], [7]. Numerous strategies are accessible for feature determination, for example, a bunch of words, TF-IDF, count vectors, and word embeddings which depend on characteristic language preparing [8], [9]. The last advance is to apply AI calculation for the order of perspective, for example, K-nearest-neighbors (KNN) and multinomial naive Bayes (MNB). Impact of two-word level features count and TF-IDF vectorizer assortment have been tended to in this paper with the utilization of two classifiers MNB and KNN for discerning opinion exactness.

In this paper, we have proposed machine learning based sentiment analysis for movie reviews in Gujarati language (MSAGL). As shown in Table 1, there is extensive research is done in language like Hindi but no sufficient research available in language like Gujarati here we reach out on that work severally [4]. In step one dataset for movie review (in Gujarati language) is created by extracting reviews from a website in https://gujarati.webdunia.com/with the extremity is ready for investigation which comprises of negative survey spoke to by 0 and positive audit spoke to by 1. All surveys are kept up in a comma separate record. In step two pre-processing of dataset is done by removing unwanted characters, words and noise as reviews are collected from internet so polishing of data is important to achieve more accurate result. In stage three tokenization process is done on dataset using TF-IDF and CV feature selection method. In stage four Separated features are feed to two unique classifiers multinomial naive Bayes classifier and k-nearest neighbor to perform sentiment analysis. This model creates a confusion grid in the wake of preparing them. The confusion association shows the positive and surveys that are accurately and wrongly anticipated. In last stage assessment of each model is performed utilizing an execution boundary such as accuracy, precision, recall and F-score. A good amount of effort and time is invested to prepare dataset which contains movie review in Gujarati language (500 reviews) and to prepare stop word list for Gujarati language, for this work, is likewise our commitment and can be made accessible and used in future for research purposes as it were.

2. LITERATURE SURVEY

Significant amount of work has been done on sentiment analysis in past few years, but we have focused on two techniques machine learning and lexicon based that are widely used for different Indian languages as shown in Table 1 [10]-[21]. For the NLP task, they have utilized vector portrayals for effective use of word vector portrayal which gives feeling examination issue arrangement [8]. They thought about part of speech and vocabulary functionality alongside the artificial intelligence (AI) approach in particular support vector machine, logistic regression, and naïve Bayes [22]. A methodology like reliance parsing is utilized in this paper which shows how this methodology is guaranteeing the perspective of short content with the social movement and changed separation, challenges are moderate through assumption structure and the notion estimation standards [23]. They have utilized the SVM classifier with preprocessing steps, for example, stemming, stop words, non-English characters, and nullification expulsion to look at the viability of the film survey dataset [24]. Execution of perspective can be improved by utilizing feature determination procedures, for that they have applied ten distinctive component choice techniques with four classifiers [1]. The perspective examination using naive Bayes (NB) classifier on collected tweets (Hindi, Bengali, and Tamil) in Indian dialects [25]. Tokenization applied on the collected tweet in Indian dialects followed by feature mining with the utilization of SentiWordNet. For a huge dataset NB classier fails to convince otherwise superior performance is given in case of smaller size dataset. For mixed code information method used in this paper follows the process of translating whole content into different dialects called English to identify the extremity of translated content for assessment examination [26]. The interpretation of the code-blended content to a solitary language has a few restrictions, for example, accomplishing theoretical comparability, linguistic, and syntactic structure of the source language [27]. Vocabulary based analysis of sentiment is done in the Telugu language by utilizing SentiWordNet [28]. Vocabulary based attitude examination using SentiWordNet and approach based on machine learning used in this paper to identify a feeling of Telugu sentence into labels called positive and negative [29]. Two different classifiers have been used in this paper to identify improved performance in a language like English. With word-level N-Gram feature for word vectorization used with logistic regression (LR) and NB to be brought improvement in performance [30]. For auditing of item valuably neural network call multilayer perceptron is used for assumption order [31].

Table 1. Studies related to sentiment analysis in Indian language

Author Citations	Techniques	Dataset used	Accuracy (%)	Language
[4]	Synset Replacement Algorithm (GUJ SentiWordNet), WordNet, Bag-of words	Gujarati Tweet	52.72 (unigram)	Gujarati
[3]	Neural Network	Microblogging	Not Measured	Gujarati-English (Gujlish)
[12]	SVM	Hindi tweets	Hindi-49.68 Bengali-43.20	Hindi, Bengali
[11]	Multinomial and Bernoulli Naive Bayes, Logistic Regression, SVM, Random Kitchen Sink	Tamil Movie Reviews	SVM-64.69 (Bigram) MNB-47.21 (Bigram)	Tamil
[17]	Synset Replacement Algorithm (Hindi SentiWordNet)	Tweets, Movie Reviews and Blogs	Not measured	Hindi
[13]	WordNet, Bag-of words	HindMonoCorp 0.5, IMDB11 Movie Review dataset	75.53	Hindi-English
[10]	TnT Tagger	Malayalam Movie Reviews	91.06	Malayalam
[21]	Lexicon based	Hindi Tweets	73.53	Hindi
[20]	SVM	Gujarati Tweets	Not Measured	Gujarati
[14]	Naive Bayes classifier	Movie Reviews	87.1	Hindi
[15]	CRF for Aspect Extraction and SVM for Classification	Product Reviews	54.05	Hindi
[18]	Dictionary Based, Naive Bayes and SVM algorithm	Hindi Tweets related to Political party in India during election 2016	62.1	Hindi
[19]	Lexicon Based, LMC classifier	Hindi speeches delivered by leaders	Not Measured	Hindi
[16]	Lexicon Based, SVM, Random Forests	Not Specified	Not Measured	Hindi, Marathi

This paper presents a successful semantic and extremity-based data recovery methodology for heterogeneous informational collections. Setting of the information inquiry is recognized and every one of the records that fulfill the extremity and setting of the info question are recovered from the information source [22]. They utilized syntactic and semantic probabilities acquired from the WordNet similitudes as the idea connection highlights to prepare the gullible Bayes classifier intended to learn idea relations. The credulous Bayes classifier is bootstrapped by utilizing an assumption augmentation method. The analysis directed utilizing benchmark datasets created promising outcomes. The viability of the proposed strategy was demonstrated by contrasting the exhibition and comparative well performing programmed philosophy development techniques [32]. In this paper, they characterized the extremity of the data article through method of methods for watching the rankings got the utilization of valence aware dictionary for sentiment reasoning (VADER). For tweets or exceptionally speedy writings, record stage slant assessment is a marvelous decision. Notwithstanding, if profiling feeling around a chose factor or highlight of a brand, item, or company is to be done, after which sentence stage or substance stage assessment is thought of [33]. In this paper they have used vocabulary and code mix based approach to accomplishes higher results for identifying sentiment [34]. The experimentation on these paintings consists of sentiment evaluation at the paragraph, sentence, and phrase stage [35].

3. PROPOSED METHOD

Figure 1 indicates the arranged informational index of film audits in the Gujarati language then we pre-handled the gathered information. In the subsequent stage for feature choice, we have utilized two techniques named TF-IDF and CountVectorizer, and those highlights are arranged utilizing two unique classifiers. Finally, we looked at the presentation of various classifiers dependent on the different exhibition measures.

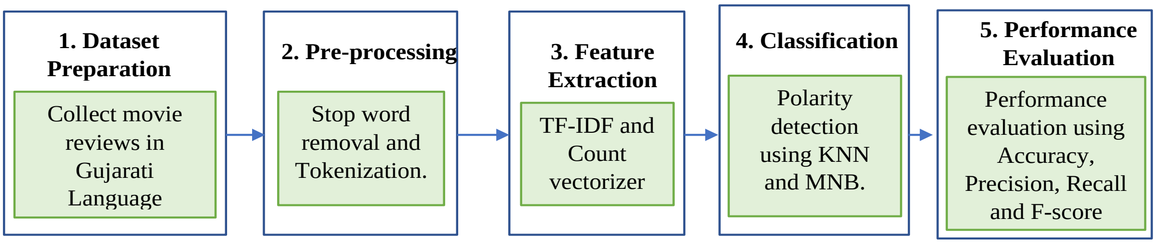
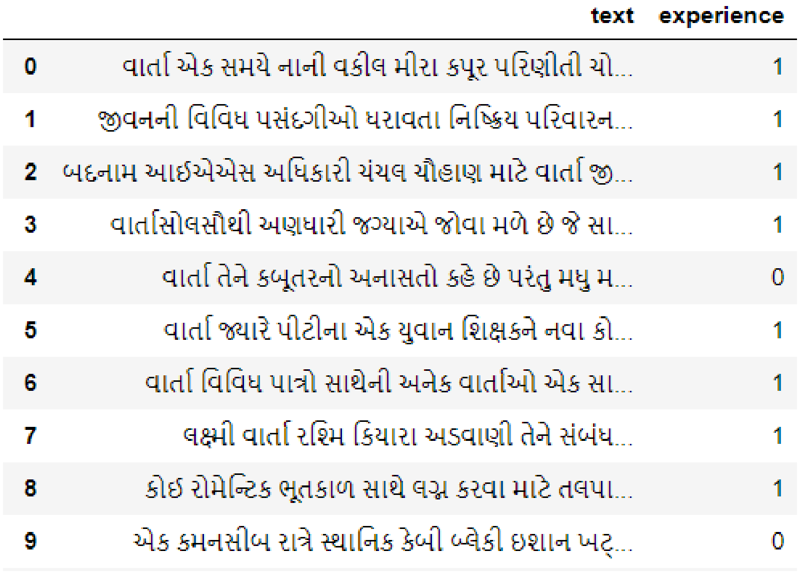

1. Dataset Preparation
Collect movie reviews in Gujarati Language
2. Pre-processing
Stop word removal and Tokenization.
3. Feature Extraction
TF-IDF and Count vectorizer
4. Classification
Polarity detection using KNN and MNB.
5. Performance Evaluation
Performance evaluation using Accuracy, Precision, Recall and F-score

Figure 1. Proposed approach

Classification steps followed are:

a. Stage 1. Informational collection with the extremity is ready for investigation which comprises of negative survey spoke to by 0 and positive audit spoke to by 1. All surveys are kept up in a comma separate record. We have created dataset of movie reviews in Gujarati language by using python crawling (beautifulsoup library used). 500 movie reviews are collected from a website called Gujarati Webidunia in https://gujarati.webdunia.com/ and labelled it with 0 and 1 where 0 represents negative and 1 represents positive as shown Figure 2.



	text	experience
0	વાર્તા એક સમયે નાની વકીલ મીરા કપૂર પરિણીતી ચો...	1
1	જીવનની વિવિધ પસંદગીઓ ધરાવતા નિષ્ક્રિય પરિવારન...	1
2	બદનામ આઈએએસ 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

Figure 2. Movie review dataset prepared in Gujarati language

b. Stage 2. Uncommon characters (@!) and pointless clear space are eliminated followed by the removal of words that do not have any estimation then tokenization is done in the pre-handling step as given in example. Original sentence = થિયેટર માં ફિલ્મ પૂરી થયા પછી અભિવાદન આવતી હોય ત્યારે પણ દર્શક દિગ્મૂઢ થઈને યવનિકા તરફ તાકી જ રહે છે. સામાન્ય રીતે જ્યારે પણ આવું બને છે, ત્યારે એ માનવું જ રહ્યું કે ફિલ્મ માં કંઈક તો બોધ છે. After removal of special character will remove characters unwanted words and we will receive output sentence as થિયેટર ફિલ્મ પૂરી અભિવાદન આવતી દર્શક દિગ્મૂઢ થઈને યવનિકા તરફ તાકી રહે સામાન્ય રીતે જ્યારે આવું બને માનવું રહ્યું ફિલ્મ કંઈક બોધ

c. Stage 3. Tokenized features are separated from clean information utilizing TF-IDF and count vectorizer techniques. Tokenization will split paragraph into sentence and sentence into word such as, ‘થિયેટર’, ‘ફિલ્મ’, ‘પૂરી’, ‘અભિવાદન’, ‘આવતી’, ‘દર્શક’, ‘દિગ્મૂઢ’, ‘થઈને’, ‘યવનિકા’, ‘તરફ’, ‘તાકી’, ‘રહે’, ‘સામાન્ય’, ‘રીતે’,’ જ્યારે’, ‘આવું’, ‘બને’,’ માનવું’,’ રહ્યું’,’ફિલ્મ’,’કંઈક’, ‘બોધ’.

d. Stage 4. Separated features are feed to two unique classifiers (KNN, MNB). This model creates a confusion grid in the wake of preparing them. The confusion association shows the positive and surveys that are accurately and wrongly anticipated.

e. Stage 5. The assessment of each model is performed utilizing an execution boundary.

4. FEATURE SELECTION

4.1. TF-IDF

Assessment technique term occurrence features weigh the noteworthiness of a word in each archive. The recurrence of term event is determined as the occasions a term shows up in a report partition by the word occurrence in the archive. reverse document rate of recurrence likewise ascertains the significance of the term. Inverse document frequency (IDF) is determined as the number of records isolated by the number of reports containing the term t [36]. for instance, there are 300 words in the record, and out of those 20 words are generally incessant than term recurrence will be 20/300=0.066 and assume there are 7000 reports and out of that lone 200 archives contains specific term than IDF=7000/200=35. TF will be 0.06*100=6 and IDF will be 35.

4.2. Count vectorizer

Text is transformed into a vector by marking the presence (1) or absence (0) of a word of a given input [36]. The calculation count vectorizer is stated in Table 2. For the following sentences. The generated matrix contains 2 rows and 3 columns’, a row represents the presence and absence of feature from a sentence.

Sentence 1=મૂવી સારી છે.

Sentence 2=મૂવી ખૂબ સારી છે.

Table 2. CountVectorizer matrix generation

Sentences	મૂવી	ખૂબ	સારી
	(Feature1)	(Feature2)	(Feature3)
Sentence 1	1	0	1
Sentence 2	1	1	1

5. CLASSIFICATION ALGORITHM

5.1. Multinomial naïve Bayes

Calculation based on probability of conditional independence between each pair of features is called Bayes’ theorem and MNB classifier follows the principle of Bayes’ theorem [37]. Consider (1):

$$P(class|feature) = P(feature|class) * \frac{P(class)}{P(feature)} \tag{1}$$

5.2. K-nearest neighbor

KNN follows the principle of similarity by calculating distance (euclidean distance) between points. euclidean distance is calculated [37] as stated in (2):

$$d(x,y) = \sqrt{\sum_{i=1}^{k}(xi - yi)^2} \tag{2}$$

6. EVALUATION PARAMETERS

6.1. Accuracy

The most natural proportion of progress is precision, and it is just the extent of accurately anticipated perception to add up to perceptions [9]. As appeared in (3):

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative} \quad (3)$$

6.2. Precision

Share in optimistic views to the complete positive perceptions anticipated. The low fake positive rate suggests high exactness [9].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

6.3. Recall

Calculation of how many positive genuine portray by our standard via marking it as constructive (true positive) is called recall [9].

$$Recall = \frac{True\ Positive}{True\ Posotive + False\ Negative} \quad (5)$$

6.4. F1-score

It is a weighted balance between recall and precision [9].

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (6)$$

7. RESULTS AND DISCUSSION

Figure 3 shows the confusion matrix generated and accuracy score generated after applying word level TF-IDF. Figure 4 shows the confusion matrix generated and accuracy score generated after applying word level CountVectorizer. Figure 5 shows the confusion matrix generated and accuracy score generated after applying word level TF-IDF. Figure 6 shows the confusion matrix generated and accuracy score generated after applying word level CountVectorizer.

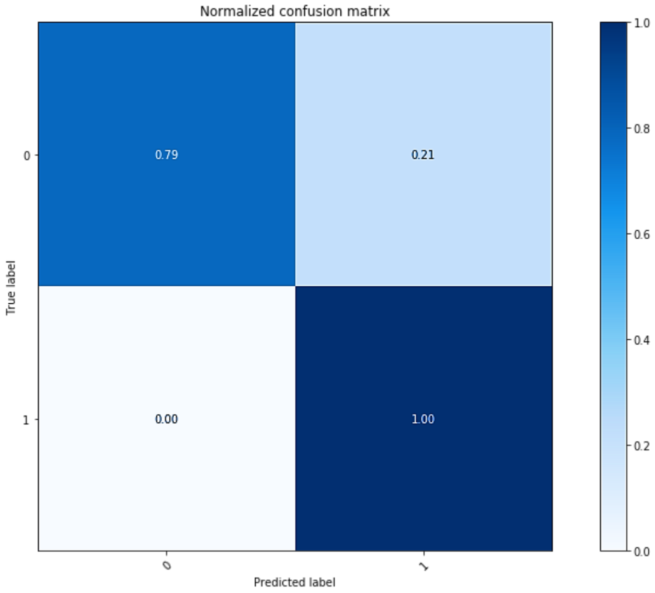


Figure 3. Confusion matrix generated for MNB with TF-IDF

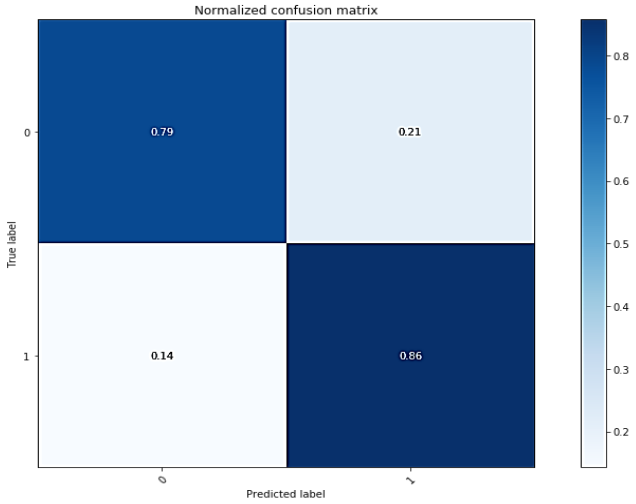


Figure 4. Confusion matrix generated for MNB with CountVectorizer

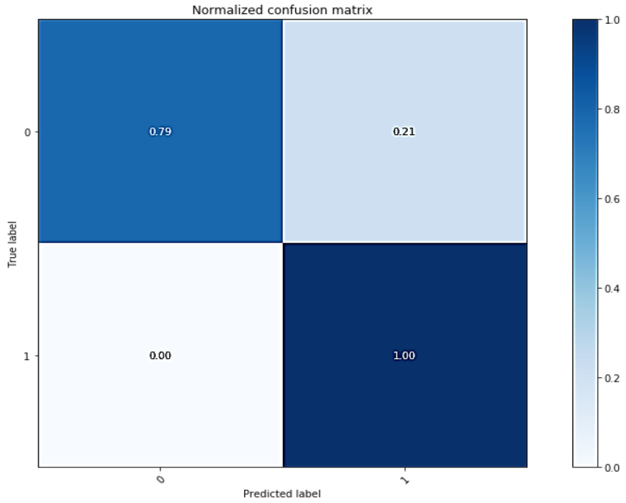


Figure 5. Confusion matrix generated for KNN with TF-IDF

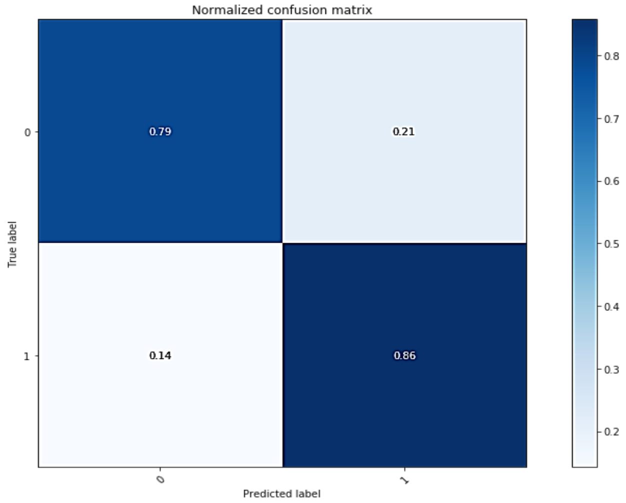


Figure 6. Confusion matrix generated for KNN with CountVectorizer

Tables 3 and 4 represent result comparison between MNB and KNN classifier based on various performance evolution parameter such as precision, recall, accuracy, and recall. Figures 7 and 8 graphical representation comparison of KNN and MNB classifier result which used CountVectorizer and TF-IDF as feature selection by considering various performance and we conclude that MNB algorithm generated more accurate result than KNN. As per Figure 9 we can represents that both algorithms are performing well using TF-IDF and CountVectorizer (CV) as feature selection, but we can say that MNB algorithm generates more accurate results with all performance parameters compare to KNN. As per Figure 9 we can represents that both algorithms are performing well using TF-IDF and CV as feature selection, but we can say that MNB algorithm generates more accurate results with all performance parameters compare to KNN.

Table 3. Order result with word-level TF-IDF

Film review dataset in the Gujarati language				
Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MNB	87.14	75.68	100	86.15
KNN	81.43	72.73	85.71	78.69

Table 4. Order result with word-level CountVectorizer

Film review dataset in the Gujarati language				
Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MNB	81.43	81.43	81.43	81.43
KNN	81.43	72.73	75.71	78.69

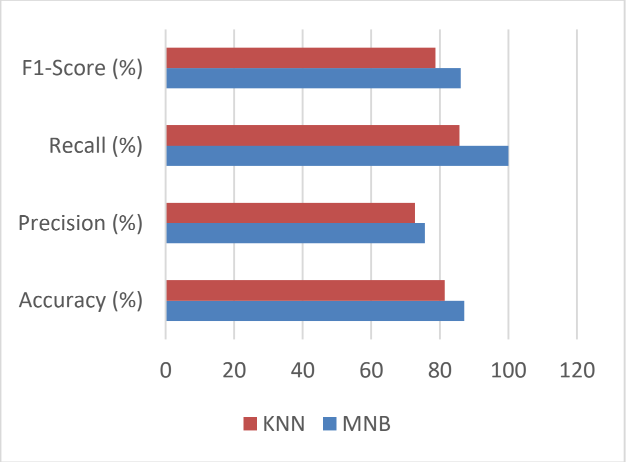


Figure 7. MNB and KNN classifier result comparison with TF-IDF

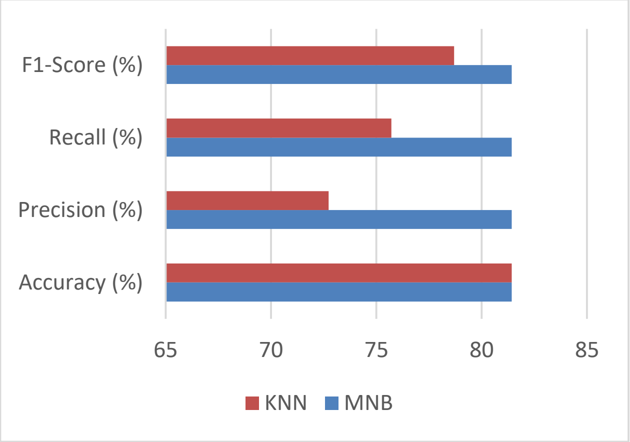


Figure 8. MNB and KNN classifier result comparison with CountVectorizer

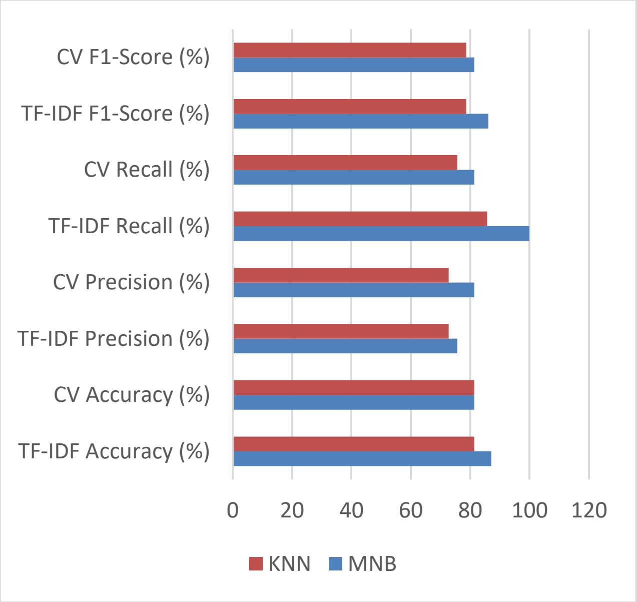


Figure 9. MNB and KNN classifier result comparison with CountVectorizer and TF-IDF

8. CONCLUSION

In this paper, the film review dataset is prepared by taking reviews in the Gujarati language, and two different machine learning-based classification techniques are applied to this data set with count and TF-IDF vectorizer elements to assess the reaction of a film review in the Gujarati language. It is concluded that TF-IDF Vectorizer features are providing improved results compared to CountVectorizer features after applying sentiment analysis. Comparing the results of two different machine learning algorithms based on Accuracy, Recall, Precision, and F-score performance parameter, we came to know MNB model forecast opinion more accurately with TF-IDF features compare to CountVectorizer features.

REFERENCES

[1] T. K. Das, D. P. Acharjya, and M. R. Patra, “Opinion mining about a product by analyzing public tweets in Twitter,” *2014 International Conference Computer Communication and Informatics (ICCCI)*, 2014, pp. 1-4, doi: 10.1109/ICCCI.2014.6921727.
[2] A. S. Kamale, P. K. Deshmukh, and P. B. Dhainje, “A survey on classification techniques for feature-sentiment analysis,” *Int. J. on Recent n Innovation Trends in Comp. and Comm.*, vol. 3, no. 7, 2015, pp. 4823-4829, doi:10.17762/ijritcc.v3i7.4744.
[3] N. Khurana, “Sentiment analysis of regional languages written in roman script on social media,” *Part of the Lecture Notes on Data Engineering and Communications Technologies book series (LNDECT)*, vol. 52, Springer, pp. 113-119, 2021.
[4] L. Gohil and D. Patel, “A sentiment analysis of Gujarati text using Gujarati SentiWordNet,” *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 9, pp. 2290-2293, 2019, doi: 10.35940/ijitee.I8443.078919.
[5] S. P. Nazare, P. S. Nar, A. S. Phate, and D. R. Ingle “Sentiment analysis in twitter,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 1, pp. 880-886, 2018.
[6] Y. Woldemariam, “Sentiment analysis in a cross-media analysis framework,” *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 2016, pp. 1-5, doi: 10.1109/ICBDA.2016.7509790.
[7] N. Spatiotis, M. Paraskevas, I. Perikos, and I. Mporas, “Examining the impact of feature selection on sentiment analysis for the Greek language,” *International Conference on Speech and Computer*, 2017.
[8] X. Fan, X. Li, F. Du, X. Li and M. Wei, “Apply word vectors for sentiment analysis of APP reviews,” *2016 3rd International Conference on Systems and Informatics (ICSAI)*, 2016, pp. 1062-1066, doi: 10.1109/ICSAI.2016.7811108.
[9] R. Ahujaa, A. Chuga, S. Kohlia, S. Guptaa, and P. Ahujaa, “The impact of features extraction on the sentiment analysis,” *Procedia Computer Science*, vol. 152, pp. 341-348, 2019, doi: 10.1016/j.procs.2019.05.008.
[10] M. Anagha, R. K. Raveena, K. Sreetha and P. C. Reghu Raj, “Fuzzy logic-based hybrid approach for sentiment analysis of malayalam movie reviews,” *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, 2015, pp. 1-4, doi: 10.1109/SPICES.2015.7091512.
[11] S. J. Arunselvan, M. A. Kumar, and K. P. Soman, “Sentiment analysis of Tamil movie reviews via feature frequency count,” *International Journal of Applied Engineering Research*, vol. 10, no. 20, pp. 17934-17939, 2015.
[12] A. Kumar, S. Kohail, A. Ekbal, and C. Biemann, “IIT-TUDA: System for sentiment analysis in Indian languages using lexical acquisition,” *Springer*, Cham, 2015.
[13] S. Jain and S. Batra, “Cross-lingual sentiment analysis using modified BRAE,” *Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 159-168.
[14] V. Jha, N. Manjunath, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, “HOMS: Hindi opinion mining system,” *IEEE 2nd Int. Conference on Recent Trends in Information Systems (ReTIS)*, 2015, pp. 366-371, doi: 10.1109/ReTIS.2015.7232906.
[15] Md S. Akhtar, A. Ekbal, and P. Bhattacharyya, “Aspect based sentiment analysis in Hindi: Resource creation and evaluation,” *Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 2703-2709.

[16] Mohammed, M. A. Ansari, and S. Govilkar, “Sentiment analysis of transliterated Hindi and Marathi script,” *Sixth International Conference on Computational Intelligence and Information*, 2016, pp. 142-149.

[17] P. Pandey and S. Govilkar, “A framework for sentiment analysis in Hindi using HSWN,” *International Journal of Computer Applications*, vol. 119, no. 19, 2015.

[18] P. Sharma and T. Moh, “Prediction of Indian election using sentiment analysis on Hindi Twitter,” *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1966-1971, doi: 10.1109/BigData.2016.7840818.

[19] S. Pundlik, P. Dasare, P. Kasbekar, A. Gawade, G. Gaikwad, and P. Pundlik, “Multiclass classification and class-based sentiment analysis for Hindi language,” *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 512-518, doi: 10.1109/ICACCI.2016.7732097.

[20] V. C. Joshi and V. M. Vekariya, “An approach to sentiment analysis on Gujarati Tweets,” *Advances in Computational Sciences and Technology*, vol. 10, no. 5, pp. 1487-1493, 2017.

[21] Y. Sharma, V. Mangat, and M. Kaur, “A practical approach to sentiment analysis of Hindi Tweets,” *1st International Conference on Next Generation Computing Technologies (NGCT2015)*, pp. 677-680, 2015.

[22] S. Subitha and S. Sujatha, “Context-based information retrieval from large heterogeneous data sources using semantics and polarity-based ranking,” *International Journal of Computer Aided Engineering and Technology (IJCAET)*, vol. 9, no. 4, 2017.

[23] J. Li and L. Qiu, “A sentiment analysis method of short texts in microblog,” *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2017, pp. 776-779, doi: 10.1109/CSE-EUC.2017.153.

[24] Y. Shi *et al.*, “Advancing innovation and development of IT and quantitative management: Preface for ITQM 2013,” *Procedia Computer Science*, vol. 17, pp. 1-7, 2013, doi: 10.1016/j.procs.2013.05.001.

[25] S. Se, R. Vinayakumar, M. A. Kumar, and K. P. Soman, “AMRITA_CEN-NLP@SAIL2015: Sentiment analysis in Indian language using regularized least square approach with randomized feature learning,” *MIKE 2015: Proceedings of the Third International Conference on Mining Intelligence and Knowledge Exploration*, vol. 9468, Dec. 2015, pp. 671–683, doi: 10.1007/978-3-319-26832-3_64.

[26] K. Denecke, “Using SentiWordNet for multilingual sentiment analysis,” *2008 IEEE 24th International Conference on Data Engineering Workshop*, 2008, pp. 507-512, doi: 10.1109/ICDEW.2008.4498370.

[27] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, “Impact of feature selection techniques for tweet sentiment classification,” In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*, 2015, pp. 299-304.

[28] R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra, “Sentiment analysis using Telugu SentiWordNet*,” Int. Conf. on Wireless Communications, Signal Processing and Networking*, 2017, pp. 666-670, doi: 10.1109/WiSPNET.2017.8299844.

[29] S. S. Mukku, N. Choudhary, and R. Mamidi, “Enhanced sentiment classification of telugu text using ML techniques,” *Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), 25th International Joint Conference onArtificial Intelligence*, New York City, USA, 2016.

[30] V. Narayanan, I. Arora, and A. Bhatia, “Fast and accurate sentiment classification using an enhanced naive Bayes model,” In *Intelligent Data Engineering and Automated Learning-IDEAL 2013*, pp. 194-201, 2013.

[31] S. Lee and J. Y. Choeh, “Predicting the help-fullness of online reviews using multilayer perceptron neural networks,” *Expert Systems with Applications*, vol. 41, no. 6, pp. 3041-3046, 2014, doi: 10.1016/j.eswa.2013.10.034.

[32] G. Sureshkumar and G. Zayaraz, “Automatic relation extraction using naïve Bayes classifier for concept relational ontology development,” *International Journal of Computer Aided Engineering and Technology (IJCAET)*, vol. 7, no. 4, 2015, doi: 10.1504/IJCAET.2015.072599.

[33] K. Machova, M. Mikula, X. Gao, and M. Mach, “Lexicon-based sentiment analysis using the particle swarm optimization,” *Electronics*, vol. 9, no. 8, 2020, doi: 10.3390/electronics9081317.

[34] K. Kour, J. Kour, and P. Singh, “Lexicon-based sentiment analysis,” *Springer*, pp. 1421-1430, 2020.

[35] Z. Turner, K. Labille, and S. Gauch, “Lexicon-based sentiment analysis for stock movement prediction,” *International Journal of Mechanical and Industrial Engineering*, vol. 14, no. 5, pp. 224-230, 2020.

[36] M. M. Fouad, T. F. Gharib, A. S. Mashat, “Efficient Twitter sentiment analysis system with feature selection and classifier ensemble,” *International conference on advanced machine learning technologies and applications*, vol. 723, 2018, pp. 516-527, doi: 10.1007/978-3-319-74690-6_51.

[37] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of sentimental reviews using machine learning techniques,” *Procedia Computer Science*, vol. 57, pp. 821-829, 2015, doi: 10.1016/j.procs.2015.07.523.