International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V70I1P236
Volume 70 Issue 1, 313-326, January, 2022

Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm

Parita Shah[1], Priya Swaminarayan[2], Maitri Patel[3], Nimisha Patel[4]

[1]*Research Scholar, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat, India.*
[1,3]*Computer Engineering Department, Gandhinagar Institute of Technology, Gujarat, India.*
[2]*Dean - Faculty of IT & CS, Parul University, Vadodara, Gujarat, India.*
[4]*Computer Engineering Department, Indus Institute of Technology & Engineering, Indus University, Gujarat, India.*

[1]*paritaponkiya@gamil.com,* [2]*priya.swaminarayan@paruluniversity.ac.in,* [3]*maitru1487288@gmail.com*

Abstract *- The study of conceptual data in an expression, that is, the assessments, evaluations, feelings, or perspectives towards a point, individual, or element, is called sentiment analysis. Expressions can be named positive, negative, or impartial. This paper authors have prepared a dataset of a movie review in Gujarati Language and introduced results generated by the proposed algorithm after performing sentiment analysis by applying different machine learning algorithms on it. The author has created numerous datasets to measure the competencies of the proposed algorithm with different machine learning classifiers. This paper describes how data are collected to create a dataset, Gujarati text pre-processing, feature selection, and classification approach is used. Minor correctness variety might take place in the challenge of applying the same model on the various dataset is likewise expressed in this paper, anyway proposed model has generated sufficient outcomes.*

Keywords - *N-gram, Feature selection, sentiment evaluation, Gujarati Language, Film Analysis, Machine classifier.*

I. INTRODUCTION

Identifying positive or negative feelings in the text is known as sentiment analysis. To perform a task such as recognizing assessment in friendly information, measuring brand notoriety, and getting clients, it is frequently utilized by organizations. The model that you build for identifying sentiments typically centres around extremity (good, negative, unbiased) yet additionally on cruciality (urgent, not urgent), goals (fascinated, not fascinated) and even on sentiments and feelings (irate, upbeat, dismal, and so on). Contingent upon how you need to decipher client criticism and inquiries, you can characterize and tailor your classifications to meet your estimation examination needs [6].

In this era where everything is available on an online platform, examination of clients input, for example, feelings in review reactions and web-based media discussions, permit brands to realize what makes clients glad or baffled is important for every business community, so they can tailor items and administrations to address their clients' issues. Thanks to web clients can easily express their considerations and emotions more transparently than any time in recent memory, feeling examination is turning into a fundamental instrument to screen and comprehend that opinion [25].

In the present scenario, life is too fast because everyone has goals and deadlines to meet; due to this, anxiety, stress and lacked motivation is increased. In this situation, entrainment plays an important role in the individual. Entrainment is an integral part of human life that helps individuals to relax, decrease anxiety, inspire motivation, and can even give you energy for real life. Many of us look forward to watching movies as they let you disconnect from your problem and allow you to feel good. To have a great time, it is important to watch a movie that is worth your time, and this will be feasible if you choose the right movie for you. Here, the importance of the sentiment analysis model comes into the picture, which will help you choose the right movie for you based on the review available.

Extensive research has been done in this field which has commonly targeted the English language. It is time to focus on Indian languages, as Indo-Aryan dialects are spoken by 78.05% of Indians, and the Dravidian dialects are spoken by 19.64% of Indians, and these are two families to which Indian language belongs. Besides the Sanskrit language, the Indian language includes other 21 languages; among all these languages, one of the important dialects is Gujarati, which belongs to the Indo-Aryan family and is the sixth most broadly communicated in the native language of Gujarat state. In recent times a high volume of information in the Gujarati language has been generated on the web, so it is essential to retrieve and analyse this information [25].

Pre-processing of data available in the Gujarati language is a challenging task due to the unattainability of resources. This paper focuses major four parts, and part one focuses on dataset preparation as there is no data set available that provides movie review in the Gujarati language, the second part focuses on pre-processing of data, i.e. removal of word, character or symbol which does not include any meaning and tokenization will generate a list of words from a given sentence/paragraph which will help to

identify features, the third part describes vector representation using TF-IDF and Count Vectorizer feature selection method by using n-gram technique, the last section explains the results are generated after applying different machine learning algorithm on it [24,26].

II. LITERATURE SURVEY

Assessment of sentiment got one of the conspicuous fields for Researchers since a decade ago. Yet very scarcely any dialects like English, Chinese, Hindi, Arabic and so forth have been significantly investigated. Due to the unavailability of lexicons and semantics, few dialects are yet neglected in this field of research.

They have used HSWN (HindiSentiWordNet), which used a Synset replacement algorithm to find the polarity of each word for sentiment analysis [27]. They have performed sentiment analysis on Hindi tweets by developing SentiWordNet, which includes adverbs and adjectives [44]. Document classification is done by the author in this paper available in the Hindi language. For classification, they have used two approaches one is based on machine learning, and another is lexicon-based classification. The machine learning approach gives an accuracy of 87.1%, which is the highest of the lexicon-based approach [17]. Aspect based sentiment analysis is performed on the dataset in the Hindi language to perform SA support vector machine and conditional random field, but the proposed system is unable to generate satisfactory results [9]. Static ontology is created with the limitation of checking max 1000 and minimum 500 words to test the documents, and multiclass classification is done to perform sentiment analysis along with HindiSentiwordnet to increase coverage of words [2]. They have used lexicon, and machine learning (NB, SVM) based approaches to classify Hindi tweets classification furthermore, they have concluded to achieve a more accurate result, Hindisetiword should be extended with a greater number of words with synonyms and antonym [30]. They have proposed an algorithm that classifies the category of given input into the area such as travel, movies, and electronics, and for this, they performed aspect-based sentiment analysis on the Hindi language [6]. Due to the lack of words, they have implemented static ontologies and used domain knowledge to perform sentiment analysis [11].

Sentiment analysis on Tamil movie review is performed by using machine learning classifiers such as SVM, NB and DT and concluded that SVM gives the highest accuracy compared to another classifier [24]. Sentiment classification is performed on Tamil movie review tweets by using feature selection method TF-IDF and domain-specific tags, but they have concluded proposed model's performance may vary due to the lack of words in a particular domain [26]. They have targeted Tamil and Bengali language for sentiment identification; for the proposed system, they have used naive Bayes classifier and C4.5 decision tree classifier and dataset size, unprocessed text may result in performance variolation of the system [10].

They have used a hybrid approach which includes Tnt Tagger and a machine learning algorithm for sentiment identification of Malayalam movie reviews [4]. A rule-based classification approach is used to calculate the polarity of a document that contains Malayalam tweets [22].

The proposed technique used to foster SentiWordNet depends on the quantitative examination of the shines related to synsets and on the utilization of the subsequent vectorial term portrayals for semi-administered synset order. The scoring trio is inferred by joining the outcomes created by a board of trustees of eight ternary classifiers, all portrayed by comparable precision levels yet unique order conduct. Authors present the consequences of assessing the precision of the naturally appointed trios on a bar likely accessible benchmark. SentiWordNet is uninhibitedly accessible for research purposes and is blessed with a Web-based graphical UI [1].

This paper offers trial results utilizing a nature-enlivened calculation—molecule swarm improvement—for marking. This improvement technique more than once names all words in a dictionary and assesses the viability of assessment order utilizing the vocabulary until the ideal names for words in the dictionary are found. The subsequent issue is that the assessment order of writings which do not contain words from the vocabulary cannot be effectively done utilizing the dictionary-based methodology. Accordingly, an assistant methodology, considering an AI strategy, is incorporated into the technique. This half and half methodology can group over 99% of writings and accomplishes preferable outcomes over the first dictionary-based methodology. The last crossbreed model can be utilized for feeling investigation in human-robot cooperation. [28].

They have collected only 40 Gujarati tweets and used POS tagging for feature extraction to perform sentiment analysis using a support vector machine [23]. Using the Indoword interface, they have created GujaratiSentoword to perform sentiment analysis using a lexicon-based approach [29]. In machine interpretation, information is gathered from different microblogging websites and changed over to the Gujarati Language to figure notions communicated Gujarati mixed with English generally this methodology is known as code-blend approach [31].

Table 1. Analysis of methods available for identifying sentiments of Indian Language.

References	Language	Method used for classification	Feature selection method	Dataset type	Accuracy
[29]	**Gujarati**	Synset Replacement Algorithm (Guj SentoWordNet), WordNet, Bag-of words	Unigram	Tweets	52.72%
[31]		Neural network	Not specified	Data collected from microblogging site and converted to Gujarati	Not specified
[23]		Support vector Machine	N-grams & POS	Normal total 40 Tweets	92%
[2]	**Hindi**	Machine Translation	TF-IDF	Movie Reviews	65.96%
		Hindisentiwordnet	Unigrams		60.31%
		Support vector Machine	TF-IDF		78.14%
[5]		Hindisentiwordnet	Unigrams	Movie Reviews	80.21%
[9]		Naive Bayes	Unigrams, Bigrams	Movie Reviews	87.1%
		Unsupervised	POS		
[10]		Decision tree C4.5 algorithm	Not specified	SAIL-2015	40.47%
[11]		Support vector Machine & J48 Decision tree	TF-IDF	SAIL-2015	42.83%(SVM)
[17]		Lexicon based	Unigrams	Movie Reviews	70%
[12]	**Hindi, Bengali**	Support vector Machine	N-grams with POS	SAIl- 2015	49.68%(Hindi), 43.20%(Bengali)
[13]		Multinomial naive Bayes	Unigrams, Bigrams, Trigrams	SAIL-2015	48.82%(Hindi), 40.40%(Bengali)
[14]	**Hindi, Tamil, Bengali**	Naive Bayes	POS using SentiWord- Net	SAIL-2015	56.67%(Hindi), 39.28%(Tamil), 33.6%(Bengali)
[18]	**Bengali**	Naive Bayes, Support vector machine, K nearest neighbour, decision tree, random forest	Unigrams, Bigrams, Trigrams	Bengali Horoscope	98.7% (SVM)
[19]	**Tamil**	Naive Bayes, Support vector machine, decision tree, Maximum entropy	POS using SentiWord- Net	Movie Reviews	75.9%(SVM)
[20]	**Konkani**	Lexicon based	POS using SentiWord- Net	Not specified	Not specified
[7]	**Malayalam**	Rule based; Lexicon based	Unigrams	Movie Reviews	85%
[15]		Support vector machine	Unigrams	Movie Reviews	91%
[23]		Dictionary Based	Unigrams	Movie Reviews	87.5%
[8]	**Punjabi**	Naive Bayes	N-grams	Blogs and News Papers	Not specified
[22]		decision tree	POS	Movie Reviews	Not specified
[16]	**Kannada**	Decision Tree (ID3)	TF-IDF	Kannada movie review	79%

III. DATASET USED

The author has created five different datasets to measure the accuracy of the different machine learning algorithms. To prepare a dataset Author has collected movie reviews from three different websites called https://gujarati.webdunia.com/movie-review,https://www.bollywoodhungama.com, https://www.filmfare.com/reviews, and https://timesofindia.indiatimes.com/entertainment/movie-reviews; collected reviews are in the English language that is translated into the Gujarati Language, and the author has performed the automated translation of reviews by creating a script. From the remaining two datasets,s one is created by collecting movie reviews from a website called https://gujarati.webdunia.com/movie-review, which provides a review of the movie in the Gujarati Language, so in this case, no machine translation approach is used. The last remaining dataset is created by collecting movie reviews manually from users who have provided reviews in the Gujarati language.

Collected data are labelled with 0 and 1 according to the ratings given on the website, 0 represents Negative, and 1 represents positive, as shown in below fig 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 છશાન ખટ...	0
10	દલિત પરિવારની વાર્તા આયં મણિ નવાઝુદ્દીન સિદ્...	0
11	વાર્તા કાજલ સ્વતંત્ર જીવન જીવવા ઇચ્છતી એક યુવ...	1

Fig 1. Dataset after text pre-processing.

IV. DATA PRE-PROCESSING

This process flows into two steps.

A. Data cleaning.

Words, characters, symbols that do not play an important role in identifying a sentiment are removed from text in this step.

B. Tokenization

Paragraphs are broken into sentences, and sentences are turned into words is called tokenization. In this step, cleaned text pass as input which converts it into tokens of words. Fig 2. shows how pre-processing of text is performed.

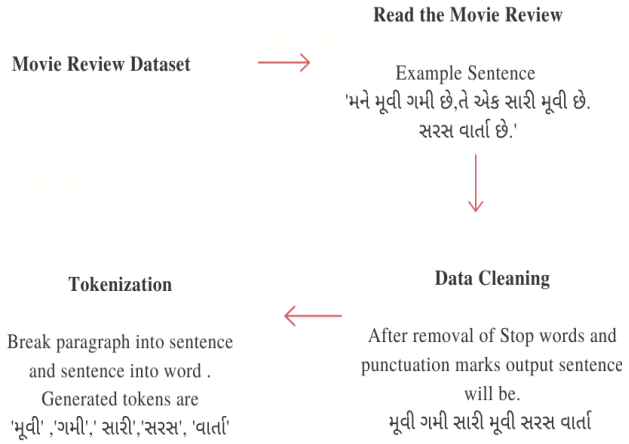

Fig 2. Gujarati text pre-processing steps.

V. FEATURE SELECTION

Feature determination is the way toward decreasing the number of info factors when fostering a prescient model. It is alluring to lessen the number of information factors to both diminish the computational expense of demonstrating and, sometimes, to improve the exhibition of the model.

In the proposed system Author have chosen TF-IDF and Count vectorizer as feature selection. To understand the working of this technique, consider fig 3., given below.

```
In [36]: sentences = ['મને મૂવી ગમી છે.',
                     'તે એક સારી મૂવી છે.' ,
                     'સરસ વાર્તા છે.',
                     'દુર્ભાગ્યે કંટાળાજનક અંત.',
                     'કંટાળાજનક મૂવી છે.'

                    ]
```

Fig 3. List of sentences given as input to TF-IDF and Count vectorizer.

A. TF-IDF

For TF-IDF calculation in the proposed code, the author has utilized TfidfVectorizer () work accessible in the sci-kit-learn library [13]. It is utilized to change an assortment of underdone records over to a lattice of TF-IDF highlights. The point of utilizing TF-IDF as opposed to crude frequencies conditions of a token in each report is proportional down the impact of tokens that frequently happen in each substance and that are henceforth noticed less useful than the angle that happens in a little part of the preparation corpus.

Evaluation procedure term event highlights we gauge the significance of a word in each chronicle. The repeat of term occasion is resolved as the events a term appears in a report segment by the word event in the document. Invert report pace of repeat moreover discovers the meaning of the term. IDF is resolved as the number of records secluded by the number of reports containing the term t [18]. For example, there are 400 words in the record, and out of those 20 words are by and large ceaseless, then term

repeat will be 20/400 =0.05 and expect there are 8000 reports and out of those 200 files contains specific terms than IDF = 8000/200=40. TF will be 0.05 * 100 =5 and IDF will be 40 [26].

Consider statements given in fig. 3 on which TF-IDF is applied, and generated result is shown in fig. 4.

	અંત	કંટાળાજનક	ગમી	દુર્ભાગ્યે	મૂવી	વાર્તા	સરસ	સારી
0	0.000000	0.000000	0.830881	0.000000	0.556451	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.000000	0.000000	0.556451	0.000000	0.000000	0.830881
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.707107	0.707107	0.000000
3	0.614189	0.495524	0.000000	0.614189	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.769447	0.000000	0.000000	0.638711	0.000000	0.000000	0.000000

Fig 4. Feature matrix generated after applying TF-IDF.

B. Counter Vectorizer

CountVectorizer is an incredible apparatus given by the sci-kit-learn library in Python. It is utilized to change a given book into a vector-based on the recurrence of each word that happens in the whole content [26]. This is useful when Author have different such messages, and the Author wishes to change over each word in every content into vectors for utilizing in the additional content investigation. CountVectorizer makes a network where every extraordinary word is addressed by a section of the lattice, and every content example from the record is a column in the framework. The worth of every cell is only the inclusion of the word in that specific content example.

Consider statements given in fig. 3 on which Countvectorizer is applied, and generated result is shown in fig. 5.

Out[37]:

	અંત	કંટાળાજનક	ગમી	દુર્ભાગ્યે	મૂવી	વાર્તા	સરસ	સારી
0	0	0	1	0	1	0	0	0
1	0	0	0	0	1	0	0	1
2	0	0	0	0	0	1	1	0
3	1	1	0	1	0	0	0	0
4	0	1	0	0	1	0	0	0

Fig 5. Feature matrix generated after applying Counvectorizer.

VI. MACHINE LEARNING ALGORITHM

A. Multinomial Naïve Bayes

Estimation dependent on the likelihood of contingent freedom between each pair of highlights is called Bayes' hypothesis, and the MNB classifier follows the guideline of Bayes' hypothesis. Think about eq. (1): which expresses that a given component should be marked for all conceivable named results by ascertaining likelihood dependent on Bayes' hypothesis [9].

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

B. K-nearest neighbour

KNN follows the guideline of similitude by ascertaining distance (Euclidean distance) between focuses. To figure distance first, it makes limit for order. After that, it will attempt to anticipate information focuses that are nearest to that limit line [26]. Euclidean distance is determined as expressed in eq. (2):

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

C. Random Forest

It is a directed learning calculation and can be utilized for both grouping and relapse reasons. It is yet the most flexible and simple to utilize calculation. A forest is comprised of trees. It has been said that the numerous trees it has, the better the timberland produces choice tree on arbitrarily picked informational collections, makes forecast from each tree, and picking the best methodology by methods for a vote. It additionally makes a decent way from of the essentialness of the capacity [3].

D. Support Vector Machine

The objective of this calculation is to discover a hyperplane that independently groups the information focuses on N-dimensional space (N-number of attributes). There are a few potential hyperplanes that could be chosen to recognize the two gatherings of information focuses. Our point is to locate a plane that has the most elevated edge, for example, the most noteworthy separation between the two classes' information focuses. Expanding the hole from the edge offers some help with the goal that further certainty can be arranged in expected information focuses [24].

E. Logistic Regression

If the dependent target value is present, then this classier is used. For instance, email identification into categorised as not spam (0) and spam email (1) [26].

VII. EVALUATION PARAMETERS

A. Accuracy

The most common extent of progress is exactness, and it is only the degree of precision expected insight to amount to discernments as showed up in eq. (3): its extent of real sure and phony negative in the occasion [24].

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

B. Precision

Offer idealistic perspectives to the total positive insights expected. The low phony positive rate proposes high precision [24].

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

C. Recall

Estimation of the number of positive veritable depict by standard through stamping it as productive (genuine positive) is called recall [24].

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

D. F1-score

It is a weighted harmony among Recall and Precision. This examines both phony positives and phony negatives. It is not as clear instinctually as precision, yet F1 is generally useful when you have an unbalanced scattering of classes. Accuracy functions admirably if there are comparable costs for sham positives and false negatives. If the expense of phony positive and phony negatives is somewhat exceptional, both Precision and Recall are easier to look at [24].

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

VIII. PROPOSED APPROACH

Fig. 6 shows the engineering and information stream model of the proposed work. It is isolated into the following stages.

Stage 1: Dataset Preparation

Stage 2: Pre-Processing

Stage 3: Feature Extraction

Stage 4: Classification

Stage 5: Performance evaluation

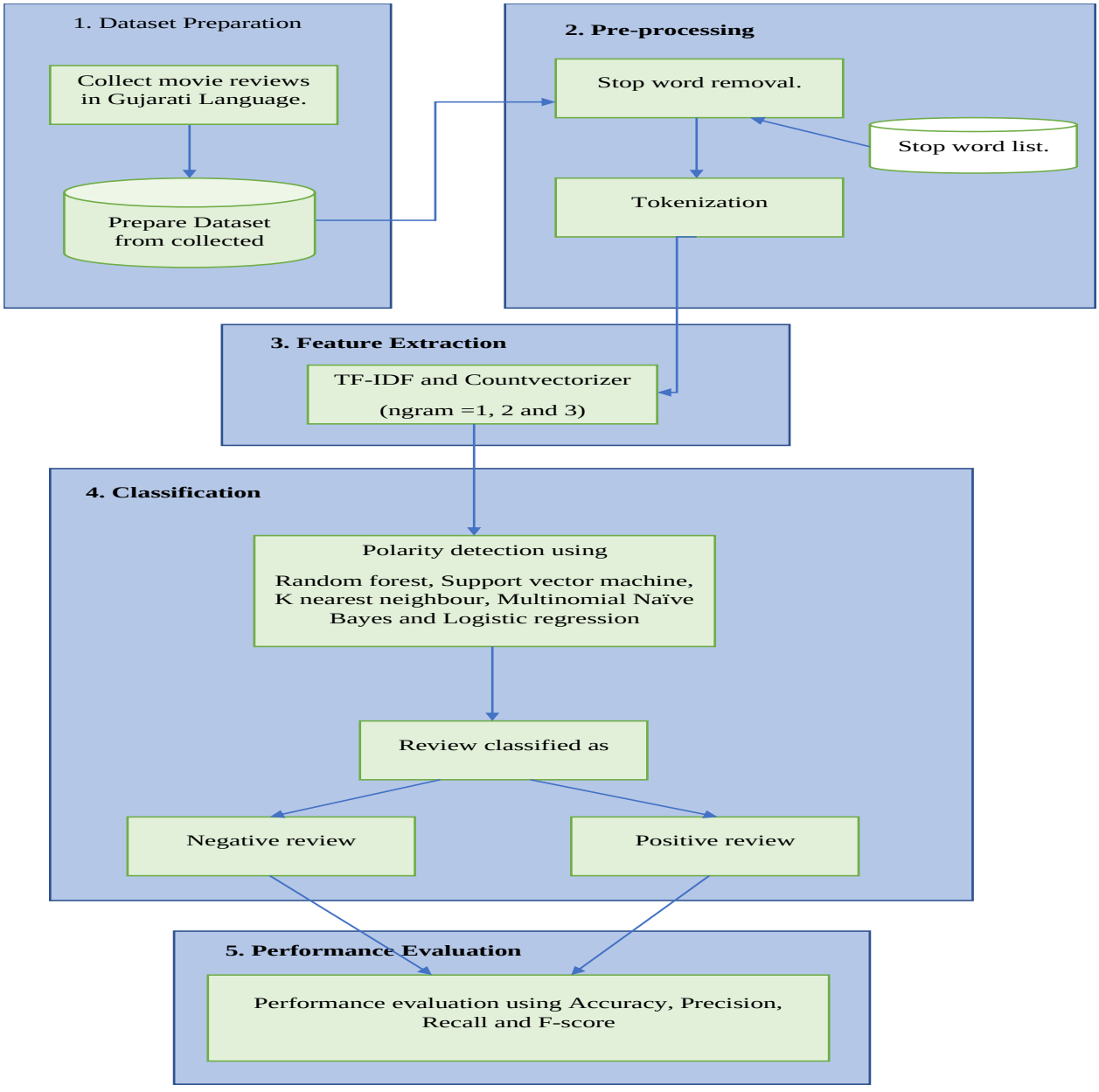

Fig 6. the detailed proposed approach used.

A. Dataset Preparation

The author has created five different datasets to measure the competencies of the proposed algorithm with different machine learning classifiers. One data set is created manually by taking reviews from different users, and other four datasets are created by collecting movie reviews from https://gujarati.webdunia.com/movie-review,https://www.bollywoodhungama.com, https://www.filmfare.com/reviews, and https://timesofindia.indiatimes.com/entertainment/movie-reviews collected reviews are in the English language that are translated into Gujarati Language, reviews that are collected from website https://gujarati.webdunia.com/movie-review is not translated into Gujarati as the reviews are available on this website is already in the Gujarati Language. The audit appraisals depend on a 1-5 scale. By and large, every film audit is 30 to 40 sentences in length. A film with in excess of 3 ratings is considered as positive, and under 3 is considered as negative. A film with a rating of 3 is expected as nonpartisan and disposed of. The corpus is implicit in a comparative way as [33] into positive and negative classes. These audits are not arbitrarily chosen; these are gathered as it is accessible. Table 2 represents the name of the source and the total review collected from that source.

Table 2. Collection of a film review to create the dataset.

no	Reviews collected from	Total reviews
1	https://gujarati.webdunia.com/movie-review	232
2	https://www.bollywoodhungama.com	592
3	https://www.filmfare.com/reviews	396
4	https://timesofindia.indiatimes.com/entertainment/movie-reviews	572
5	Collected manually from users	293

B. Gujarati text Pre-Processing

a) Stop word list creation.

Stop words are continuous, uniformly conveyed work words in any report corpus, which does not add any importance to the content substance. Data recovery from the corpus is not getting influenced by the expulsion of these words. It has been demonstrated that eliminating the stop words lessens the report size to a significant degree and saves time in text preparation [34] in Natural Language Processing. In this research, the author has created a stop word list manually with around 300 words as it was not readily available. Some examples of stop words in Gujarati are represented in fig 7.

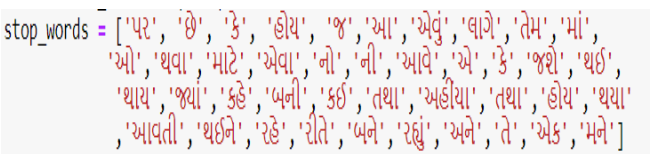

Fig 7. Example of some stop words in the Gujarati Language.

b) Removal of stop words and special characters.

Stop words and accentuation expulsion is fundamental from gathered information. Information is gathered from the web, so it might contain undesirable characters and word that is not significant for distinguishing the extremity of the word. Evacuation of undesirable images, words or characters will limit the length of the archive without bargaining extremity of feelings in this manner improves result and reduction time needed to handle the information. The author has arranged a stop word and accentuation list for the expulsion of stop words and accentuation from given information as demonstrated in the beneath model. Consider fig 8. encompass sentence in Gujarati.

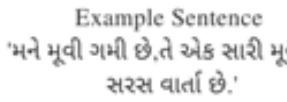

Fig 8. Example of the sentence in the Gujarati Language.

Consider the sentence given in fig 8. after the expulsion of a unique character will eliminate the characters undesirable words, and Author will get a yield sentence as stated in fig 9.

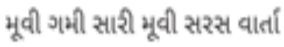

Fig 9. Sentence after removal of stop words and special characters.

c) Tokenization

Tokenization will part section into sentence and sentence into word as shown below [21].

Consider sentence stated fig 8, after tokenization string of words will be created which is nourished as a feature to classification model as shown in fig 10.

Fig 10. List of tokens generated.

C. Feature Extraction

The author has used TF-IDF and Count vectorizer as feature selection as this method will convert tokenized features into the form of vector.

D. Classification

For experiment purposes, the Author has used five different classifiers, which include Random Forest, Support vector machine, K nearest neighbour, Multinomial Naïve Bayes, and Logistic regression. 5000 most frequent features are created by us with n-grams which ranges from 1 to 3: trigrams (n=3), bigrams(n=2) and unigrams(n=1). Afterwards, these features are converted into a matrix and fed as input to machine learning classifiers which measures accuracy by creating a confusion matrix that shows the ratio of true positive, true negative, false positive and false negative.

E. Performance evaluation

Assessment of each model is done using various performance measures such as Precision, Accuracy, F1 Score and Recall.

IX. RESULTS

The proposed framework is tried for execution examination utilizing the split proportion for determination of the preparation and test sets. On normal, the framework is performing better with 70% of the informational collections preparing informational collection and 30% as testing informational index, and these outcomes have just appeared in the paper. Results additionally show critical improvement after pre-processing of the underlying surveys, which is supporting effectively notable discoveries. With five separate datasets, the author compared all five classifiers utilising unigram, bigram, and trigram features. The accuracy-based performance comparison of all five datasets with distinct five classifiers employing unigram, bigram, and trigram features with Tf-idf and n-gram (Countvectorizer) as feature selection technique is shown in Figures 11,12 and 13.

Table 3,4,5,6, and 7 indicates the result generated by the proposed algorithm on the basis of different performance evaluation criteria. Because it's mainly utilised in circumstances where there are different qualities, MNB performs better than other algorithms (for model - word includes in a text characterization issue). It mostly works with the number of considerations that each word generates. The highlights are presented in a multinomial format. In these cases, TF-IDF (Term Frequency, Inverse Document Frequency) is also beneficial. The irregular forests are an order calculation made up of a variety of different trees. When constructing each individual tree, it employs packing and element irregularity to create an uncorrelated forest of trees whose board expectation is more precise than that of any single tree. As a result, random forest beats TF-IDF when using the n-grams (bag of words) technique, providing higher accuracy. TF-IDF is preferable to Count Vectorizers because it not only focuses on the recurrence of words in the corpus but it also provides the meaning of the words. We may then exclude the words that are less important for inspection, resulting in a less difficult model structure by reducing the information aspects; as a result, LR, SVM, and KNN perform well with TF-IDF. Table 3,4, and 5 signifies the outcomes as exactness rate Accuracy of all classifiers is appeared for all datasets with Unigram, Bigram and Trigram features.

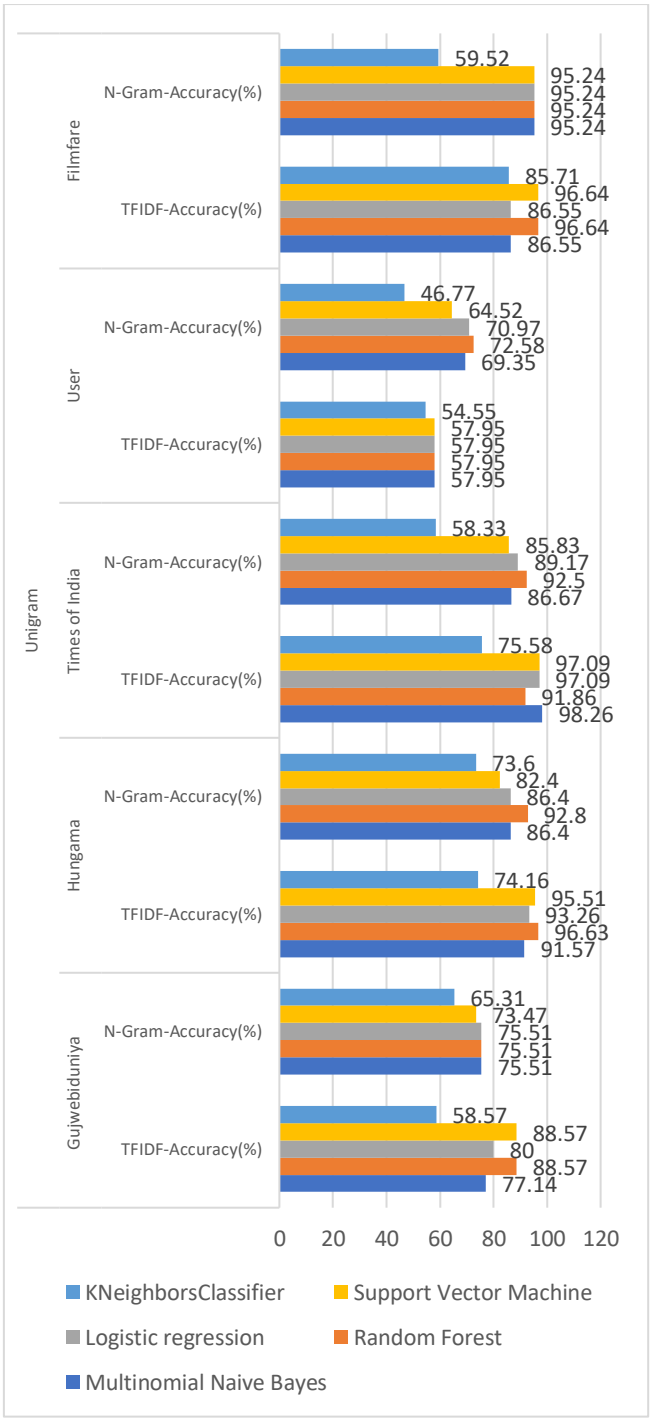

Fig 11. Dataset wise accuracy comparison of all classifiers using TF-IDF and CV with unigram Feature

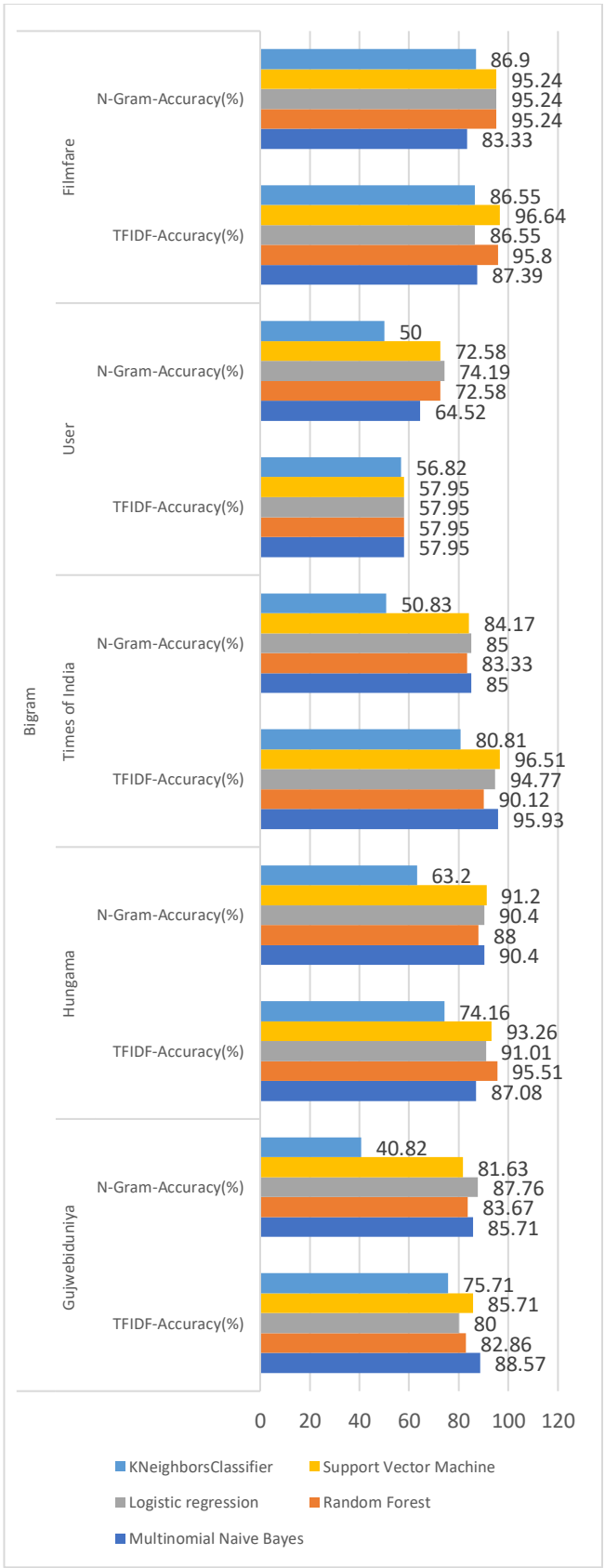

Fig 12. Dataset wise accuracy comparison of all classifiers using TF-IDF and CV with Bigram Feature

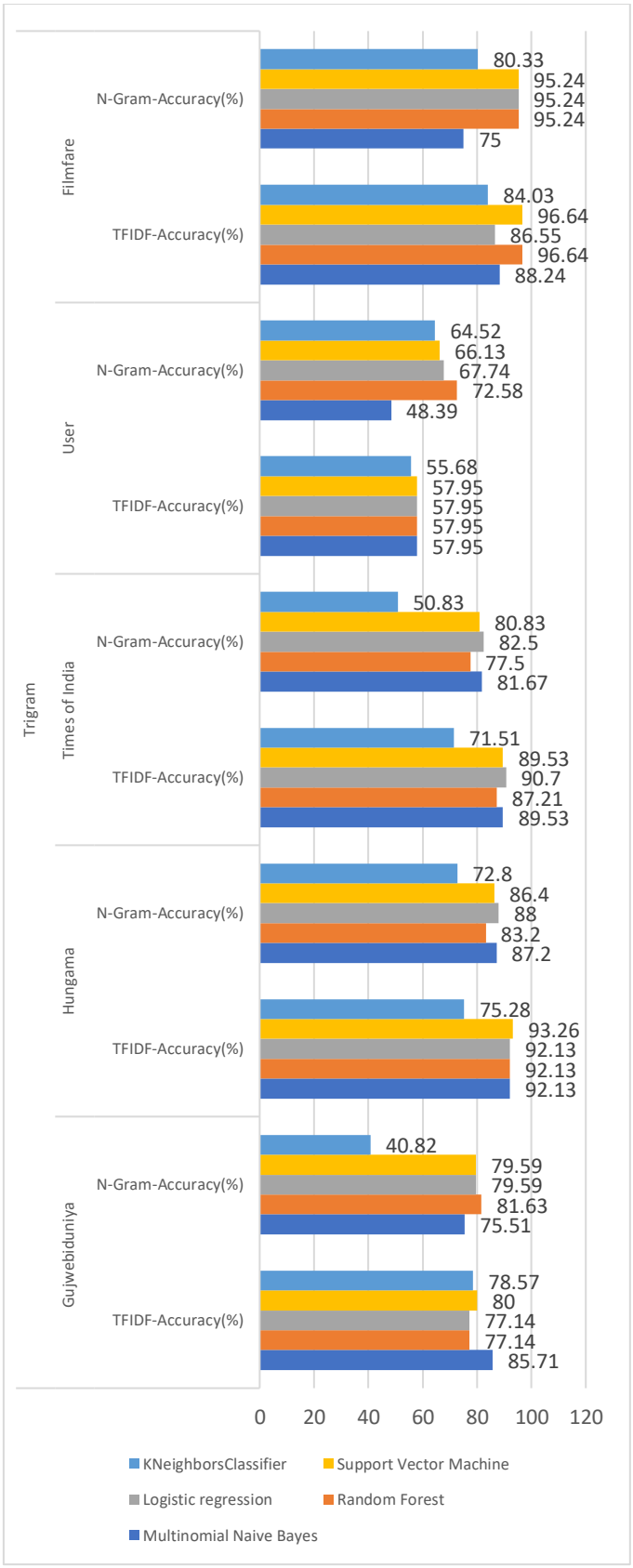

Fig 13. Dataset wise accuracy comparison of all classifiers using TF-IDF and CV with Trigram Feature

Table 3. Performance comparison of Gujwebiduniya Dataset using TF-IDF and CV as feature Selection.

Feature	Performance evaluation	MNB	RF	LR	SVM	KNN
Unigram	Accuracy (TF-IDF)	77	89	80	89	59
	Accuracy(CV)	76	76	76	73	65
	Precision (TF-IDF)	100	94	100	100	66
	Precision (CV)	67	72	76	63	54
	Recall (TF-IDF)	59	85	64	79	54
	Recall (CV)	80	65	76	85	95
	F-score(TF-IDF)	74	89	78	89	59
	F-score(CV)	73	68	76	72	69
Bigram	Accuracy (TF-IDF)	89	83	80	86	76
	Accuracy (CV)	86	84	88	82	41
	Precision (TF-IDF)	100	100	100	100	84
	Precision (CV)	76	80	88	76	41
	Recall (TF-IDF)	79	69	64	74	69
	Recall (CV)	95	80	88	80	100
	F-score (TF-IDF)	89	82	78	85	76
	F-score (CV)	84	80	88	78	58
Trigram	Accuracy (TF-IDF)	86	77	77	80	79
	Accuracy (CV)	76	82	80	80	41
	Precision (TF-IDF)	100	100	100	100	88
	Precision (CV)	63	82	80	68	41
	Recall (TF-IDF)	74	59	59	64	72
	Recall (CV)	95	70	80	95	100
	F-score (TF-IDF)	85	74	74	78	79
	F-score (CV)	76	76	80	79	58

Table 4. Performance comparison of Hungama Dataset using TF-IDF and CV as feature Selection.

Feature	Performance evaluation	MNB	RF	LR	SVM	KNN
Unigram	Accuracy (TF-IDF)	92	97	93	96	74
	Accuracy (CV)	86	93	86	82	74
	Precision (TF-IDF)	87	96	90	94	70
	Precision (CV)	90	94	86	89	80
	Recall (TF-IDF)	98	98	98	98	86
	Recall (CV)	84	93	86	76	67
	F-score (TF-IDF)	92	97	94	96	77
	F-score (CV)	87	93	86	82	73
Bigram	Accuracy (TF-IDF)	87	96	91	93	74
	Accuracy (CV)	90	88	90	91	63
	Precision (TF-IDF)	83	98	86	90	73

	Precision (CV)	91	89	90	95	60
	Recall (TF-IDF)	95	93	98	98	78
	Recall (CV)	91	88	90	88	94
	F-score (TF-IDF)	88	96	92	94	76
	F-score (CV)	91	89	90	91	73
Trigram	Accuracy (TF-IDF)	92	92	92	93	75
	Accuracy (CV)	87	83	88	86	73
	Precision (TF-IDF)	91	90	87	92	75
	Precision (CV)	93	81	88	89	66
	Recall (TF-IDF)	93	96	100	96	78
	Recall (CV)	82	90	88	85	100
	F-score (TF-IDF)	92	93	93	94	76
	F-score (CV)	87	85	88	87	80

Table 5. Performance comparison of Times of India Dataset using TF-IDF and CV as feature Selection.

Feature	Performance evaluation	MNB	RF	LR	SVM	KNN
Unigram	Accuracy (TF-IDF)	92	97	93	96	74
	Accuracy (CV)	86	93	86	82	74
	Precision (TF-IDF)	87	96	90	94	70
	Precision (CV)	90	94	86	89	80
	Recall (TF-IDF)	98	98	98	98	86
	Recall (CV)	84	93	86	76	67
	F-score (TF-IDF)	92	97	94	96	77
	F-score (CV)	87	93	86	82	73
Bigram	Accuracy (TF-IDF)	96	90	95	97	81
	Accuracy (CV)	85	83	85	84	51
	Precision (TF-IDF)	99	88	100	100	77
	Precision (CV)	89	88	85	90	75
	Recall (TF-IDF)	93	91	89	93	85
	Recall (CV)	80	76	85	76	83
	F-score (TF-IDF)	96	90	94	96	81
	F-score (CV)	84	82	85	83	79
Trigram	Accuracy (TF-IDF)	90	87	91	90	72
	Accuracy (CV)	82	78	83	81	51
	Precision (TF-IDF)	94	86	95	92	69
	Precision (CV)	89	70	83	80	65
	Recall (TF-IDF)	83	88	85	85	73
	Recall (CV)	71	95	83	81	70
	F-score (TF-IDF)	88	87	90	89	71
	F-score (CV)	79	81	83	81	66

Table 6. Performance comparison of user Dataset using TF-IDF and CV as feature Selection.

Feature	Performance evaluation	MNB	RF	LR	SVM	KNN
Unigram	Accuracy (TF-IDF)	58	58	58	58	55
	Accuracy (CV)	69	73	71	65	47
	Precision (TF-IDF)	58	58	58	58	57
	Precision (CV)	75	73	71	73	93
	Recall (TF-IDF)	100	100	100	100	90
	Recall (CV)	87	100	71	80	29
	F-score (TF-IDF)	73	73	73	73	70
	F-score (CV)	80	84	71	77	44
Bigram	Accuracy (TF-IDF)	58	58	58	58	57
	Accuracy (CV)	65	73	74	73	50
	Precision (TF-IDF)	58	58	58	58	59
	Precision (CV)	79	73	74	75	75
	Recall (TF-IDF)	100	100	100	100	86
	Recall (CV)	69	100	74	93	47
	F-score (TF-IDF)	73	73	73	73	70
	F-score (CV)	74	84	74	83	58
Trigram	Accuracy (TF-IDF)	58	58	58	58	56
	Accuracy (CV)	48	73	68	66	65
	Precision (TF-IDF)	58	58	58	58	58
	Precision (CV)	72	73	68	71	73
	Recall (TF-IDF)	100	100	100	100	84
	Recall (CV)	47	100	68	91	80
	F-score (TF-IDF)	73	73	73	73	69
	F-score (CV)	57	84	68	80	77

Table 7. Performance comparison of Filmfare Dataset using TF-IDF and CV as feature Selection.

Feature	Performance evaluation	MNB	RF	LR	SVM	KNN
Unigram	Accuracy (TF-IDF)	87	97	87	97	86
	Accuracy (CV)	95	95	95	95	60
	Precision (TF-IDF)	87	96	87	96	86
	Precision (CV)	95	95	95	95	94
	Recall (TF-IDF)	100	100	100	100	99
	Recall (CV)	100	100	95	100	60
	F-score (TF-IDF)	93	98	93	98	92
	F-score (CV)	97	97	95	97	73
Bigram	Accuracy (TF-IDF)	87	96	87	97	87
	Accuracy (CV)	83	95	95	95	87
	Precision (TF-IDF)	87	95	87	96	87
	Precision (CV)	96	95	95	95	95

	Recall (TF-IDF)	100	100	100	100	99
	Recall (CV)	86	100	95	100	91
	F-score (TF-IDF)	93	98	93	98	93
	F-score (CV)	90	97	95	97	93
Trigram	Accuracy (TF-IDF)	88	97	87	97	84
	Accuracy (CV)	75	95	95	95	8.3
	Precision (TF-IDF)	88	96	87	96	88
	Precision (CV)	98	95	95	95	85
	Recall (TF-IDF)	100	100	100	100	95
	Recall (CV)	74	100	95	100	89
	F-score (TF-IDF)	94	98	93	98	91
	F-score (CV)	84	97	95	97	90

X. CONCLUSION

Sentiment detection is an interesting but challenging task while focusing on Indian languages and trickier when you try to analyse sentiment from a language like Gujarati due unavailability of sufficient recourses. In this paper sentiment analysis model is applied to movie reviews that are prepared in the Gujarati Language. The experiment was conducted on the different datasets and showed that language-specific enhanced results were achieved. To achieve desired results, the Author has performed data pre-processing which provides a list of tokens that is helpful in the feature selection task. Feature vector generated using TF-IDF and Count Vectorizer technique is feed as input to different machine learning-based classifier which generates confusion matrix based on which accuracy of the different classifier is measured. Minor accuracy variation may occur after applying the same model on the different datasets is also stated in this paper; the however proposed model generated adequate results. In future, more reviews can be collected to analyse generated results by applying the same model on a large dataset.

REFERENCES

[1] Esuli, F. Sebastiani, SentiWordNet, a high-coverage lexical resource for opinion mining Kluwer Academic Publishers, (2007).

[2] Joshi, A. R. Balamurali, P. Bhattacharyya, A fall-back strategy for sentiment analysis in hindi, a case study, Proceedings of the 8th ICON, (2010).

[3] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM, 56 (4) (2013) 82.

[4] N. Medagoda, S. Shanmuganathan, and J. Whalley, A comparative analysis of opinion mining and sentiment classification in non-english languages, 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), (2013).

[5] N. Mittal, B. Agarwal, G. Chouhan, N. Bania, P Pareek, Sentiment Analysis of Hindi Reviews based on Negation and Discourse Relation, In Proceedings of the 11th Workshop on Asian Language Resources, (2013) 45-50.

[6] J. Kaur and J. R. Saini, 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, 4(2) (2014) 53.

[7] D. S. Nair, J. P. Jayan, R. R. R, and E. Sherly, SentiMa - Sentiment extraction for Malayalam, International Conference on Advances in Computing, Communications, and Informatics (ICACCI), (2014).

[8] Kaur and V. Gupta, N-gram Based Approach for Opinion Mining of Punjabi Text, Lecture Notes in Computer Science Multidisciplinary Trends in Artificial Intelligence, (2014) 81–88

[9] V. Jha, N. Manjunath, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, HOMS, Hindi opinion mining system, 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), (2015).

[10] S. S. Prasad, J. Kumar, D. K. Prabhakar, and S. Pal, Sentiment Classification, An Approach for Indian Language Tweets Using Decision Tree, Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, (2015) 656–663.

[11] M. Venugopalan and D. Gupta, Sentiment Classification for Hindi Tweets in a Constrained Environment Augmented Using Tweet Specific Features, Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, (2015) 664–670.

[12] Kumar, S. Kohail, A. Ekbal, and C. Biemann, IIT-TUDA, System for Sentiment Analysis in Indian Languages Using Lexical Acquisition, Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, (2015) 684–693.

[13] K. Sarkar and S. Chakraborty, A Sentiment Analysis System for Indian Language Tweets, Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, (2015) 694–702.

[14] S. Se, R. Vinayakumar, M. A. Kumar, and K. P. Soman, AMRITACEN@SAIL2015, Sentiment Analysis in Indian Languages, Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, 703–710 (2015).

[15] D. S. Nair, J. P. Jayan, R. R.r, and E. Sherly, Sentiment Analysis of Malayalam film review using machine learning techniques, International Conference on Advances in Computing, Communications, and Informatics (ICACCI), (2015).

[16] D. N. and R. K. P., Polarity detection of Kannada documents, 2015 IEEE International Advance Computing Conference (IACC), (2015)

[17] D. Mumtaz and B. Ahuja, Sentiment analysis of movie review data using Senti-lexicon algorithm, 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), (2016).

[18] T. Ghosal, S. K. Das, and S. Bhattacharjee, Sentiment analysis on (Bengali horoscope) corpus, 12th IEEE Int. Conf. Electron. Energy, Environ. Commun. Comput. Control (E3-C3), INDICON, (2015) (2016) 1–6.

[19] S. Se, R. Vinayakumar, M. A. Kumar, and K. P. Soman, Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms, Indian Journal of Science and Technology, 9(45) (2016).

[20] D. T. Miranda and M. Mascarenhas, KOP, An opinion mining system in Konkani, IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), (2016).

[21] V. Rohini, M. Thomas, and C. A. Latha, Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm, IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), (2016).

[22] M. P. Ashna and A. K. Sunny, Lexicon based sentiment analysis system for malayalam language, 2017 International Conference on Computing Methodologies and Communication (ICCMC), (2017).
[23] V. C. Joshi, V. M. Vekariya, An Approach to Sentiment Analysis on Gujarati Tweets, Advances in Computational Sciences and Technology, pp.1487-1493, 2017.
[24] Fouad M.M., Gharib T.F., Mashat A.S. (2018) Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. In, Hassanien A., Tolba M., Elhoseny M., Mostafa M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). AMLTA., Advances in Intelligent Systems and Computing, l (2018) 723.
[25] Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof. Dr. D. R. Ingle Sentiment Analysis in Twitter International Research Journal of Engineering and Technology (IRJET) e-ISSN, 2395-0056 , 05(1) (2018) 880-886.
[26] Ravinder Ahujaa, Aakarsha Chuga, Shruti Kohlia, Shaurya Guptaa, and Pratyush Ahujaa. The Impact of Features Extraction on the Sentiment Analysis, International Conference on Pervasive Computing Advances and Applications PerCAA Elsevier, (2019) 341-348.
[27] Lata Gohil, Dharmendra Patel, A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net, International Journal of Innovative Technology and Exploring Engineering (IJITEE), (2019) 2290-2293
[28] Kristina Machova, Martin Mikula, Xiaoying and Marian Mach, Lexicon-based Sentiment Analysis Using the Particle Swarm Optimization, Electronics, 2020, pp. 1-2
[29] Kavleen Kour, Jaspreet Kour, and Parminder Singh, Lexicon-Based Sentiment Analysis, Springer, (2020) 1421-1430
[30] Zane Turner, Kevin Labille, Susan Gauch, Lexicon-Based Sentiment Analysis for Stock Movement Prediction, International Journal of Mechanical and Industrial Engineering, (2020) 185-191
[31] Nisha 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) Springer, 52 (2021) 113-119.