iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55149

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Review on Sentiment Analysis of Marathi Language of Maharashtra

Ramnath Mahadeo Gaikwad[1], Rajashri Ganesh Kanke[2], Manasi Ram Baheti[3]
[1, 2, 3]*Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad*

Abstract: ***Sentiment analysis on numerous Regional languages is performed, and classification algorithms based on Lexicon, Dictionary, and Machine Learning are employed. Because of the widespread usage of social media platforms, people are rapidly turning to the internet to find and discuss information, thoughts, opinions, feelings, perspectives, facts, and suggestions, resulting in a plethora of user-generated emotion enormous amounts of text data available for analysis. A large number of individuals in India express themselves in multiple languages, resulting in a massive amount of Natural Language Processing text data for (NLP) researchers. Sentiment Analysis (SA) of code-mixed text provides valuable information in politics, education, services marketing, business, health, sports, and other sectors. Work on Indian Language Sentiment Analysis Textual Data, particularly in Hindi, has gained steam in the previous decade in comparison to code-mixed Indian language text. However, due to a lack of language and vocabulary (linguistic and lexical) tools and annotated resources, the process of Sentiment Analysis of Regional Languages becomes very difficult. The goal of this research was to present a complete summary of the Sentiment Analysis of Regional languages, with a focus on code-mixed Regional languages.***
Keywords: NLP, SA, Marathi language, Opinion mining, Lexicon based, Machine learning

I. INTRODUCTION

With the growing usage of cellphones, social media and e-commerce have had a tremendous impact on human lives. These services enable users to exchange and access the opinions of people all around the world. As a result, there is an explosion of data on the internet. Opinions serve as the primary motivators for nearly all human activities and behaviours. We seek comments before making a decision. People used to rely completely on their friends and family for opinions and remarks on any subject. When a company wanted public feedback on its products and services, it conducted surveys and opinion polls, which were time-consuming, labor-intensive, and expensive [1].

The rapid rise of social media platforms such as Facebook, Twitter, Fora, and others allows users to freely share their ideas on the Web, which may be used for decision-making by others. Individuals use these sites to find out what other people think about a product or service before purchasing it, or to understand what other people think about political candidates. Businesses must strengthen their marketing efforts. However, even when huge volumes of data are available, the knowledge that may be obtained from the data must still be extracted, and this extraction is not easy [2].

Each site has a great amount of opinionated language about goods, places, films, and so on, but the viewer is unable to extract the viewpoints when the data becomes large. Sentiment Analysis (SA) methods are thus used to extract and summarise significant information for the reader. Sentiment Analysis has gained attention in recent decades as an active study area in the field of Natural Language Processing (NLP) to examine people's opinions, attitudes, and feelings. assessments, appraisals, attitudes, and feelings towards services, institutions, people, events, problems, subjects, and their attributes [3]. The primary reason that Sentiment Analysis is such an active research topic is that it has a wide range of applications in every area. Sentiment Analysis study has a huge impact on NLP, management sciences, political science, economics, and social sciences because they are all influenced by people's thoughts.

Document, phrase, or aspect-level sentiment analysis is possible. Sentiment is assessed at the document level by treating the entire document as a basic information unit. In the sentence level analysis, each sentence is considered separately, whereas the aspect level is utilised to classify sentiments in relation to entity aspects. Researchers select them based on the nature of the analysis [4]. With the increasing popularity of exchanging ideas in native languages across websites, there is a requirement for Sentiment categorization in local languages, such as positive, negative, harm, no harm, and so on.In English, there are numerous ways to classify emotions [3], However, they are uncommon in Indian languages such as Tamil. Because no review papers were found in our survey search, this review article will investigate the techniques and corpora used in Sentiment Analysis research in Tamil. This survey categorises current articles according on methodology, key sentiment difficulties, and resources used.

This could help academics choose applicable tactics for specific applications and educate newbies to the field with a broad perspective. In addition, the available benchmark data sets are investigated. It includes the most often used Sentiment Analysis techniques and applications, as well as assessments [5].

II. RELATED SURVEYS

Despite the fact that Sentiment Analysis (SA) research has reached mainstream product scenarios, significant advances in the discipline have been accomplished from all research perspectives. We will examine the work that has previously been completed in order to attain the objectives specified in this dissertation.Code-mixing techniques and associated studies have been around for decades. The early research attempted to determine the language in mixed code texts by learning the structure of the language from the informant and/or the given text. It was discovered that rule detection by pure text alone is unsuccessful and is heavily reliant on information containing predefined linguistic rules [6].Several corpora and usages on the Indian subcontinent were discovered to use maximum entropy classifiers for text classification, with an accuracy rate of around 80%. This groundbreaking study in the field of language identification paved the way for Indian languages to flourish. The project produced a SentiWordNet for the English language, which was crucial for all later sense-based lexical analysis work. Character-based sequence modelling and technologies such as conditional random fields were used to offer a strategy for transliterating languages [8]. Thus, lexical analysis tools were utilised to improve the accuracy of transliteration generated by statistical methodologies, and the results were proven by [7].

SVM and Naive Bayes were used in their work to categorise and identify millions of comments for sentiment polarity.They proposed employing both lexicon-based and emoticon-based detectors to classify sentiment. The work[12] established the efficacy of a text normalisation strategy focused on managing abbreviations, spelling errors, missing punctuation, slang, wordplay, censor evasion, and emoticons. During a recent study on the problem, I coined the term "textual code-switching" and developed a method for identifying texts that contain code-switching [9]. Another of their papers from 2011 revealed that lexical analysis using SentiWordNet sense has the potential to outperform traditional word-based analysis. They subsequently created a strong sentiment analysis system based on that strategy [10]. The only subjective lexical resources used for adverbs and adjectives are WorldNet and the graph traversal approach. A fuzzy logic membership function [12] can be used to measure the level of emotion polarity for a specific POS-tagged preposition. In a survey, machine translation techniques for transliteration were utilised, and it was suggested that the phoneme-based approach be used to generate transliteration utilising bilingual corpora by using a CRF-based classification algorithm [13].

Negation and discourse analysis aided sentiment analysis in moving forward and achieving 80.21% accuracy using supervised techniques to categorise the word and multi-grams for feature modelling, which was a remarkable achievement that established a standard in the field of language recognition and labelling. Maximum Entropy outperformed the others, with Naive Bayes close behind [16]. The algorithms Max Entropy, Naive Bayes, and Decision tree were used. The attempt was made to conclude that POS Tagging had overall benefits, but the current research shows differently [17]. Clustering algorithms can improve sentiment classification accuracy by applying strategies such as sense-based and cross-lingual word clustering by word sense [18]. A conditional random field model and a weakly supervised learning model can be used to label tokens with approximately 90% accuracy [22]. Uses social media data to perform mixed-script language recognition and concludes that supervised learning outperforms dictionary-based approaches [19].

Language detection using multi-class regression classifiers can be accomplished with approximately 54% accuracy.HSWN and negation discourse are employed for sentiment analysis of text corpora in Hindi, with an accuracy of close to 80%. Text normalisation can be performed using methods such as phonetics-based, slang, and spelling correction procedures, according to a different, more limited study [20]. To demonstrate the usage of sentiment analysis techniques on transliterated content, they used bilingual dictionary approaches and HSWN for sentiment score computation, reaching 80% accuracy. Another person utilised the genetic algorithm to recognise Marathi and Sanskrit words [21].

TABLE I
Details of the papers reviewed

Ref	Language	Technique used	Lexicon Type	Dataset Size	Dataset	Result
[22]	Telugu	GMM, SVM, NB (doc2vec)	Movie Song	300 Songs	Telugu movie song	SVM+GMM = 91.2%

[23]	Bengali, Hindi & Tamil	Lexicon based, SV, SVM, LR	Microbogging	SAIL 2015 Tweets Datasets	SAIL 2015 Bengali, Hindi & Tamil Tweets Datasets	Bengali 67.83% Hindi 81.57% Tamil 62.16%
[24]	Tamil	SVM, Maxent classifier, Decision tree & NB	Review Collected form website	534- Tamil Review from website	Tamil Review	SVM 71.91% Maxent classifier 59.55% Decision tree 64.04% NB 66.17%
[1]	Hindi, Marathi	SVM, Random forests	Not Specified	Not Specified	Not Specified	Not Measured
[25]	Hindi	Lexicon based, LMC classifier	Author dataset	Not Specified	Hindi speeches delivered by Leaders	Not Measured
[26]	Tamil, Hindi & Bengali	RNN	Microbogging	SAIL 2015 Tweets Datasets	SAIL 2015 Tweets Datasets Tamil, Hindi & Bengali	Tamil 88.23% Hindi 72.01% Bengali 65.16%
[27]	Bengali, Tamil	NB & C4.5 Decision Tree	Microbogging	999 Bengali tweets & 1103 - Tamil	Bengali & Tamil tweets	Not Measured
[28]	Hindi	Dictionary Based, NB & SVM	Micro bogging	42,235 tweets	Hindi tweets	Dictionary Based - 34% NB – 62.1% SVM – 78.4%

III. APPLICATIONS AND IMPORTANT OF SENTIMENT ANALYSIS

The process of analysing a text to determine the sentiment (positive, negative, or neutral) expressed is known as sentiment analysis. Social media monitoring, market research and analysis, and listening to the voice of the consumer (VoC) are some popular Sentiment Analysis uses.

A. Monitoring of Social Media

Because they are unwelcome, social media posts frequently offer some of the most candid thoughts regarding your products, events, services, organisations, and enterprises.

B. Pay attention to the customer's voice (VoC).

All of your customers' input from the web, customer surveys, chats, call centres, and emails should be evaluated together. Sentiment analysis helps you to categorise and compose this data in order to identify trends and discover reoccurring subjects and worries.

C. Market Analysis and Research

SA is used in Business Intelligence to understand the subjective reasons why customers respond or do not respond to something, whether it is a product, a user experience, or customer assistance.

Sentiment Analysis tools are essential for recognising and analysing client sentiments. Companies can improve their Customer Experience (CX) by employing these ways to learn how their customers feel. SA tools generate insights into how businesses may improve their customer service and experience.

IV. CHALLENGES FOR SENTIMENT ANALYSIS

When we consider sentiment analysis issues, there are a few areas where businesses struggle to attain sentiment analysis accuracy. Analysing emotions or feelings in natural language processing can be difficult because machines, like the human brain, must be trained to assess and understand emotions.

Some of the difficulties that must be overcome while undertaking sentiment analysis are listed below.

1) Tone: Tone is difficult to translate literally and even more difficult to recognise in writing.
2) Polarity: Positive (+1) and negative (-1) polarity scores for words like "love" and "hate" are high.
3) Sarcasm: Irony and sarcasm are used in informal chats and memes on social media
4) Emojis: One of the problems with text-based social media content, such as Twitter, is that it is dense with emojis. Natural Language Processing (NLP) tasks for specific languages are trained.
5) Idioms: Machine learning programmes are incapable of comprehending figures of speech. For example, the algorithm will be perplexed by the statement "not my cup of tea" because it takes things literally.
6) Negations: Negative terms such as not, never, can't, were not, and so on might cause confusion in the Machine Learning model.
7) Comparative sentences: Comparative sentences can be tricky because they don't always give opinions. A lot has to be taken from it.
8) Employee bias: staff input is crucial for developing corporate culture, boosting sales strategies, and lowering staff turnover.
9) Multilingual sentiment analysis: When languages are mixed into multilingual sentiment analysis, all of the issues stated above compound.
10) Audio-Visual Data: Videos and text data are not the same thing. The challenge is that the video must not only be transcribed, but it may also have captions that must be evaluated for brand logos.

V. FRAMEWORK FOR SENTIMENT ANALYSIS

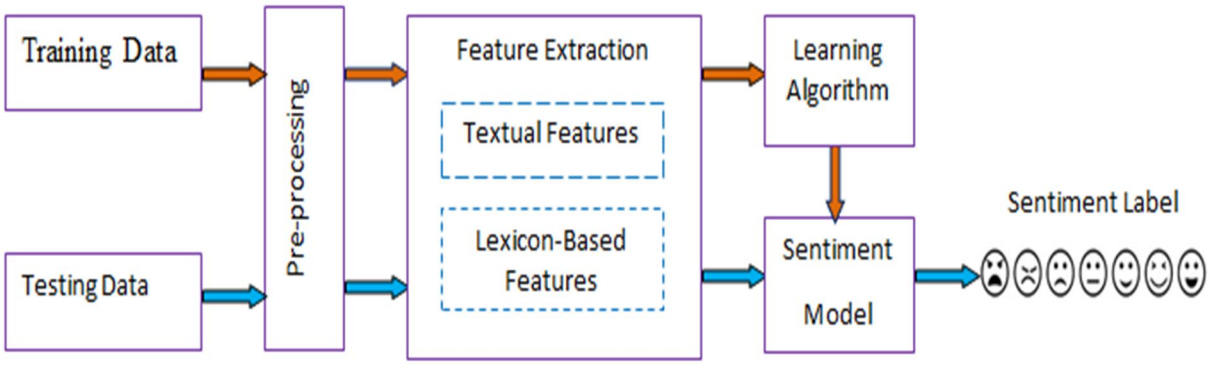


Fig. 1 Sentiment Analysis Framework in General.

VI. PRE-PROCESSING FUNCTIONS THAT CAN PERFORM ON TEXT DATA SUCH A

Data pre-processing is the process of preparing raw data for machine learning (ML) models. This is the first and most crucial stage in developing a machine learning model.

This walks through some of those steps, including Bag-of words (Bow) Model, Stemming, Removing Stop Words, Tokenization, Lemmatization, Displaying Document Vectors, Removing Low-Frequency Words, Distribution of words Across Different sentiment, Creating count vectors for the dataset

VII. CONCLUSION

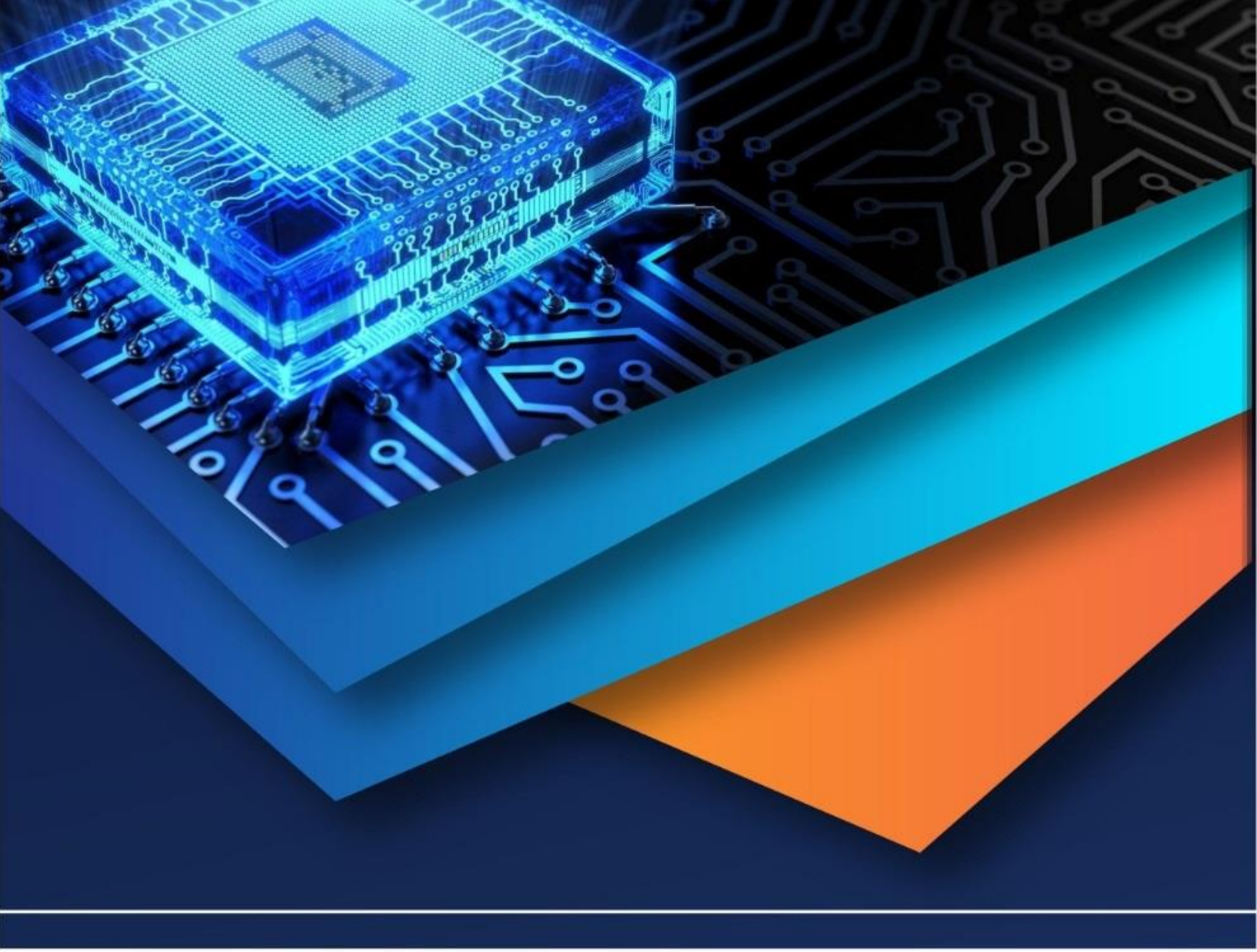
The goal of this review essay is to assess critically recent writing in the area of sentiment analysis with Regional languages. The review takes into account pre-processing, corpus, methodologies, and success rates. The pre-processing processes, such as stop word removal and negation handling strategies, affect how well the SA model performs. When comparing Character level and word level feature representation approaches that leverage the existence of words, Bow, or TF, TF-IDF, and Word2vec demonstrate a considerable improvement. Compared to other classifiers, SVM and RNN produce the best results. SVM performance is poor compared to the corpus's size.

VIII. ACKNOWLEDGEMENT

My respected guide for preparing this paper. I would like to thank Dr. Manasi Ram Baheti Madam for her help and providing infrastructure and experimental facilities to carry out the above work. Hon'ble Prof. Dr. Sachin N. Deshmukh Sir, Head of Department Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad is also deeply thankful and likewise The Dr. Babasaheb Ambedkar Research and Training Institute (BARTI) in Pune is an autonomous institute of the Maharashtra Government's Department of Social Justice and Special Assistance. I am also deeply grateful to him for giving me a scholarship to do research.

REFERENCES

[1] Ansari, J.A.N., Khan, N.A. Exploring the role of social media in collaborative learning the new domain of learning. Smart Learn. Environ. 7, 9 (2020). https://doi.org/10.1186/s40561-020-00118-7
[2] Appel, G., Grewal, L., Hadi, R. et al. The future of social media in marketing. J. of the Acad. Mark. Sci. 48, 79–95 (2020). https://doi.org/10.1007/s11747-019-00695-1
[3] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
[4] Kapoor, K.K., Tamilmani, K., Rana, N.P. et al. Advances in Social Media Research: Past, Present and Future. Inf Syst Front 20, 531–558 (2018). https://doi.org/10.1007/s10796-017-9810-y
[5] Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. International journal of qualitative methods, 16(1), 1609406917733847.
[6] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. Mathematical and Computational Applications, 23(1), 11.
[7] Ansari, M. A., & Govilkar, S. (2018). Sentiment analysis of mixed code for the transliterated hindi and marathi texts. International Journal on Natural Language Computing (IJNLC) Vol, 7.
[8] Bhargava, R., Sharma, Y., & Sharma, S. (2016, September). Sentiment analysis for mixed script indic sentences. In 2016 International conference on advances in computing, communications and informatics (ICACCI) (pp. 524-529). IEEE.
[9] Khan, J., & Lee, S. (2021). Enhancement of Text Analysis Using Context-Aware Normalization of Social Media Informal Text. Applied Sciences, 11(17), 8172.
[10] Vu, L., & Le PhD, T. (2017). A lexicon-based method for Sentiment Analysis using social network data. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE) (pp. 10-16). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
[11] Joshi, A., Balamurali, A. R., & Bhattacharyya, P. (2010). A fall-back strategy for sentiment analysis in hindi: a case study. Proceedings of the 8th ICON.
[12] Rana, S. (2014). Sentiment Analysis for Hindi Text using Fuzzy Logic. Indian Journal of Applied Research, 4(8).
[13] Dhore, M., Dixit, S., & Dhore, R. (2012, December). Optimizing transliteration for Hindi/Marathi to English using only two weights. In Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology (pp. 31-48).
[14] Schroeder, S. R., & Chen, P. (2021). Bilingualism and COVID-19: using a second language during a health crisis. Journal of communication in healthcare, 14(1), 20-30.
[15] Kundu, B., & Chandra, S. (2012, December). Automatic detection of English words in Benglish text: A statistical approach. In 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI) (pp. 1-4). IEEE.
[16] Shelke, R., & Thakore, D. (2020). A novel approach for named entity recognition on Hindi language using residual bilstm network. International Journal on Natural Language Computing (IJNLC), 9(2), 1-8.
[17] Stratos, K., Collins, M., & Hsu, D. (2016). Unsupervised part-of-speech tagging with anchor hidden markov models. Transactions of the Association for Computational Linguistics, 4, 245-257.
[18] Jardim, S., & Mora, C. (2022). Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning. Procedia Computer Science, 196, 199-206.
[19] Babu, N. V., & Kanaga, E. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: A review. SN Computer Science, 3(1), 1-20.
[20] Golmaei, S. N., & Luo, X. (2021, August). DeepNote-GNN: predicting hospital readmission using clinical notes and patient network. In Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 1-9).
[21] Rao, L. (2022). Sentiment Analysis of English Text with Multilevel Features. Scientific Programming, 2022.
[22] Abburi, H., Akkireddy, E. S. A., Gangashetti, S., & Mamidi, R. (2016, January). Multimodal sentiment analysis of telugu songs. In SAAIP@ IJCAI.
[23] Patra, B. G., Das, D., Das, A., & Prasath, R. (2015, December). 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.
[24] Chu, C. T., Takahashi, R., & Wang, P. C. (2005). Classifying the sentiment of movie review data.
[25] Shah, P., Swaminarayan, P., & Patel, M. (2022). Sentiment analysis on film review in Gujarati language using machine learning. International Journal of Electrical and Computer Engineering, 12(1), 1030.
[26] Seshadri, S., Madasamy, A. K., Padannayil, S. K., & Kumar, M. A. (2016). Analyzing sentiment in indian languages micro text using recurrent neural network. IIOAB J, 7, 313-318.
[27] Prasad, S. S., Kumar, J., Prabhakar, D. K., & Tripathi, S. (2016, August). Sentiment mining: An approach for Bengali and Tamil tweets. In 2016 Ninth International Conference on Contemporary Computing (IC3) (pp. 1-4). IEEE.Mangat, V. (2017). Dictionary based Sentiment Analysis of Hinglish text. International Journal of Advanced Research in Computer Science, 8(5).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY