IJCSEInternational Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Sentiment Analysis on Indian Regional Languages: A Comprehensive Review

Sunil D. Kale[1]

[1]Computer Engineering Department, Pune Institute of Computer Technology, Pune, Maharashtra, India

Author's Mail Id: kalesunild@gmail.com

Available online at: www.ijcseonline.org

Accepted: 16/Jan/2019, Published: 31/Jan/2019

Abstract— Sentiment Analysis is the extraction of emotions from written or spoken sentences to get a broader and clearer view of the user's point of view. Their emotions significantly impact people's lives. Organizations can benefit from these feelings by gaining enormous earnings, the confidence of their clients, and their devotion. Sentiment analysis is gaining popularity in implementing better CRM functionalities for large and small firms. This paper presented a comprehensive literature review of various Indian regional Languages. Moreover, it presented challenges like Explicit rejection of feelings, diagnosing sarcasm, etc. This paper also provides future direction for improving the result of accuracy by the right mix of algorithms.

Keywords— Sentiment Analysis, Emotion analysis, Indian regional Languages, Hindi, Marathi

I. INTRODUCTION

In today's world, technology has advanced so that people increasingly engage in their hobbies on vibrant social media platforms and voice and text messaging in their native tongues to express their ideas. Many data is produced online in numerous formats, frequently on social networking sites. Enhancing their business was the main objective. An analysis of the tone's emotions is presented. People use Twitter, Facebook, and Blogs to convey their ideas and feelings. Customers now leave reviews for specific services businesses provide in the form of voice and text messages, which must be assessed.

Many companies and organizations must improve their customer service and consumer experience. This study analyzes the emotional core of various people's opinions for various uses by analyzing the emotional core of call centers' tone. This study will benefit many company enterprises by enhancing CRM capabilities. In order to improve customer happiness, various business organizations can immediately examine client feedback and reframe their offering. Using sentiment analysis from speech approaches, sales representatives for various commodities can gauge the emotions of their clients and engage them in conversation more skillfully. By modifying the product based on customer feedback, user satisfaction can be increased.

Text is used because it can be difficult to manually interpret people's feelings. Instead, a device that can recognize it by designating a polarity as positive or negative is needed. India is known as a linguistic and cultural melting pot. People communicate with one another in various languages. Hence, analyzing the polarity of such a language is difficult when data are available since the dataset must be in the right format. However, the internet contains a large amount of unprocessed data. To determine if the data is positive, negative, or neutral, sentiment analysis, a technique of Natural Language Processing, is employed. Businesses can contextualize identifying information in online interactions to better understand the social feelings of their clients. In addition to polarity, it also considers intentions (interested or not interested), sentiments and emotions (happy, sad, furious, etc.), and urgency (urgent or not urgent). Businesses use it to measure brand reputation, social media sentiment analysis, and customer happiness. Figure 1 shows the generalized framework of sentiment analysis.

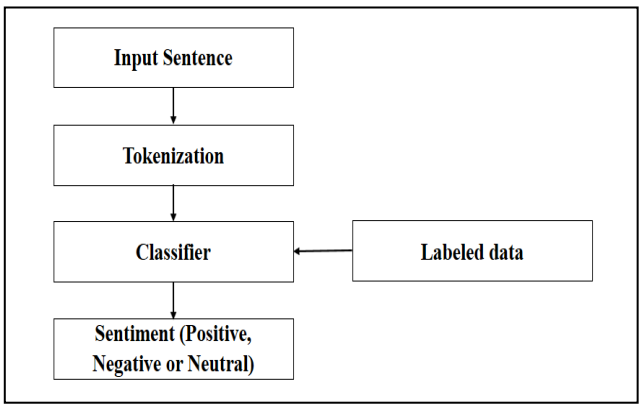

Figure 1. Framework of Sentiment Analysis

II. RELATED WORK

[1] This study article mainly aims to categorize tweets into positive, negative, and neutral polarity. The model is evaluated using data from the real world, which is the

paper's main benefit. Sentiment analysis has been done by six teams in both restricted and unconfined environments. Teams were allowed to use SentiWordNet for Indian languages developed by Das and Bandyopadhyay in the Confined System. Three languages—Hindi, Bengali, and Tamil—are used for the sentiment analysis, with the three having the highest levels of accuracy at 43.2% for Bengali, 55.67% for Hindi, and 39.28% for Tamil. Using the TWITTER4J (Application), more than 2000 tweets were used to get the data. The algorithms used for sentiment analysis include Naive Bayes, Decision Trees, Support Vector Machines, and Multinomial Naive Bayes. As a result of the inaccessibility of several NLP (Natural Language Processing) technologies, including POS taggers and NER, the accuracy of the unconfined system is lower than the accuracy of the confined system. The authors of this report suggest future research on code-mixed tweets.

[2] Finding the hidden sentiments in Marathi-language literature is the main purpose of this paper. Lexicon-based and machine learning-based sentiment analysis are said to be the two main subcategories of sentiment analysis by the authors of this research. The authors describe the Support Vector Machine, Naive Bayes, and Maximum Entropy machine learning methods. The sentences have been divided into two categories, positive and negative, in the author's proposed work. The authors argued that improved results require both strategies and resources.

[3] The paper suggests using R to conduct sentiment analysis on tweets related to demonetization. The polarity of the paragraph and the polarity of the sentence are the two different sorts of polarity that the authors utilize to categorize the tweets. The suggested approach divides 12974 tweets into 5064 neutral, 4936 neutral, and 2974 neutral categories. The writers use the R studio's graphical tools to further examine the tweets. The results of a graphic analysis suggest that 72% of tweets are favorable or acceptable.

[4] The study aims to analyze tweets for sentiment to forecast the outcome of India's general elections in 2014. By using the Twitter Archiver tool, 42,235 tweets in Hindi were gathered. The authors used the dictionary-based, Naive Bayes and SVM algorithms to categorize the tweets into positive, negative, and neutral categories. The Bharatiya Janata Party has 78.4%, 62.1, and 34% chances of winning the elections, respectively, according to the SVM, Naive Bayes, and dictionary-based models. In the future, new social media platforms like Facebook may be considered for data collection, and more machine learning algorithms may also be employed for analysis, according to the author.

[5] This study uses a hybrid approach based on machine learning and lexicon to analyze sentiment in Kannada. The information is gathered from several Kannada movie review websites. The features are then extracted once each tokenized word has been tagged with its part of speech. The decision tree classifier is employed once the material has been translated into English using Google Translator. Kannada has a recall score of 0.78 and a precision score of 0.79, while English has a recall score of 0.86 and a precision score of 0.67.

[6] The study aims to do sentiment analysis using a straightforward, reliable, scalable, and language-neutral method. This research uses two confined and unconfined environments to do the analysis. While restricted, Participants can access SentiWordNet from Das and Bandyopadhyay (2010). Participants in unconfined are allowed to use POS taggers and other data. Positive and Negative (class 2) and Positive Negative and Neutral (class 3) are used. For classes two and three, Bengali accuracy was 67.83% and 51.25%, respectively. For classes two and three, the accuracy of Hindi was 81.57% and 56.96%, respectively. Tamil accuracy in the second and third grades was 62.16% and 45.24%, respectively.

[7] The main objectives of this research are the definition and inference of the probabilistic characteristics of emotions, such as happiness, anger, sadness, and neutrality. The Author proposes to implement this system using the hidden Markov model. The model outperformed more traditional machine learning techniques, it was discovered. The model is also more precise and scalable when assessing user-generated data and defining attitudes and beliefs. The model has several shortcomings, including the fact that it becomes highly complex as more states and interactions between states are introduced, even though it has proven to be quite effective at explaining the probabilistic features of states. Opinion mining, another name for sentiment analysis, studies how people feel, think, and act.

[8] Dynamic time warping is the central concept of this investigation. This work aims to measure similar patterns across various time zones. When comparing the distance created between two time zones, it was discovered that the distance is reduced the closer the two sound patterns are to one another. The advantage of this paradigm is that it offers precise time alignment between the reference and test patterns. However, locating the optimal time alignment path necessitates extensive computational labor.
[9] This study aims to achieve uniformity in the POS tagging of all Indian languages. In order to pre-process linguistic and lexical resources in the Indian language for sentiment analysis, it is necessary to locate the relevant resources, such as annotated datasets. It focuses on employing sentiment analysis to use the explosion of data. The main advantage of this article is the focus on user-generated content, which might originate from users who may also be content developers. This study's limitation is that only a few different languages are covered in these surveys.

[10] Dynamic time warping is the central concept of this investigation. This work aims to measure similar patterns across various time zones. The closer the two sound

patterns are to one another, the greater the distance produced between the two time zones.

[11]This paper's main objective is to comprehensively summarize Konkani NLP work. The bigger background of this paper is that NLP resources for many Indian languages lag behind the rapid digitization of the economy. Although the technologies are still not generally available, NLP has the potential to promote the rapid expansion of Indian languages. The newspaper estimates that 2.3 million people speak Konkani, or about 0.19% of India's population. This survey focused on the work and resources required for more research. This review provides an overview of the history of Konkani, pertinent linguistic sources, and NLP investigations conducted in several languages. NLP research in the Konkani language was conducted in four areas as part of the Text-to-Speech transformation. The paper's intelligibility and natural sensation of sound increased from 1.8 to 2.5 and 3.5 on a 5-point scale, respectively. According to claims, 64% of text-to-speech conversions are accurate. Parts of speech (PoS) tags: To choose the best tag for a given word, a PoS tagger is suggested to use the Hidden Markov Model (HMM) and Viterbi algorithm as its architecture. Using a corpus of 268,000 words, a finite state machine (FSM) is constructed, and the system reports a 95% success rate. Sentimental evaluation: A native Bayes classifier was developed and tested using 50 Konkani poems. Its accuracy for available data was 82%, and for unobserved data was 70%. The challenges that Konkani faces as a language are discussed in this essay. This study focuses primarily on the Parts of Speech tagging, a crucial technique for natural language processing. Other names for POS include word-category disambiguation and grammatical tagging. In this study, each input is properly categorized according to its kind, including noun, verb, adjective, pronoun, conjunction, etc. It was divided into supervised and unsupervised taggers. In this paradigm, there are three main steps. 1. Data gathering Data collecting for Konkani text is covered here. 2. Understanding tag sets and frequencies can be useful during the training stage. 3. By attempting to predict the tag for unseen data, the model is put to the test. The HMM, a set of states, the probability of a state change, and the likelihood of an observation serve as the basis for this study. The system uses the POS for information extraction and question-answering. The suggested model annotated the terms in the text using the HMM and Viterbi methods to minimize the distance created. The advantage of this paradigm is that it offers precise time alignment between the reference and test patterns. However, locating the optimal time alignment path necessitates extensive computational labor.

[12] The HMM-based voice recognition system technique for India's Gujarati language was defined in this study. Constructing and applying SRS for the common Gujarati language is challenging because of the language barrier, complicated linguistic structure, and morphological variation. The main objective of this article is to overcome this challenge and train HMM-based SRS. The 650 frequently used words were collected from 40 speakers in South Gujarat who were selected based on their gender and who represented a cross-section of the region's 40 speakers. The accuracy was achieved with a 12.7% average error rate of around 87.23%.

[13] The primary goal of this research is to extract the sentiment for the Hindi language using simple classifiers. This work first proposed estimating the sentiment of the document using lexical resources. SentiWordNet is a word-based lexical database with hostile, neutral, and positive polarity scores. With the aid of sentiment scores from English-SWN, which led to the creation of HSWN and sentiment-related scores, H-SWN, or Hindi SentiWordNet, is developed. For the sentiment analysis, the following three approaches were taken into account:

Language-specific sentiment analysis Hindi-language sources are used for analysis. Use RapidMiner5.0 to categorize the documents. For classification, LibSVM's learner with CSVC and Vanilla attributes is utilized. It has a 78.14% accuracy rate. Sentiment analysis based on MT: Assuming the sentiment is maintained, Google Transform employs a translation mechanism to transform Hindi into English. Using translated information as input, a classifier produces the polarity with an accuracy rate of 65.96%. Sentiment analysis based on resources: H-SWN was used to do majority-based sentiment analysis for this method. This document has stop words eliminated, and the sentiment score for each word is calculated. Authors demonstrates that the first method yields better results than a corpus annotated in the same language. In the future, it may be necessary to create a new dataset because there aren't enough accurate datasets with better coverage.

[14] This study focuses on automatically classifying tweets for positive and negative sentiment, which is useful for people trying to buy things and companies looking to use sentiment analysis to assess customer feedback. These are the methods that are used. Data Pre-processing: The following procedures must be taken to extract meaningful tokens from tweets: I.Emoticons Handling: Internet emoticons are important because they are significant ways of elaborating the user's mood. II. Negation Handling: Not is used in place of all negative sentences. III. Spelling correction: Many terms have repeated letters on Twitter and other social media platforms. These words are changed to single letters, and their weight in the final score is doubled IV. Stop words are eliminated since they are useless for assessing sentiment. Some examples include, therefore, that, my, etc. V. Slang Handling: Many short versions of words absent from Hindi lexical resources must be replaced with words with the same meaning, such as asap, which stands for as soon as feasible. Three-stage hierarchical approach It is connected to emoticons in the first stage, predetermined lists in the second, and a subjectivity lexicon in the third.

I.Label With a Predefined Lexicon The easiest way to communicate a user's mood is through their emoticons, which may be used to categorize a tweet as good or negative. II. Label using Predefined List: A predefined list of strongly positive and negative terms is evaluated against the tweet to see if it contains either of the two and, if so, whether the semantic orientation is in line with the scores by scoring the positive and negative words.

III.Token Weighing based on Subjectivity Lexicon: A subjectivity lexicon is employed to ascertain the overall semantic orientation if the testing tweet is not labeled in the first two phases. Several phases are involved in this, including token creation, which neglects weightless words; negation handling, which reverses the polarity to avoid making mistakes; and discourse analysis, which establishes coherent relationships. Two techniques are used to assign tokens weight and polarity:

The first technique uses the SWN dataset, which has the drawback that 90% of the vocabulary has higher object values, making it difficult to discern most of the polarity.
The second probability-based approach assigns a positive polarity to the sentence in which the chance that the word can be positive is calculated.

This study demonstrates that adding discourse indicators improves sentiment categorization precision. This study examined several methods, from pre-processing the data to determining the token's polarity.

[15] Adjectives and adverbs are occasionally used to express feelings and opinions. In this project, WordNet is used as a resource to build a subjective lexicon for the Hindi language using the breadth-first graph traversal method. Bi-lingual Dictation and translation of dictation are other methods for creating lexicons. A WordNet seed list of 45 adjectives and 75 adverbs is the foundation for this algorithm. The antonym is assumed to exhibit the original word's opposite polarity. If words are connected based on their synonymy or antonymy, then an adjective and adverb lexicon should result. Two different kinds of datasets are produced, and they are as follows:

I. Hindi Subjective Lexicon: The II has 888 adverbs and 8048 adjectives comprising positive, negative, and objective words. Dataset for Product Reviews: Google Translate was used to translate the Amazon product reviews into English, and the translation was then submitted for manual validation by judges.

The Hindi subjectivity lexicon is built using a graph traversal method in this research, which also uses the relationship between synonyms and antonyms. This method has been tested for Hindi and has a 79% classification and review accuracy rate. In order to support product reviews, marketing campaigns, social events, etc., it is essential to obtain public opinion. In this work, each word is assigned a score using a subjective lexicon, machine learning is used to perform supervised or semi-supervised learning by extracting textual features, and n-gram modeling is used to train the n-gram model. The processing of sentiment analysis for a text is separated into five parts, including the creation of the lexicon. Sentiment polarity detection uses a network overlap technique to classify sentiments, sentiment structuration based on 5W, sentiment summarization visualization-tracking, and subjectivity detection, in which text can be subjective or objective in which subjective contains opinion and objective contains no opinion. While the preceding polarity assignment is carried out manually in this paper, coverage extension benefits from employing an automated procedure.

[16] The focus of this study is on some extra SA initiatives for Urdu. Urdu writing is said to need to be more accessible due to concerns with irregular space utilization and space omission. The following are the steps for sentiment analysis in Urdu: Data pre-processing involves several stages, including noise removal, sentence boundary detection, tokenization, and, occasionally, part-of-speech tagging. Polarity Positive, negative, and neutral connotations are found for words and statements. Good training data must be available in large quantities for proper algorithms to be provided. For Urdu, data can be gathered from free sources like news and other media. The sentence is divided into tokens in the subsequent stage of tokenization. By comparing the tokens to the sentiment lexicon, tokens are utilized to determine the polarity. Then, by adding them up, the allocated single polarity is calculated. This result illustrates the findings that are favorable, unfavorable, and neutral. Numerous difficulties arise as social networks, media, and online forums produce the dataset. Despite having few resources, Urdu is a lively language. Different strategies were employed, including lexicon-based, machine-learning, and hybrid strategies. These publications examine how several approaches and prospective strategies are available, but little work is done in Urdu. Enhancing Urdu SA using knowledge from different Urdu web forums is feasible.

[17] With data mining and machine learning techniques, this research article intends to determine the attitude of the sentence in the Hindi language from the information gathered from newspapers. India's official language is Hindi, which is used and understood throughout the nation. Reviews, news articles, and blogs are just a few examples of text types mined for sentiment data and categorized as positive, negative, or neutral based on their polarity. In order to determine the greatest accuracy, this study uses naive Bayes as a probabilistic model of words on text classification and draws a graph of features and their accuracy. Numerous Hindi sentences are pulled from the internet in order to gather data. Tagging is employed for the mining portion of speech and serves as a tagger for the review words. The polarity is identified with the aid of the seed list and a Hindi dictionary, and negation is handled by flipping the polarity. The main source for experiments is movie reviews. Precision, recall, and

accuracy were evaluated using a polarity matrix as three metrics.

Data mining for sentiment analysis is becoming more popular since it is crucial for today's humans, who rely heavily on the internet. The best k is discovered for news phrases in Hindi at an occurrence 850, and accuracy is approximately 0.66.

[18] Instead of being a pioneer, this work aims to expand that research for future advancement. In order to boost precision, this study evaluated the status, accomplishments, and standards of the researchers in the field and replaced them without a recommended approach. Even though the current work is a proposal that enhances existing methodologies, it will also be fairly comparable in light of the previous findings. The goal is to enhance what has already been created or demonstrated to be true and determine whether the simplest course of action is still the best to take. It refers to the currently used direct supervised learning for sentiment analysis, devoid of extensive NLP or language-specific research. This work is certain to produce significant advancements in the field of sentiment analysis because it experiments with the current state of the art and ventures into a previously unexplored territory (Marathi transliterated text).

[19] The suggested technique finds hidden sentiments in texts written in Marathi. For the system to perform as intended, sentiment analysis methodology is used. In this method, a corpus-based strategy is suggested, which calls for producing an updated, diverse corpus of Marathi keywords and each keyword's unique polarity about the Word Net, which is considered a corpus. The method determines if a sentence is positive, negative, or neutral on a predetermined polarity scale.

[20] This paper offers a backup method for performing sentiment analysis on Hindi documents, an issue for which no previous research has been conducted.

(A) After researching three techniques for performing SA in Hindi, a sentiment-annotated corpus was created in the domain of Hindi movie reviews. The first method entails using this annotated Hindi corpus to train a classifier that will be used to categorize fresh Hindi documents. (B) In the second method, translate the provided content into English and categorize it using a classifier trained on typical English movie reviews. (C) In the third method, authors prepared the Hindi-SentiWordNet (H-SWN) lexical resource and used a majority score-based classification algorithm to categorize the given material. According to a performance comparison of different approaches, a fallback strategy for performing sentiment analysis for a new language is to (1) Train a sentiment classifier on in-language labeled corpora and use this classifier to categorize new documents. (2) If in-language training data is unavailable, use crude machine translation to translate the new material into a language with abundant resources, such as English, and, assuming polarity is not lost in translation, identify the polarity of the translated document using a classifier for English. (3) If translation is impossible, create a SentiWordNet-like resource for the new language and use a majority approach to classify the material. Creating a Hindi SentiWordNet lexical resource and establishing an emotion-tagged corpus for Hindi movie reviews. There are two further contributions by the author.

[21] This study primarily focuses on sentiment analysis of Twitter data, which is useful for analyzing information in tweets when opinions are very unstructured, varied, and occasionally neutral. This study presents a survey, a comparative analysis, and evaluation metrics of existing methods for opinion mining, such as lexicon-based and machine learning methods. Research on Twitter data streams using several machine learning techniques like Naive Bayes, Max Entropy, and Support Vector Machine. This paper also discusses the general difficulties and uses of sentiment analysis on Twitter.

[22] This study discusses a score-based sentiment mining method for Hindi that extracts the emotion hidden in sentences from book reviews. First, prospective scores for opinion words were extracted using their parts-of-speech tags from the HindiSentiWordNet (H-SWN) scores in three tests. Then, the authors concentrated on word-sense disambiguation (WSD) to improve the system's accuracy. Finally, managing morphological changes improved the classification outcomes. The results achieved an overall accuracy of 86.3% when tested against human annotations. The work was further expanded with the help of the Hindi Subjective Lexicon (HSL). Additionally, I created an annotated corpus of Hindi book reviews.

[23] This article introduces the WKWSCI Sentiment Lexicon, a new general-purpose sentiment lexicon. It contrasts it with five current lexicons: Hu & Liu Opinion Lexicon, Semantic Orientation Calculator (SO-CAL), Multi-perspective Question Answering (MPQA) Subjectivity Lexicon, General Inquirer, and NRC Word-Sentiment Association Lexicon. Using data sets from Amazon product reviews and news headlines, the efficacy of the sentiment lexicons for categorizing sentiment at the document and sentence levels was assessed. When proper weights are employed for various categories of sentiment terms, WKWSCI, MPQA, Hu & Liu, and SO-CAL lexicons achieve accuracy rates of 75%–77% when used to categorize the sentiment of product reviews.

[24] Recently, deep learning has used its autonomous feature learning approaches to outperform most cutting-edge conventional techniques for sentiment analysis. This essay seeks to give previous analyses of attitudes in the most widely spoken yet underappreciated Indian languages. In addition, potential issue areas that can be resolved in the multilingual field are studied.

[25] This essay critically examines the difficulties of conducting sentiment analysis on tweets in English and

Indian regional languages. This study considered Tamil, Malayalam, Telugu, Hindi, and Bengali, five Indian languages. And several issues with the conceptualization and identification of Twitter sentiment analysis challenges in those languages through a systematic review, this research developed a framework.

[26] A mixed corpus of Kannada and English was constructed by crawling Facebook comments. The literature and related corpus for code-mixed Kannada-English sentiment analysis are currently unavailable. Researchers employed the Sentiment Analysis for Indian Languages (SAIL)-2017 code-mixed corpus in addition to the crawling corpus comprising Bengali-English and Hindi-English languages. The effectiveness of distributed representation techniques in sentiment analysis is also covered in this work. Reported contrasts between several deep learning and machine learning methods.

[27] To identify the sentiments of Hindi-English (Hi-En) code-mixed data, presented an ensemble of character-trigram-based LSTM and word-n-gram-based Multinomial Naive Bayes (MNB) models. The ensemble model combines the strengths of rich sequential patterns from the LSTM model and the polarity of keywords from the probabilistic n-gram model to find attitudes in sparse and inconsistent code-mixed data. Research on user code-mixed data shows that compared to numerous baselines and other deep learning-based proposed solutions, the proposed method produces state-of-the-art outcomes.

[28] The task of sentiment analysis on text data is the main emphasis of this study. Done feature extraction and ranking, classify the sentiment expression to categorize the polarity of the text review on a scale from negative to positive, and then utilize these features to train a proposed classifier to categorize the text data into its correct label.

[29] Sentiment analysis of tweets in the Malayalam language of South India was used in this work. Recurrent Neural Networks Long Short-Term Memory, a deep learning method, is the model utilized to forecast the analysis of feelings. Accuracy goals were found to rise with dataset quality and depth.

[30] The authors are concentrating on pre-processing the words provided by the user through their reviews in the Malayalam language on social networking sites. After completing the pre-processing steps, the authors calculated the decrease in word count, and the studies revealed that more than 20% of the word count was reduced.

[31] This study article describes the (i) importance of sentiment analysis utilizing Machine Learning (ML) techniques, (ii) significance of sentiment analysis in predictions, and (iii) categorization of data based on ML approaches. (iii) a survey on using microblogging platforms to forecast epidemics and outbreaks is conducted through significant research articles published between 2010 and 2017.

[32] Attempts to create a Telugu sentence corpus that meets the highest standards of annotation and support for Telugu Sentiment Analysis are discussed in this work. The ACTSA (Annotated Corpus for Telugu Sentiment Analysis) corpus is a collection of Telugu sentences obtained from various sources, pre-processed, and manually annotated by Telugu native speakers using pre-established annotation rules. Prepared corpus is the largest dataset currently accessible, with 5457 annotated sentences in total. The annotation standards and the corpus are made available to the general audience.

[33] The study of sentiment in Indian languages, including Hindi, Telugu, Tamil, Bengali, Malayalam, etc., has attracted more attention from researchers. This paper used Telugu SentiWordNet to propose a two-phase sentiment analysis for Telugu news sentences. It first defines subjectivity categorization, categorizing statements as subjective or objective. Because they lack any sentimental significance, objective phrases are viewed as having neutral sentiments. The subjective sentences are then further divided into positive and negative

[34] Here, the architecture of the autoencoder helps generate sentiment analysis in two languages. Sentiment Analysis of two languages can be performed by using the Bilingually Constrained Recursive Auto-encoder (BRAE) model and also with the help of linked Wordnet datasets.

[35] This research proposes a unique approach that considers linguistic code flipping and grammatical transitions between the two languages under consideration and is intended to perform efficient sentiment analysis of bilingual sentences written in Hindi and English. According to experimental evaluation using real-world, code-mixed datasets taken from Facebook, the proposed approach attained extremely good accuracy and was very effective in performance.

[36] The caliber of the training corpus produced directly impacts the supervised learning method's accuracy. To classify the sentiment, a pre-annotated seed list of terms and their WorldNet-sourced synonyms and antonyms is used to generate a sentiment lexicon. Dictionary-based learning approaches require less processing time than supervised learning methods do. However, there isn't a sentiment lexicon easily accessible for Malayalam. Therefore, a sentiment lexicon must be constructed to analyze sentiment using a lexicon-based technique.

[37] Here, a location-based study of the campaign was conducted, and together with a monthly and weekly analysis of the tweets, it was possible to anticipate how polarizing each tweet would be. The studies were carried out in five stages: tweet extraction and pre-processing, tokenization, line sentiment analysis, blog sentiment analysis, and analysis. The suggested tool can also handle transliterated words. Unbiased tweets about this particular campaign were pulled from Twitter, and when compared to hand tagging, the unigram machine learning approach achieved an accuracy rate of 84.47%. This strategy aids

the government in carrying out social programs for societal improvement.

[38] This study employs support vector machines to analyze sentiment in Punjabi news articles. Because there is a rising volume of Punjabi content online, sentiment analysis on the Punjabi language is necessary. This is because it allows researchers, organizations, and governments to examine user-generated content and extract valuable information. Supervised machine learning techniques called support vector machines are used for classification and regression issues. The investigation focuses on determining if Punjabi content is positive or unfavorable. The proposed system's results show astounding precision. Support vector machine sentiment analysis on Punjabi news articles is found to be 90% accurate.

[39] The model for categorizing Hindi-speaking documents into different classes using ontology is proposed in this study. Additionally, HindiSentiWordNet (HSWN) is used for sentiment analysis to assess the polarity of each class. Combining HSWN with LMClassifier improved the accuracy of the results of the polarity extraction.

[40] In this article, an attempt was made to create a system for extracting sentiments from coded mixed sentences for English that combine Tamil, Telugu, Hindi, and Bengali with four other Indian languages. The technique utilized is split into two steps, namely Language Identification and Sentiment Mining Approach, due to the complexity of the problem. Outcomes are compared to a baseline of English sentences that were automatically translated, and it is discovered that the evaluated outcomes are around 8% more precise. The suggested method is adaptable and reliable enough to handle additional identification languages and unusual foreign or superfluous words.

III.CHALLANGES

Identifying subjective material in a text: Sentimentality is represented by subjective content. The same word may be used subjectively in one situation, while in another, it may be used objectively. This makes it difficult to identify the text's subjective passages.

Domain dependence: In other contexts, the same words or phrases may mean something completely different. For instance, the word "unpredictable" has a positive connotation when used concerning plays, movies, etc., yet it serves a terrible function when used to describe a car's steering.

Diagnosing Sarcasm: Sarcasm is the first use of positive language to express an opponent's negative opinion. For instance: "Nice scent. You must wash it. The sentence is absolutely positive, but its underlying meaning is the opposite.

Explicit rejection of feeling: Various techniques can explicitly deny sentiment rather than using the basic negatives no, not, never, etc. These negations are difficult to distinguish.

IV.CONCLUSION AND FUTURE SCOPE

This presents comprehensive review on various Sentiment Analysis techniques, linguistic alternatives, resources, and challenges. Specific experiment findings show that the hybrid approach has advanced the accuracy of results. To improve the system's accuracy and efficiency, the authors implemented a novel pattern-matching algorithm. By applying various approaches offered by other researchers who may afterward assist Indian society, this study will make it simpler for academics to develop effective Sentiment Analysis for their regional Indian languages. Since a person may alter their facial emotions more quickly than their speech, spoken emotion is more significant than looks. Future multimodal detection systems might integrate bio-signals, auditory signals, and visual data to categorize human emotional states.

Future Scope: The difficulties open the door for enormous future research in this field. The right mix of algorithms can remove barriers and improve accuracy. Regional languages must also catch up in this field of study because they require pertinent datasets for analysis. The creation of linguistically appropriate datasets can create a number of new opportunities in this field. Text sentiment extraction is a laborious process. Voice datasets can occasionally lose their original nature when they are converted to text due to the voice notes, pitch, tone, etc. Research in sentiment analysis from speech rather than text can be continued in order to preserve the authenticity of the original dataset. As sarcasm and word polarity can be easily identified based on voice note transitions, using voice datasets will also eliminate problems like these. As a result, there are still a lot of unexplored areas in the area that need to be investigated.

REFERENCES

[1] Patra BG, Das D, Das A, Prasath R. Shared task on sentiment analysis in Indian languages (sail) tweets-an overview. InMining Intelligence and Knowledge Exploration: Third International Conference, MIKE 2015, Hyderabad, India, December 9-11, 2015, Proceedings 3 2015 (pp. 650-655). Springer International Publishing.

[2] Snehal Pawar, Swati Mali, "Sentiment Analysis in the Marathi Language," International Journal on Recent and Innovation Trends in Computing and Communication vol. 5, no. 8, pp. 21-25, Aug. 2017

[3]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): 252258.

[4]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.

[5]Rohini, V., Merin Thomas, and C. A. Latha. "Domain-based sentiment analysis in regional Language-Kannada using a

machine learning algorithm." In 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 503-507. IEEE, 2016.

[6]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.

[7]Kaur, Parneet, Parminder Singh, and Vidushi Garg. "Speech recognition system; challenges and techniques." International Journal of Computer Science and Information Technologies 3, no. 3 (2012): 989-3992.

[8]Juang, B-H. "On the hidden Markov model and dynamic time warping for speech recognition—A unified view." AT&T Bell Laboratories Technical Journal 63, no. 7 (1984): 1213-1243.

[9]Sardesai, Madhavi, Jyoti Pawar, Shantaram Walawalikar, and Edna Vaz. "BIS annotation standards about Konkani language." In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, pp. 145-152. 2012.

[10]Ding Jr, Ing, Chih-Ta Yen, and Yen-Ming Hsu. "Developments of machine learning schemes for dynamic time-wrapping-based speech recognition." Mathematical Problems in Engineering 2013 (2013).

[11]Dessai, Nilesh Fal, Gaurav Naik, and Jyoti Pawar. "Development of Konkani TTS system using concatenative synthesis." In the 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 344-348. IEEE, 2016

[12]Tailor, Jinal H., and Dipti B. Shah. "HMM-based lightweight speech recognition system for the Gujarati language." In Information and Communication Technology for Sustainable Development, pp. 451-461. Springer, Singapore, 2018.

[13]Pandey, Pooja, and harvari Govilkar. "A framework for sentiment analysis in Hindi using HSWN." International Journal of Computer Applications 119, no. 19 (2015).

[14]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.

[15]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.

[16]Khan, Khairullah, Wahab Khan, Atta Ur Rahman, Aurangzeb Khan, Asfandyar Khan, Ashraf Ullah han, and Bibi Saqia. "Urdu sentiment analysis." International Journal of Advanced Computer Science and Applications 9, no. 9 (2018).

[17]Sharma, Sheetal, S. K. Bharti, and Raj Kumar Goel. "Sentiment analysis of Indian language." International Research Journal of Engineering and Technology 5, no. 5 (2018): 4251-53.

[18] Ansari MA, Govilkar S. Sentiment analysis of transliterated Hindi and Marathi Script. InSixth International Conference on Computational Intelligence and Information 2016 (pp. 142-149).

[19] Deshmukh S, Patil N, Rotiwar S, Nunes J. Sentiment analysis of Marathi language. International Journal of Research Publications in Engineering and Technology [IJRPET]. 2017 Jun;3:93-7.

[20] Joshi A, Balamurali AR, Bhattacharyya P. A fallback strategy for sentiment analysis in Hindi: a case study. Proceedings of the 8th ICON. 2010 Apr.

[21] Kharde V, Sonawane P. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971. 2016 Jan 26.

[22] Hussaini F, Padmaja S, Sameen S. Score-based sentiment analysis of book reviews in Hindi language. International Journal on Natural Language Computing. 2018 Oct;7(5):115-27.

[23] Khoo CS, Johnkhan SB. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science. 2018 Aug;44(4):491-511.

[24] Chakraborty K, Bag R, Bhattacharyya S. Relook into sentiment analysis performed on Indian languages using deep learning. In2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) 2018 Nov 22 (pp. 208-213). IEEE.

[25] Soman SJ, Swaminathan P, Anandan R, Kalaivani K. A comparative review of the challenges encountered in sentiment analysis of Indian regional language tweets vs English language tweets. International Journal of Engineering & Technology. 2018;7(2):319-22.

[26] Shalini K, Ganesh HB, Kumar MA, Soman KP. Sentiment analysis for code-mixed Indian social media text with distributed representation. In2018 International conference on advances in computing, communications and informatics (ICACCI) 2018 Sep 19 (pp. 1126-1131). IEEE.

[27] Jhanwar MG, Das A. An ensemble model for sentiment analysis of Hindi-English code-mixed data. arXiv preprint arXiv:1806.04450. 2018 Jun 12.

[28] Thakur P, Shrivastava DR, DR A. A review on text-based emotion recognition system. International Journal of Advanced Trends in Computer Science and Engineering. 2018 Sep;7(5).

[29] Thomas M, Latha CA. Sentimental analysis using recurrent neural network. International Journal of Engineering & Technology. 2018;7(2.27):88-92.

[30] Mathews DM, Abraham S. Effects of Pre-processing Phases in Sentiment Analysis for Malayalam. Int. J. of Comp. Sci.and Engg. 6(7) July 2018

[31] Singh R, Singh R, Bhatia A. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. Int. J. Adv. Sci. Res. 2018 Mar;3(2):19-24.

[32] Mukku SS, Mamidi R. Actsa: Annotated corpus for Telugu sentiment analysis. InProceedings of the First Workshop on Building Linguistically Generalizable NLP Systems 2017 Sep (pp. 54-58).

[33] Naidu R, Bharti SK, Babu KS, Mohapatra RK. Sentiment analysis using telugu sentiwordnet. In2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) 2017 Mar 22 (pp. 666-670). IEEE.

[34] Impana P, Kallimani JS. Cross-lingual sentiment analysis for Indian regional languages. In2017 International conference on electrical, electronics, communication, computer, and optimization techniques (ICEECCOT) 2017 Dec 15 (pp. 1-6). IEEE.

[35] Pravalika A, Oza V, Meghana NP, Kamath SS. Domain-specific sentiment analysis approaches for code-mixed social network data. In2017 8th international conference on computing, communication and networking technologies (ICCCNT) 2017 Jul 3 (pp. 1-6). IEEE.

[36] Ashna MP, Sunny AK. Lexicon-based sentiment analysis system for Malayalam language. In2017 International conference on computing methodologies and communication (ICCMC) 2017 Jul 18 (pp. 777-783). IEEE.

[37] Tayal DK, Yadav SK. Sentiment analysis on social campaign “Swachh Bharat Abhiyan” using unigram method. AI & SOCIETY. 2017 Nov;32:633-45.

[38] Kaur G, Kaur K. Sentiment detection from Punjabi text using support vector machine. International Journal of Scientific Research in Computer Science and Engineering. 2017 Dec;5(6):39-46.

[39] Pundlik S, Dasare P, Kasbekar P, Gawade A, Gaikwad G, Pundlik P. Multiclass classification and class based sentiment analysis for Hindi language. In2016 international conference on advances in computing, communications and informatics (ICACCI) 2016 Sep 21 (pp. 512-518). IEEE.

[40] Bhargava R, Sharma Y, Sharma S. Sentiment analysis for mixed script indic sentences. In2016 International conference on advances in computing, communications and informatics (ICACCI) 2016 Sep 21 (pp. 524-529). IEEE.

Authors Profile

Sunil D. Kale received his graduate degree from the Government College of Engineering, Aurangabad, Maharashtra, India, and a master's degree from M. Tech. (CSE) from Visvesvaraya Technological University, Karnataka, India. He is a Research Scholar in Computer Engineering Department of Smt. Kashibai Navale College of Engineering, Vadgaon (Bk), and working as Assistant Professor in Computer Engineering Department of Pune Institute of Computer Technology, Pune, India. His area of interest is text analytics and pattern recognition. He has published 17 papers in national and international journals and conferences. He is a life member of Indian Society of Technical Education.

