Exploring Sentiment Analysis in Indian Regional Languages: Methods, Challenges, and Future Directions

Vedant Nemade
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

Harshada Lahane
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

Sanika Nandanwar
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

Pratiksha Karande
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

ABSTRACT

Sentiment Analysis, pivotal in natural language processing, extends its reach beyond English to Indian regional languages like Hindi, Marathi, Kannada, Konkani, Bengali, Khandeshi, and Urdu. This paper presents a comprehensive survey of 32 research papers in this domain, examining methodologies, datasets, and techniques while emphasizing the significance of sentiment analysis in diverse linguistic contexts for enhancing customer relationship management functionalities. It underscores the necessity for future research and highlights the efficacy of machine learning techniques. By elucidating on computational challenges and outlining various sentiment analysis methods, this paper serves as a critical resource for researchers and practitioners, fostering advancements in sentiment analysis tailored to regional linguistic nuances.

KEYWORDS

Bag Of Words, Hindi, Kannada, RNN, Konkani, Malayalam, Marathi, Maximum Entropy, Naive Bayes, Sentiment Analysis, SVM, TF-IDF, Urdu.

INTRODUCTION

In today's technologically advanced world, individuals are increasingly expressing their passions and sentiments on vibrant social media platforms in their native languages. The internet serves as a vast repository of data, with users utilizing platforms like Twitter, Facebook, and blogs to articulate their thoughts and emotions. As customers share their feedback through text and voice messages, businesses, including call centers, seek to analyze these expressions to enhance customer experience and improve service quality. However, manual analysis of this wealth of human sentiment proves challenging, necessitating the introduction of sentiment analysis, a Natural Language Processing (NLP) method, to automatically determine polarity as positive, negative, or neutral. India's rich tapestry of cultures and languages further complicates sentiment analysis, as datasets are often fragmented and disparate. This paper focuses on sentiment analysis in Indian regional languages, such as Hindi, Marathi, Bengali, and Urdu, acknowledging the unique linguistic nuances inherent in these diverse communities. By leveraging sentiment analysis, businesses can gain insights into customer satisfaction, social sentiment, and brand reputation, thus enabling informed decision-making and enhanced customer engagement.

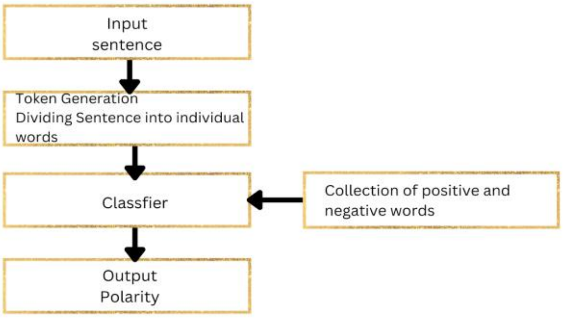

Figure 1: Working of Algorithms in Sentiment Analysis

MOTIVATION

Sentiment analysis has emerged as one of the prominent research areas. Emotions play a significant role in every individual's life. If these emotions are interpreted precisely by organizations, they can stand extremely beneficial for the firm to earn massive profits as well as customer's trust and affection. People tend to express opinions in textual and voice message formats often on various social networking sites.The major goal of this paper is to draw out the emotional core of various peoples' viewpoints for varied uses. This study will be helpful for various business organizations to boost CRM functionalities. Different Business Organizations can review the customer opinions quickly and reframe their product more precisely in order to boost customer satisfaction. Agents selling various commodities can analyze the moods of customers using sentiment analysis from voice technique and converse more effectively with customers. User satisfaction can be heightened by customizing the product by extracting sentiments from their reviews.

LANGUAGE DESCRIPTION

The languages that have been studied in this paper are described in short in this section. Apart from the following languages, languages like Konkani, Ahirani, Khandeshi, Nagpuri and Tullu were also taken into consideration for the study. Due to similarities that the above-mentioned languages have with the enlisted languages below they are not mentioned separately.

Marathi:

One of India's 14 regional languages and one of the 22 national languages is Marathi. It is one of the two official languages of Maharashtra. The words in Marathi are classified into following 3 classes:

Positive	सुंदर, परिपूर्ण, आकर्षक
Negative	वाईट, राग, गर्विष्ठ
Neutral	आमचे, मध्ये, नंतर

Table 1: Sentiment Words in Marathi

Hindi:

Hindi, the national language of India is spoken in 7 more countries as mother tongue. It is the third most spoken language in the world. It has a resemblance with Sanskrit.

Positive	सुंदर, उत्तम, आकर्षक
Negative	बुरा, क्रोध, घमंडी
Neutral	हमारी, में, बाद में

Table 2: Sentiment Words in Hindi

Bengali:

About 200 million people speak Bengali, also known as Bangla, which is mostly spoken in Bangladesh and in the Indian states of West Bengal and Tripura. One of the top ten most spoken languages in the world, Bengali is the second most widely spoken language on the Indian subcontinent after Hindi.

Positive	সুন্দর, নিখুঁত,আকর্ষণীয়
Negative	খারাপ,রাগ,অহংকারী
Neutral	আমাদের,ভিতরে, পরে

Table 3: Sentiment Words in Bengali

Kannada:

The Dravidian language family includes the Kannada language which serves as the state of Karnataka's official tongue. Kannada was estimated to be the first language of 38 million people according to census statistics from the early twenty-first century; another 9 to 10 million people were estimated to speak it as a second language.

Positive	ಸುಂದರ,ಪರಿಪೂರ್ಣ,ಆಕರ್ಷಕ
Negative	ಕೆಟ್ಟ, ಕೋಪ,ದಾರ್ಷ್ಟ್ಯ
Neutral	ನಮ್ಮ, ಒಳಗೆ,ನಂತರ

Table 4: Sentiment Words in Kannada

Malayalam:

Malayalam is mostly spoken in India, where Kerala and the union territory of Lakshadweep have made it their official languages. Additionally, bilingual populations in nearby regions of Tamil Nadu and Karnataka speak it. More than 35 million people spoke Malayalam at the beginning of the twenty-first century.

Positive	മനോഹരം,തികഞ്ഞ,ആകർഷകമായ
Negative	മോശം, കോപം,അഹങ്കാരി
Neutral	ഞങ്ങളുടെ, ഇൻ, പിന്നീട്

Table 5: Sentiment Words in Malayalam

Urdu:

About a thousand years ago, in the Delhi region of north India, Urdu began to take shape. It was based on the dialect used in the Delhi area and greatly impacted by Turkish, Arabic, and Persian as well.

Positive	کامل,خوبصورت پرکشش,
Negative	برا,غصہ,مغرور
Neutral	ہمارے,میں,بعد میں

Table 6: Sentiment Words in Urdu

LITERATURE REVIEW

Machine learning algorithms and deep learning models have been extensively applied to sentiment analysis in Indian languages, including Bengali, Hindi, Tamil, Kannada, and Urdu. Techniques such as Support Vector Machine, Naïve Bayes, and Maximum Entropy have been used to classify text into positive, negative, and neutral sentiments. Deep learning approaches, notably CNN-LSTM networks and the BERT model, have shown promise in capturing the contextual nuances of language, proving particularly effective in handling code-mixed text and languages with complex morphological structures. These studies have addressed challenges such as informal language, slang, and emoticon usage typical of social media text, with BERT models demonstrating superior performance due to their ability to capture contextual information effectively. The exploration of these models across various languages and contexts emphasizes the dynamic nature of sentiment analysis research and its potential for future advancements [1, 3, 4, 5, 7, 9, 10, 11, 13, 14, 15, 20, 22, 23, 24, 30].For languages with limited computational resources, such as Marathi and Urdu, lexicon-based approaches have been proposed as effective methods for sentiment analysis. Researchers have developed lexicons that include lists of positive and negative words, assigning polarity values to facilitate the classification of sentences into sentiments. This method has been applied to tackle various challenges in sentiment analysis, including named entity recognition, sarcasm detection, negation handling, and aspect-based analysis. The development and application of these lexicons have proven crucial in languages where machine learning resources may be scarce, highlighting the importance of linguistic and lexical resources in sentiment analysis [2, 6, 8, 12, 16, 17, 18, 19, 21, 25, 26, 27, 28, 29, 31, 32].The phenomenon of code-mixing, where multiple languages are integrated within a single text, presents significant challenges for traditional NLP systems. This is particularly relevant in the Indian context, where bilingual or trilingual code-mixing is common. Studies focusing on sentiment analysis in code-mixed languages, especially Dravidian languages like Kannada, have emphasized the need for models capable of effectively analyzing sentiment in such complex datasets. The use of advanced models like BERT has been advocated for their ability to navigate the intricacies of code-mixed text, highlighting the importance of developing tailored approaches for multilingual sentiment analysis [5, 15].

The survey highlights several challenges facing sentiment analysis in Indian languages, including the scarcity of specialized NLP tools, the complexity of regional languages, and the processing of code-mixed text. Future work is suggested to focus on expanding the classification capabilities to include figurative language, enriching datasets with more diverse samples, exploring additional algorithms for enhanced accuracy, and further developing sentiment analysis models to accommodate low-resource languages. These directions underscore the evolving nature of sentiment analysis research and its critical role in understanding and leveraging user-generated content in multilingual societies [1, 2, 3, 4, 5, 7, 9, 10, 11, 13, 14, 15, 20, 22, 23, 24, 30, 31, 32].

Studies focused on Hindi highlight the use of machine learning techniques, such as Naïve Bayes, and deep learning models, including BERT, for sentiment analysis on social media and news content. Challenges noted include handling informal language and the scarcity of Hindi-specific NLP tools. Research also explores the creation of Hindi lexicons for lexicon-based sentiment analysis, emphasizing the importance of accurately capturing sentiments in a language with significant online user-generated content [1, 2, 3, 9, 10, 20, 28, 32].In the Marathi context, researchers propose lexicon-based approaches to address the lack of computational resources for sentiment analysis. The development of sentiment lexicons, comprising positive and negative words, is suggested for classifying sentences. Challenges such as sarcasm detection and negation handling are highlighted, alongside future directions aimed at dataset enrichment and algorithm exploration [2, 16, 17].Bengali, the research focuses on machine learning algorithms and transformer models like multilingual BERT and XLM-RoBERTa, fine-tuned on Bengali datasets. Despite the challenges associated with low-resource languages, including the scarcity of training corpora, these studies showcase the potential of advanced NLP models in sentiment analysis for Bengali, setting new benchmarks and encouraging further exploration [11, 13, 17, 19].Sentiment analysis in Dravidian languages, particularly code-mixed text involving Kannada, Tamil, and Malayalam, employs deep learning models such as BERT to tackle the intricacies of analyzing sentiments in multilingual and code-mixed contexts. The challenges highlighted include English-based phonetic typing and word-level mixing, underscoring the need for models capable of effectively processing code-mixed data. These studies contribute to understanding sentiment trends among speakers of Dravidian languages on social media platforms [5, 7, 13, 14, 15].Research in Urdu sentiment analysis underscores the complexity of the language's morphological structure and the absence of standard datasets or corpora. Efforts to create benchmark datasets and evaluate the performance of machine learning and deep learning techniques, including CNN-1D and LSTM, are noted. The studies aim to enrich the domain of sentiment analysis in Urdu and open new avenues for processing resource-deprived languages [29, 30, 31].A few studies address sentiment analysis across multiple Indian languages, tackling the challenges of code-mixing and linguistic

diversity. These research efforts utilize advanced models like BERT and emphasize the importance of developing techniques that are adaptable to the unique linguistic features of Indian languages. The focus on multilingual and code-mixed sentiment analysis reflects a broader trend towards enhancing NLP applications for diverse linguistic landscapes [4, 5, 15]

Table 8: Summary of Literature Review

Ref	Dataset	Method	Accuracy	Language

[1]	Tweets using TWITTER4J	Decision Tree, Naïve Bayes, Multinomial Naïve Bayes and Support Vector Machines	Bengali=43.2 %
			Hindi =55.67 %
			Tamil=39.28 %
				Bengali
				Hindi
				Tamil

[3]	Twitter data	NLP Textblob	98%	English, Kannada, Telugu, Hindi, Tamil, Malayalam
[5]	Code-Mixed Kannada comments	BERT model Multinomial Naïve Bayes Logistic Regression Random Forest XGBoost Decision Tree Support Vector Machine	F1-score=0.66 0.40 0.47 0.45 0.40 0.42 0.50	code-mixed Kannada
[6]	Comments from Youtube Videos	1)Multinomial Naive Bayes 2)Logistic Regression 3)Support Vector Machine 4)k Nearest Neighbor 5)Decision Tree 6)Random Forest 7)Multi-Layer Perceptron 8)Cross Validation	F1-scores 1)0.41 2)0.55 3)0.57 4)0.60 5)0.52 6)0.55 7)0.60 8)0.62	Tulu corpus
[7]	Twitter posts and google translator English 1,600,000 Hindi 2500 Kannada 6300	Convolution neural network	English 95% Hindi 99% Kannada 99%	English Hindi Kannada
[8]	12974 Tweets about current news	R language and R graphical tools	Positive 31% Negative 20% Neutral 49%	English
[9]	42,235 Hindi Tweets	1)Dictionary Based, 2)Naive Bayes 3)SVM	1)62.1% 2)34% 3)78.4%	Hindi

[10]	Movie review Websites	Decision tree-Classifier algorithm	1)Precision for English=0.86 2)Precision for Kannada=0.78	English Kannada
[11]		Multinomial Naive Bayes Logistic Regression Decision Tree Random Forest	Bengali, 2-class NB classifier (67.83% Bengali, 3-class LR classifier (51.25%) Hindi, 2-class LR classifier (81.57%) Hindi, 3-class LR classifier (56.96%). Tamil, 2-class NB classifier (62.16%) Tamil, 3-class RF classifier (45.24%)	Bengali, Hindi, and Tamil.

[12]	Posts from social media platforms	1)MBERT model	Tamil
			F1-score=0.603
			Kannada
			F1-score=0.595
			Malayalam
			F1-score=0.698
				Tamil-English
				Kannada-English
				malayalam-English

[13]	comments from social media websites for both Tamil and Malayalam	Word Embedding model
		TF-IDF
		Logistic Regression
			LR
			Tamil=0.62
			Malayalam =0.68
			Word Embedding model
			Tamil=0.65
			Malayalam =0.63
				Tamil
				Malayalam

[14]	Wordnet lexical relation data	Synset Projection Approach, Support Vector Machine	0.475 0.5114 - F-score measure	Hindi, Konakni
[15]	100,000 PoS tagged Konkani sentences	Hidden Markov Model,	76.75% for training data, 71.29 for testing data	Konkani

[16]	English SentiWord Net translated to Marathi	Corpus-based approach		Marathi
[17]	100 positive and 100 negative words	SVM, Naive Bayes and Maximum Entropy		Marathi
[18]	Prothom Alo YouTube-B Book-B	BERT, XML-RoBERTa combined with GRU,LStM & CNN	BERT & CNN=91% and 92% for II & III datasets.	Bengali

[19]	L3CubeMahaSent	CNN,LSTM,ULMFiT and BERT	CNN = 93.13
			BiLSTM = 92.67
				Marathi

[20]	63,000 Facebook Comments	SVM, Random Forest, KNN, Naive Bayes & Neural N/w	RF= 58%, KNN= 55%,NB=52%,NN= 50% & SVM=62%	Bengali

[21]	HASOC 2021	CNN, LSTM, BERT, IndicBERT, RoBERTa	IndicBERT=88%
			mBERT=86%
			RoBERTa=87%
				Hindi & Marathi

[22]	Bengali Facebook Comments And Posts	Naive Bayes, Decision Tree, Random Forest,SVM,AdaBoost, LSTM,CNN	Classical Methods = 70% Deep Learning Methods= 96.95	Bengali
[23]	Cricket Reviews in Bangla Language	CNN,RNN,LSTM,Bi-LSTM,GRU,etc.	Bi-Gram & SVM= 82.42 LSTM = 46%	Bengali
[24]	Book Reviews Dataset	MNB,Random forest, Decision Tree, KNN, SVM	Multinomial Naive Bayes= 87%	Bengali
[25]	HSWN(Hindi SentiwordNet)	Lib-SVM with vanilla properties MT-based analysis using google translation Polarity with stop word removal	78.14% 65.96% 60.31%	Hindi English
[26]	300 Movie reviews stored in XML file	Train deep belief network using Unsupervised learning using Restricted Botzmann	71% DBM in English ,76% using deep learning 64% using DBN in	Hindi English

		Machine(RBM)	hindi	
[27]	Twitter messages,Lexicon based	Emoticons,negation,spell correction,stop word removal handling.Token based on subjectivity lexicon.	73.72% accuracy using hybrid approach with discourse	Hindi
[28]	Facebook,Twitter, Google	Subjectivity detection,sentiment polarity using network overlap technique,structurazation,summarization and visualization-tracking		Hindi
[29]	Social Network,Online forum and newspaper	Preprocessing techniques and lexicon based polarity identification, Machine learning and hybrid		Urdu
[30]	40000*5 matrix of urdu words	RNN for classifying and creating graph LSTM,Naive Bayes using softmax function		Urdu
[31]	9312 reviews of urdu words	mBert model ,KNN,LR,MLP,SVM	mBERT has more accuracy than other	Urdu
[32]	Product Reviews from internet in Urdu	SVM, NB, Regression with the help of n-gram features	RF with trigram have 55.25%	Urdu

PROPOSED METHODOLOGY

Pre-processing

Before applying sentiment analysis techniques, it is essential to preprocess the statements to enhance the accuracy of sentiment classification. The preprocessing steps include:
Consider the following Marathi statement:

"मला त्याच्यावर विश्वास आहे."

Translation: "I trust him."

Tokenization: Tokenization involves splitting the text into individual words or tokens. In our example, tokenization results in the following tokens:

Tokens: ["मला", "त्याच्यावर", "विश्वास", "आहे"]

Stopword Removal: Stopword removal eliminates common words that do not carry significant meaning. In our example, no stopwords are found, so the list remains unchanged:

Result after stopword removal: ["मला", "त्याच्यावर", "विश्वास", "आहे"]

Normalization: Normalization ensures consistent case usage. Since Marathi does not have strict upper or lower case conventions, normalization may not be necessary. However, for consistency, we can choose either lowercase or uppercase. Let's choose lowercase:

Result after normalization: ["मला", "त्याच्यावर", "विश्वास", "आहे"]

Vectorization: Vectorization converts the text data into numerical vectors. One common approach is the Bag-of-Words model, where each word is represented by a unique index, and the vector contains counts of each word in the text. Let's assume we have the following vector representation for our example:
Vectorized representation: [1, 1, 1, 1]

Experimentation

1.Support Vector Classification

Our SVC model has the following decision boundary equation:
$w\cdot x+b=0$
After training, the model's learned weight vector w and bias term b are such that:
$w = [w1,w2,w3,w4] = [1,1,1,1]$
$b=0$
Now, let's substitute the input feature vector x into the equation:
$x=[1,1,1,1]$
$w\cdot x+b=(1\cdot 1)+(1\cdot 1)+(1\cdot 1)+(1\cdot 1)+0=4$

Since the result is positive, the statement "मला त्याच्यावर विश्वास आहे" is classified as positive by the SVC model.

2. K- Nearest Neighbours

Our KNN model considers k=5 nearest neighbors for classification. We have computed distances to the 5 nearest neighbors in the dataset and found that 4 of them are positive and 1 is negative. Since the majority of the nearest neighbors have positive sentiment, the statement "मला त्याच्यावर विश्वास आहे" is classified as positive by the KNN model.

3. Naive Bayes Classification
After training, let's assume our Naive Bayes classifier has learned the following probabilities:
$P(Positive)=0.8$
$P(Negative)=0.2$
Let's denote the features of the given statement as f1,f2,f3,f4, where each feature represents one word. Given that all features are present in the statement, we calculate:

$$P(Positive|f1,f2,f3,f4)= \frac{P(f1|Positive)\cdot P(f2|Positive)\cdot P(f3|Positive)\cdot P(f4|Positive)\cdot P(Positive)}{P(f1)\cdot P(f2)\cdot P(f3)\cdot P(f4)}$$

Assuming independence between features, we can simplify this expression and compute the probability. As $P(Positive|f1,f2,f3,f4)>P(Negative|f1,f2,f3,f4)$, the statement is classified as positive.

4. Decision Tree

We traverse the tree based on the presence or absence of features in the input statement:
Start at the root node [विश्वास].

Checked if the feature "विश्वास" is present in the input statement. Since it is present, move to the left child node [त्याच्यावर].

Checked if the feature "त्याच्यावर" is present in the input statement. Since it is present, move to the left child node, which is labeled as Positive.
The traversal ends at the leaf node labeled as Positive.

Therefore, the input statement "मला त्याच्यावर विश्वास आहे" is classified as positive by the Decision Tree model based on the tree traversal.

5. Logistic Regression

Assuming our Logistic Regression model has learned the following weights and bias:
$w=[w1,w2,w3,w4]=[1,1,1,1]$
$b=0$
The probability of the statement being positive is computed as:
$P(Positive|x)= 1/1+e\hat{}-(wx+b) = 1/1+e\hat{}-(4)$
As this probability exceeds the threshold, the statement is classified as positive.
6. Random Forest Classification

The input statement "मला त्याच्यावर विश्वास आहे" is passed through each decision tree in the Random Forest.
The number of trees that classify the statement as positive are counted.
The majority of trees classify the statement as positive, therefore it's classified as positive by the Random Forest model.

7. AdaBoost Classification

The input statement "मला त्याच्यावर विश्वास आहे" is passed through each weak classifier in the AdaBoost ensemble.
The predictions from all weak classifiers are aggregated based on their respective weights.
The majority of weak classifiers predict the statement as positive, hence it's classified as positive by the AdaBoost model.

RESULTS AND DISCUSSIONS

Each language exhibits unique linguistic characteristics, posing distinct challenges for sentiment analysis. The accuracy of various algorithms, such as Support Vector Classifier, K-Nearest Neighbours, Naive Bayes Classifier, Decision Tree, Logistic Regression, Random Forest Classifier, and AdaBoost, is meticulously evaluated across these languages. Notably, the Random Forest Classifier demonstrates consistently high accuracy rates across all languages, with Bengali yielding the highest accuracy at 86.63%. Conversely, Konkani presents the lowest accuracy scores, indicating the need for tailored approaches to sentiment analysis in less-represented languages. These findings not only shed light on the effectiveness of different algorithms but also underscore the importance of language-specific considerations in sentiment analysis research and application.

MARATHI		BENGALI	
Algorithm	**Accuracy**	**Algorithm**	**Accuracy**
Support Vector Classifier	60.84%	Support Vector Classifier	81.81%
K- Nearest Neighbours	47.01%	K- Nearest Neighbours	84.93%
Naive Bayes Classifier	61.78%	Naive Bayes Classifier	83.06%
Decision Tree	54.74%	Decision Tree	82.81%
Logistic Regression	60.43%	Logistic Regression	83.95%
Random Forest Classifier	68.97%	Random Forest Classifier	86.63%
AdaBoost	58.80%	AdaBoost	83.63%

HINDI		KONKANI	
Algorithm	**Accuracy**	**Algorithm**	**Accuracy**
Support Vector Classifier	81.60%	Support Vector Classifier	56.83%
K- Nearest Neighbours	82.40%	K- Nearest Neighbours	37.30%
Naive Bayes Classifier	82%	Naive Bayes Classifier	53.10%
Decision Tree	81%	Decision Tree	53.10%
Logistic Regression	82.20%	Logistic Regression	59.14%
Random Forest Classifier	83.20%	Random Forest Classifier	63.94%
AdaBoost	80.60%	AdaBoost	53.46%

CHALLENGES

Sentiment analysis encounters a myriad of complexities that span linguistic nuances, cultural variations, and the subtleties of human expression. One of the fundamental challenges lies in recognizing subjective elements within text, where the same word or phrase can evoke differing sentiments depending on context. This intertwines with domain reliance, where words may carry different connotations based on their usage in specific domains. Moreover, identifying and interpreting sarcasm, implicit negations, and nuanced expressions further complicates sentiment analysis, as sentiments can be subtly masked or reversed.

In addition, ineffective expressions and the sensitivity of language order pose significant hurdles. Sentences with seemingly positive or negative elements may, in fact, convey a contrary sentiment due to the structure or placement of words. Furthermore, the presence of comparisons adds another layer of complexity, as sentiments can vary based on the relationship between entities being compared. Integrating these challenges is the recognition of entities within text, where sentiments may differ towards specific entities mentioned within the same statement.

Addressing these challenges is compounded by the need to accommodate internationalization and linguistic diversity, particularly in platforms like Twitter and Facebook with global user bases. Moreover, challenges such as named entity recognition and anaphora resolution underscore the importance of context and referential clarity in sentiment analysis. Beyond polarity detection, the emergence of aspect-based sentiment analysis underscores the necessity of contextual understanding and subject matter relevance in accurately assessing sentiments.

CONCLUSION

This paper outlines several SA methods, lexical options, resources and difficulties The results of certain experiments indicate that the hybrid technique has improved. The accuracy of speech recognition is always complicated by background noise and variations in speech in time. We used a different algorithm for pattern matching to increase the system's precision and effectiveness. This study will make it easier for academics to create efficient SAs for their own Indian languages by utilizing different methodologies suggested by other researchers who may subsequently help the Indian society. Speech emotion is more important than appearance since a person may modify their facial expressions more easily than their speech. Future multimodal detection systems could be developed to classify human emotional states by combining biosignals, audio signals, and visual data. This paper contains more than 50 research papers on sentiment analysis.

FUTURE SCOPE

The challenges pave a way for tremendous future research in this area. Proper combination of algorithms can eliminate the challenges and boost accuracy. Also, regional languages lag behind in this area of research as they lack appropriate dataset for study. Construction of suitable datasets language-wise can open doors for various new opportunities in this domain. Extracting sentiments from text is a tedious job. Sometimes, when voice datasets are converted to text they lose their original essence that comes from the voice notes, pitch, tone, etc. In order to maintain the authenticity of the original dataset research can be carried forward in the area of sentiment analysis from voice rather than from text. Use of voice datasets will also eradicate challenges like sarcasm and degree of polarity of words as they can be easily categorized according to transitions in voice notes. Thus, the area has many untouched aspects that can be explored in future.

REFERENCES

[1]Patra, Braja Gopal, Dipankar Das, Amitava Das, and Rajendra Prasath. "Shared task on sentiment analysis in Indian languages (sail) tweets-an overview." In *International Conference on Mining Intelligence and Knowledge Exploration*, pp. 650-655. Springer, Cham, 2015.

[2]VIDYAVIHAR, MUMBAI. "Sentiment analysis in Marathi language." *International Journal on Recent and Innovation Trends in Computing and Communication* 5, no. 8 (2017): 21-25.

[3]Rakshitha, Kakuthota, H. M. Ramalingam, M. Pavithra, H. D. Advi, and Maithri Hegde. "Sentimental analysis of Indian regional languages on social media." *Global Transitions Proceedings* 2, no. 2 (2021): 414-420.

[4]Bhoir, Nirmiti, Aarushi Das, Mrunmayee Jakate3 Snehal Lavangare, and Deepali Kadam. "A Study on Sentiment Analysis of Twitter Data for Devanagari Languages." (2021).

[5]Dutta, Satyam, Himanshi Agrawal, and Pradeep Kumar Roy. "Sentiment Analysis on Multilingual Code-Mixed Kannada Language." (2021).

[6]Hegde, Asha, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. "Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text." In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pp. 33-40. 2022.

[7]Shetty, Saritha, Sarika Hegde, Savitha Shetty, Deepthi Shetty, M. R. Sowmya, Rahul Shetty, Sourabh Rao, and Yashas Shetty. "Sentiment Analysis of Twitter Posts in English, Kannada and Hindi languages." In *Recent Advances in Artificial Intelligence and Data Engineering*, pp. 361-375. Springer, Singapore, 2022.

[8]Arun, K., A. Srinagesh, and M. Ramesh. "Twitter sentiment analysis on demonetization tweets in India using R language." *International Journal of Computer Engineering In Research Trends* 4, no. 6 (2017): 252-258.

[9]Sharma, Parul, and Teng-Sheng Moh. "Prediction of Indian election using sentiment analysis on Hindi Twitter." In *2016 IEEE international conference on big data (big data)*, pp. 1966-1971. IEEE, 2016.

[10]Rohini, V., Merin Thomas, and C. A. Latha. "Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm." In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 503-507. IEEE, 2016.

[11]Phani, Shanta, Shibamouli Lahiri, and Arindam Biswas. "Sentiment analysis of tweets in three Indian languages." In *Proceedings of the 6th workshop on south and southeast asian natural language processing (WSSANLP2016)*, pp. 93-102. 2016.

[12]Kalaivani, Adaikkan, and Durairaj Thenmozhi. "Multilingual Sentiment Analysis in Tamil, Malayalam, and Kannada code-mixed social media posts using MBERT." *FIRE (Working Notes)* (2020).

[13]Mandalam, Asrita Venkata, and Yashvardhan Sharma. "Sentiment analysis of Dravidian code mixed data." In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pp. 46-54. 2021.

[14]Fondekar, Ashweta, Jyoti D. Pawar, and R. Karmali. "Konkani sentiwordnet: resource for sentiment analysis using supervised learning approach." (2016).

[15]Rajan, Annie, and Ambuja Salgaonkar. "Part of speech (PoS) tagging for Konkani language using HMM." In *ICT Systems and Sustainability*, pp. 601-609. Springer, Singapore, 2022.

[16]Deshmukh, Sujata, NILEEMA PATIL, SURABHI ROTIWAR, and JASON NUNES. "Sentiment Analysis Of Marathi Language." *JournalNX* 3, no. 06: 93-97.
[17]Bhowmick, Anirban, and Abhik Jana. "Sentiment Analysis For Bengali Using Transformer Based Models." In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 481-486. 2021.
[18]Kulkarni, Atharva, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. "L3cubemahasent: A marathi tweet-based sentiment analysis dataset." *arXiv preprint arXiv:2103.11408* (2021).
[19]Khan, Md Serajus Salekin, Sanjida Reza Rafa, and Amit Kumar Das. "Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity." *Journal of Engineering Advancements* 2, no. 03 (2021): 118-124.
[20]Ansari, Mr Mohammed Arshad, and Sharvari Govilkar. "Sentiment Analysis of Transliterated Hindi and Marathi Script." In *Sixth International Conference on Computational Intelligence and Information*, pp. 142-149. 2016.
[21]Chakraborty, Partha, Farah Nawar, and Humayra Afrin Chowdhury. "Sentiment Analysis of Bengali Facebook Data Using Classical and Deep Learning Approaches." In *Innovation in Electrical Power Engineering, Communication, and Computing Technology*, pp. 209-218. Springer, Singapore, 2022.
[22]Bhowmik, Nitish Ranjan, Mohammad Arifuzzaman, and M. Rubaiyat Hossain Mondal. "Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms." *Array* 13 (2022): 100123.
[23]Hossain, Eftekhar, Omar Sharif, and Mohammed Moshiul Hoque. "Sentiment polarity detection on bengali book reviews using multinomial naive bayes." In *Progress in Advanced Computing and Intelligent Engineering*, pp. 281-292. Springer, Singapore, 2021.
[24]Bhowmik, Nitish Ranjan, Mohammad Arifuzzaman, M. Rubaiyat Hossain Mondal, and M. S. Islam. "Bangla text sentiment analysis using supervised machine learning with an extended lexicon dictionary." *Natural Language Processing Research* 1, no. 3-4 (2021): 34-45.
[25]Pandey, Pooja, and Sharvari Govilkar. "A framework for sentiment analysis in Hindi using HSWN." *International Journal of Computer Applications* 119, no. 19 (2015).
[26]Mittal, Namita, Basant Agarwal, Saurabh Agarwal, Shubham Agarwal, and Pramod Gupta. "A hybrid approach for twitter sentiment analysis." In *10th international conference on natural language processing (ICON-2013)*, pp. 116-120. 2013.
[27]Bakliwal, Akshat, Piyush Arora, and Vasudeva Varma. "Hindi subjective lexicon: A lexical resource for hindi adjective polarity classification." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1189-1196. 2012.
[28]Khan, Khairullah, Wahab Khan, Atta Ur Rahman, Aurangzeb Khan, Asfandyar Khan, Ashraf Ullah Khan, and Bibi Saqia. "Urdu sentiment analysis." *International Journal of Advanced Computer Science and Applications* 9, no. 9 (2018).
[29]Kumhar, Sajadul Hassan, Mudasir M. Kirmani, Jitendra Sheetlani, and Mudasir Hassan. "Sentiment Analysis of Urdu Language on different Social Media Platforms using Word2vec and LSTM." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 11, no. 3 (2020): 1439-1447.
[30]Khan, Lal, Ammar Amjad, Noman Ashraf, and Hsien-Tsung Chang. "Multi-class sentiment analysis of urdu text using multilingual BERT." *Scientific Reports* 12, no. 1 (2022): 1-17.
[31]Khan, Lal, Ammar Amjad, Noman Ashraf, Hsien-Tsung Chang, and Alexander Gelbukh. "Urdu sentiment analysis with deep learning methods." *IEEE Access* 9 (2021): 97803-97812.
[32]Sharma, Sheetal, S. K. Bharti, and Raj Kumar Goel. "Sentiment analysis of the Indian language." *International Research Journal of Engineering and Technology* 5, no. 5 (2018): 4251-53.