2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)

Recognizing Sentiments In Regional Language Text From Social Media Using Machine Learning Approach : A comparative Study to understand the state of the art tools and associated functionalities available

Sudarshan Sirsat
Department. of Data Science and Technology,
K J Somaiya Institute of Management
Somaiaya Vidyavihar University
Mumbai, India
sudarshan@somaiya.edu

Dr. Nitish Zulpe
College of Computer Science and Information Technology
Swami Ramanand Teerth Marathwada univesity
Latur, India
nitishzulpe@gmail.com

***Abstract*—Regional language contents gain the highest traction in today's social media network marketing and from users' perspectives since nearly 90% of users consume contents in regional language. Those who are only focusing on English as the major language of social media or website contents are clearly missing the majority. Social media currently is touching the untouched digital population of the world through creative contents, easy to use convincing features and infinite possibilities of enhancing personal social networks using regional language like Hindi, Marathi, and Gujarati and many other state and regional languages. This has created immense opportunities in the area of text mining, text analysis and text generation too many scholars and researchers to visualize and analyses the contents to gain insights into what are the multidimensional perspectives of the authors of any such contents. This research work is going to understand the current work done in the area of regional language sentiment analysis in India and all across the globe to understand the status and research gap in the domain to work further in the Maharashtra state regional language contents available on social media pages. This research work will try to explore all possible tools and technologies available and used by the recent research scholars and industry leaders to decide the way ahead to work in the Marathi regional language domain in near future.**

Keywords— Regional Language, Natural Language Processing, Sentiment Analysis, Social Media, Social Media, Marathi Language Sentiment Analysis, Web Scrapping

I. INTRODUCTION

Language plays a very important role in deciding the convenience factor amongst audience, viewers, online blog readers and social media users. Irrespective of media like newspaper, radio, television or internet, people are highly inclined towards their regional language for accessing the contents or putting the contents somewhere. The reasons for the same can be high understanding, convenience to write and speak, getting the exact context of speakers, writers or authors and many other factors which play a direct or indirect role in choosing the language for communication. Many global leaders preferred their national or regional language for passing messages to the world on global platforms like the United Nations and any such global platform during their foreign visits. When it comes to using social media, a huge chunk of the digital population is comfortable and feels convenient to listen, read, write and share the social media contents in their regional language[1]. This leads to use of code mix language or purely regional language fonts for writing and reading any such contents at bulk. From many global and local leaders to the common social media users, the pattern of using regional language contents on a regular basis is found very strongly on most of the social media platforms.

This enormously generated content on social media has many hidden meanings, messages and emotions attached to it by the author or writer even one who is sharing the same trying to pass the emotions forward. These contents have many research opportunities like text mining , text analysis, opinion mining and sentiment analysis which can unfold many text analysis and social media users' aspects and perspectives in the near future. This study is trying to understand any such work done by the Indian and abroad research scholars to know their point of view and understanding about the regional language content usage on social media. This research study will try to investigate and try to understand the technologies available and used by various researchers for their regional language analysis, also the dictionaries and datasets used by them to do the sentiment analysis on regional language text. In the special scenario wherein if datasets have to be prepared then what can be the right approach in preparing a regional language dictionary or dataset for Marathi language text sentiment analysis in specific. This research study is more focused towards identifying the research gap in the specified domain, up till now there is no systematic work carried out in the proposed system, also the algorithms and techniques were not able to achieve high accuracy levels so there is need to develop enhanced models based on automated machine learning techniques that resolves the challenges to fulfil the research gap.

II. RESEARCH METHODOLOGY APPROACHES

A. Sentiment Analysis Approach

Sentiment Analysis(SA) is a subfield of broader area Natural Language Processing (NLP) also sometimes referred as opinion mining, used for extracting features of text data to do further analysis. Features like attitude, opinion, sentiments can be extracted from text data from various digital file storage

2023 International Conference on Advanced Computing Technologies and Applications (ICACTA) | 979-8-3503-4834-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACTA58201.2023.10393104

979-8-3503-4834-7/23/$31.00 © 2023 IEEE

and web page sources like emails, chats messengers, social media posts, reviews given on various products and services online etc. this type data have hidden meanings, opinions, expressions and propaganda by its creator[2].

Sentiment Analysis approaches vary depending on research objectives: Naive Bayes approach, Deep Learning LSTM , Pre trained rule based VADER Models

These sentiment analysis approaches can be used for solving classification problems when data is factored or bivariate, regression problems when precision is expected in the outcomes or result. Sentiment analysis can be used for expressing and interpreting the sentiment polarity and intensity differently from textual data[2].

Polysemy (context) plays a very important role in sentiment analysis since one word or statement may have different meaning in different contexts where it is used or said.

Social media is being explored and exploited across demography be it rural or urban, young or old, elite or novice, professional or amateur giving rise to the opportunity for analysis of the content to better understand the changes that our society is currently witnessing. There are quite a few microblogging sites and social media applications that enable the users to post their contents, comments and reviews about the product and services in the language of their choice be it English, Hindi or any other regional language. The trend to offer reviews by the user on what they buy from e-commerce sites and view over OTT platform is far more increasing, providing immense analytical and research opportunities to the research scholars across the globe. These contents are the sources that can provide insights into content writers attitude, thought processes and can be further worked upon to determine whether the thinking is positive, negative or neutral. The qualitative methods such as textual analysis can further help the researcher understand and interpret what each of the subtext, symbolism, assumptions reveal and in turn what value it creates. The ultimate aim of the analysis is to examine structure, derive new meaning if possible and relate it to the historical or cultural settings under which the content must have been produced. The paper focuses on studying various tools and approaches used for the sentiment analysis of such regional language contents generated through microblogs. The paper also attempts at comparing various tools and techniques available for performing textual analysis.

B. Machine Learning Based Approach

Sentiment analysis models can be developed and trained using machine learning algorithms to identify or detect features beyond definition like context, sarcasm or emotion without human intervention. How sentiment analysis works with machine learning, is demonstrated with a simple diagram below. Supervised machine learning approach can be used to data labelling and tagging which is up to subjectivity and bias present in the text data.

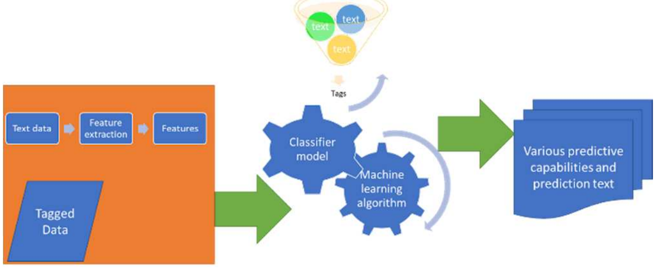


Fig. 1. Sentiment analysis with machine learning approach

The Fig. 1. Depicts the machine learning approach to achieve various predictive capabilities of machine learning algorithm through basic pre-processing.

III. Literature Review

There has been limited work carried out by researchers and scientist in the Marathi regional language text sentiment analysis which will be reviewed systematically and comparatively on the time scale(in the order of years) as mentioned below:

M. K. Patil(2021) et.al, focuses on the opportunities and challenges the Mapathy language provides to the researchers in India being 4th most spoken language in India and 15th most spoken global language. This work is mostly based on review work and context-based sentiment analysis and shaded-based sentiment analysis which is categorical sentiment-corpus based classification problem[3]. S. Pundlik (2016) et.al, Use of mother tongue and state wise regional languages on internet and social media is increasingly reported on various platforms, like giving reviews, passing comments etc. this gives researchers an area to do the analysis with machine learning techniques like natural language processing and sentiment analysis. Analysis may vary with types of contents like classification problems in terms of speech, sentiment analysis in terms of product or service reviews and polarity checking in case of some comments on particular topics. Researchers proposed classifications models using HindiSentiWordNet(HSWN), LM Classifier and a combined approach to get better results[4]. C. Nanda(2018) et.al, use of social media for posting comments and reviews on products, services, movies etc. is largely in trend, though EEnglish is mostly and globally used language for this. In India Hindi is a widely used language to give reviews and share information on social media. This study focuses on sentiment analysis of movie reviews in Hindi language by researchers and also to check the polarity of them with machine learning techniques[5]. V. Yadav(2021) et.al, the scholars focus on the sentiment analysis of product and service reviews by their consumers and customers to get the sense of how these products and services are up to the customer satisfaction marks. Determinative recommendations from the Aspect based sentiment analysis through categorising and analysing the opinionated text from e-commerce portals. This proposed research work focuses on polarity check, sentiment analysis and conflict opinions amongst consumers of the products and services by developing aspect based sentiment analysis models for Hindi language reviews[6]. K. Yadav(2020) et.al, research scholars focused Hindi-English based code language for informal communication on various social media platforms. The challenge is at a different level according to the authors due to sentiment analysis of two language combined approaches by social media users to write the comments, reviews and posts. The code-mix language usage is increasing rapidly due to easy to share opinions for larger groups of social media users and easy to type features which creates the challenges for machine learning based research enthusiasts to clearly identify and analyse the sentiments. Hence conventional sentiment analysis approach will not work here as per researchers, so they used Ensembling based hybrid Naive Bays approach and SVM, Linear regression techniques for the same. They have also used SGD classifiers and a bidirectional LSTM based approach which gave them desired output in their research study[7]. A. Goel(2020) et.al, authors did some basic work of polarity check on Hindi language based text from social media like twitter using sentiment analysis and some deep learning methods, the paper focuses on polarity check only by classifying them on positive negative sentiment polarity[8]. A. Madan(2021) et.al, the research scholar focuses on use of Sentiment Analysis(SA) to get the insights into the opinionated text mined from social media or documents available for the same. Twitter is one of the most microblogging social media websites used majorly for such text mining in recent decades by a number of researchers. The scholars talk about Natural Languages techniques used for basic initial processing of such opinionated text by Lexicon based approach(LBA) and Machine Learning approach(MLA) based on supervised algorithms. Use of HindiSentiWordNet based enhanced dictionary for with lexicon based approach along with Hybrid based Approach(HBA) which combines LBA and MLA to solve the classification problem of movie review tweets in Hindi language. They did the comparative analysis of the LBA and MLA techniques for the same[9]. AlBadani, B(2022) et.al, used Universal Language Model Fine-tuning (ULMFit) and SVM for Twitter based social media text and Twitter Sentiment Detectors (TSD's) for better accuracy and reliable performance for classification techniques and achieved 99.78% accuracy[10].

K. Thapar, (2022) et.al, did Natural Language Processing and sentiment analysis of COVID'19 using a hybrid AI model with social media network dataset. Study also used the bi-directional LSTM, CNN and hybrid model of combining the strengths of both the algorithmic techniques for implementing and achieving the research objectives[11]. M. Divate(2021) et.al, the scholar worked on machine learning based sentiment analysis techniques for Maharashtra state regional and mother tongue language Marathi. When it comes to sharing news online on various social media networks, research scholars focus on helping social media users reading positive news to avoid depression and negativity using sentiment analysis techniques. This study proposes use of filters before sharing any news on social media before posting it on social media using three approaches, sentiment analysis on text, audio, emotions, polarity-based sentiment analysis check and LSTM algorithm of deep learning techniques. The study identified polarity with accuracy of 72% level[12]. S. C, S. Adak(2021) et.al, use of microblog based mined text like text from twitter and various other such websites has increased over recent few years, since the data collected can directly be used for feedback and sentiment analysis by processing it a little bit. This text can be helpful in understanding the features and characteristics of effectiveness of marketing campaigns and its performance over other strategic policies. Subjective detection and sentiment analysis can be such basic two steps, along with machine learning algorithms like Support Vector Machine(SVM), Naive-Bayesian and deep learning etc. through this research work researchers focused on Hindi language based text for classification. The Recurrent Neural Network(RNN) and Convolution Neural Network(CNN) and combined approach has been used by scholars with a Hindi language based subjective data extracted from a dataset made from movie reviews[8]. R. Naukarkar(2021) et.al, in this study scholars focused on polarity checking on Marathi text based tweets through sentiment analysis and machine learning concepts. Tweets are being classified as positive, negative and neutral with higher accuracy outcomes[13].

S. Tammina(2020) et.al, used machine learning and lexicon based mixed sentiment analysis approach using Telugu SentiWordNet dictionary for regional language sentiment analysis using movie review. Aamazon review and microblog text based datasets. Study did subjective classification and TF-IDF function based SVM algorithm for polarity check on the similar datasets[14]. A. Prasad(2020) et.al, worked in the field of information retrieval and text processing which has various application areas like business analytics, politics, social reforms etc. Hindi language is mostly spoken regional language in India which has challenges in the area of sentiment analysis like lack of annotated corpus and lexical resources as compared to the English language. The scholars have developed a semantically and syntactically annotated corpus in the domain of home remedies written in Hindi language weblogs. The corpus can also be useful to other researchers working in the Hindi language opinion mining system in various such domains[15]. P. Bafna(2020) et.al, many websites enable regional language contents to make users understand the information in their mother tongue or regional known language. Hindi is one of the majorly spoken languages in India, especially in the government domain. The classification and clustering operations can be carried out using prosed and verses as input data source. Natural Language Processing and machine learning techniques can be applied over such Hindi language based text using BaSa technique to identify context based common tokens from given corpus. Corpus of 820 proses and 710 verses was processed and scholars observed significant output in the area[16]. S. K Bharti(2017) et.al, to avoid direct negativity and use of intensified positive words to express sarcasm over social media network based opinionated text is one of the upcoming challenges for the sentiment analysts over the last few years. This is more visible on microblogging websites like twitter than other social media or messaging web pages like Facebook, Instagram and WhatsApp, due to easy sharing limited text capability of twitter. The scholars proposed an automated sarcasm detector system for text based data, specifically for regional languages like Hindi, Marathi, Bengali etc. since they are less explored to any such machine learning techniques due to their unique lexicon structure and features. The researchers also proposed context-based pattern finders for "sarcasm as a contradiction between a tweet and the context of its related news" to detect sarcasm out of Hindi language text/tweets and they have achieved 87% accuracy in the same[17]. S. Pawar(2017) et.al, the author focuses on the requirement of sentiment analysis in Marathi language in current scenarios of internet boom and huge textual information availability, the scholar used Support Vector Machine, Naive Bayes and Maximum Entropy machine learning approach for sentiment analysis and classification problem through lexicon based sentiment analysis. The study had faced many challenges namely polarity, negation handling, aspect based sentiment analysis etc. the scholar mentions unavailability of the resources rather than techniques to accomplish the work[18].

Ansari(2016) et.al, textual data available through the advent of social media gives various research opportunities in sentiment analysis and various other textual analysis. The paper is trying to assess the current status, standards and achievements of researchers and their studies. Also scholars trying to propose the provisional improvement techniques and comparative findings. Like direct and simplest approach of using sentiment analysis without NLP or machine learning language specific methods. The study just mentions the Marathi transliterated text and its area of applications in sentiment analysis[19]. V. Jha(2015) et.al, the textual information shared by social media, commercial websites and government web pages can be used for understanding the sentiments they have expressed in textual form. Such textual information can be broadly categorised into opinion and facts and can also be used to make well studied decisions for business process or public policy making. This text based data can also be categorised or classified as subjective or objective, this also can be applied to Hindi language based Unicode standards UTF-8 contents. Hindi language subjectivity analysis(HSAS) is proposed in this study via two methods, first is English language opinion finder subjectivity lexicon and another one is to use a small seed word list of Hindi language and expand it to generate Hindi language subjectivity lexicon. Different techniques were used by scholars to evaluate Hindi language lexicons in which they have achieved 71.4% agreement with human annotators and approximately 80% accuracy in classification on parallel language dataset. The simulation techniques used by researchers validated the test conducted for the same[20].

V. Rohini(2017) et.al, did the domain based sentiment analysis of Kannada language like attitude, opinion detection in the regional language text using machine learning algorithms for polarity check, opinion mining and other basic total sentiment score calculations[21]. R. Naidu(2017) et.al, worked on the sentiment analysis and polarity check of Telagu corpus data and identified the positive, negative and neutral statements and used TeluguSentiWordNet dictionary for the same[22]. AlBadani, B(2022) et.al, used Universal Language Model Fine-tuning (ULMFit) and SVM for Twitter based social media text and Twitter Sentiment Detectors (TSD's) for better accuracy and reliable performance for classification techniques and achieved 99.78% accuracy[23]. E. D. Madyatmadja (2022) et.al, a case study based on Indonesian regional language text explored citizen science with the help of contextual text analytics framework for Indonesian citizen report classification and Context based approach called Con-TOP to minimize various confliction errors. The study focused on the public participation to solve the local governance problems through in the information system domain. The scholars have used various ML algorithms with varying accuracy and precision scores with respect to proposed methodology [24]. J. Wang(2020) et.al, proposed the dimensional sentiment analysis over the focused binary approach which explores only positive and negative aspect of the dataset. The study proposes the use of tree structured regional CNN-SLTM model which consists of regional CNN and LSTM to predict the valence-arousal i.e. VA space scores. The study explored the Lexicon based, regression based, CNN, Structured Representation models as the part of research methodology [25]. U. Naqvi (2021) et.al, the study focuses on the increased used of their regional language Urdu because of social media usage substantial growth and how scholars are considering the exploration of this regional language sentiment analysis for their research work. The scholars proposed and explored the applications of Deep Learning algorithm and methods like CNN, LSTM and hybrid models and approach for Urdu Text Sentiment Analysis [26].

A. Anggraini (2020) et.al, an Indonesian regional language conjunction rule based sentiment analysis study was conducted for analysing the citizens/user complaints about regional drinking water company called PDAM and its water services. Accuracy of model enhanced by 13% when scholars adapted the rule based sentiment analysis. The approach required to build the word based dictionary for implementing the rule based approach [27]. K. T. -K. Phan(2021) et.al, the study explores the zero-shot cross-lingual transfer learning through pre-trained multilingual models like mBERT and XLM-R for implementation of Aspect Based Sentiment Analysis and opinion target expression approach. The experiment was conducted on benchmarked datasets on six languages like English, Russian, Dutch, Spanish, Turkish ad French [28].

A. *State of the Art tools available languages they suppurt and functionalities they provide)*

The three libraries which are currently available for the technical implementation of the regional language sentiment analysis basic implementation are iNLTK, Indic NLP and StandfordNLP[29].

1. **iNLTK**- Hindi, Punjabi, Sanskrit, Gujarati, Kannada, Malayalam, Nepali, Odia, Marathi, Bengali, Tamil, Urdu
2. **Indic NLP Library**- Assamese, Sindhi, Sinhala, Sanskrit, Konkani, Kannada, Telugu,
3. **StanfordNLP**- Many of the above languages

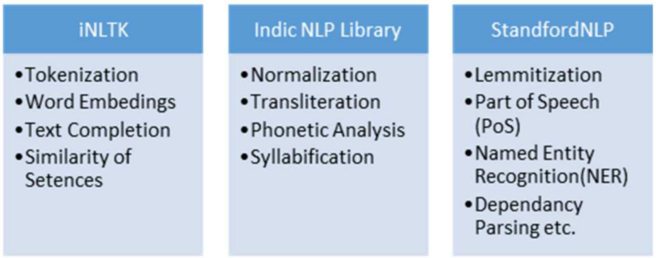


Fig. 2. Tools and functionality they support for regional languges

Fig. 2. Refers to the three different tools available for regional language text analysis and various functionalities they provide, it has been observed that not one tools provides all necessary support for sentiment analysis of regional language text hence we propose the hybrid and combinational approach.

B. *Comparative Literature study in table format*

The Table. 1. shows various techniques used by respective scholars for their respective regional language sentiment analysis studies and issues faced by them or accuracy achieved using the current techniques.

TABLE I. COMPARATIVE STUDY OF DIFFERENT METHODS USED BY VARIOUS SCHOLARS FOR THEIR RESPECTIVE REGIONAL LANGUAGE SENTIMENT ANALYSIS STUDY

Author ,Year	**Techniques**	**Findings**	**Research Gap**
M. K. Patil (2021)[3]	Context based sentiment analysis, Sharded based sentiment analysis	categorical sentiment-corpus based classification problem	Insufficient resources(language dictionary)
S. Pundlik (2016) [4]	NLP,HindiSentiWordNe, LMClassifier	Analysis of types of contents and techniques applicable	

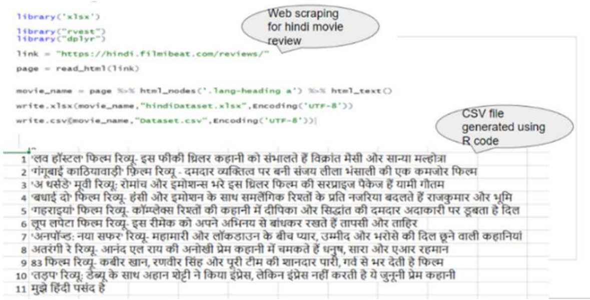
Author ,Year	Techniques	Findings	Research Gap
C. Nanda (2018) [5]	Sentiment analysis, Polarity check	Sentiment analysis of Hindi language based movie reviews	
V. Yadav (2021)[6]	Polarity check, Aspect Based Sentiment Analysis	Determinative Recommendations on product/service reviews	Worked on single line Hindi text
K. Yadav (2020) [7]	Hindi-English code mix language, SVM, SLTM, Stochastic Gradient Descent (SGD)	Sentiment analysis of Hindi-English code mix language	Code-mix language adds challenges of two languages
A Goel (2020) [8]	Hindi-Sentiment analysis, Polarity check	Basic work and concept paper	No outcomes provided
A. Madan (2021)[9]	Lexicon Based Approach (LBA), Machine Learning Approach (MLA), and comparison	Used for basic initial processing of twitter based opinionated text	Worked on Hindi, and sentiment analysis part can be done
AlBadani, B. (2022) [10]	ULMFit and SVM based approach and Twitter Sentiment Detectors	Higher accuracy and performance	Used English text 140 Character limitations of twitter data
K. Thapar (2022) [11]	Hybrid approach using Bi-directional LSTM and CNN	COVID'19 sentiment analysis and polarity check	Sentiment dictionary was not used for detailed sentiment analysis
M. Divate (2021) [12]	Machine learning based sentiment analysis, Polarity check,LSTM	Created filters to avoid sharing fake news in regional language	Accuracy can be increased up to 90% to 95% for reliable results
S. Adak (2021) [13]	Sentiment analysis on Hindi language text, SVM, RNN,CNN	The analysis on effectiveness of marketing campaigns	Focus is on comparative analysis
S. Tammina (2020) [14]	Lexicon based and machine learning based sentiment analysis TF-IDF based approach using SVM, naive Bayes etc. algorithms	Subjective classification and polarity check of Telagu Lang text	Sentiment analysis is not done
A Prasad (2020) [15]	Hindi language based Information retrieval & processing	Developed syntactical and semantically annotated corpus for Hindi language	Corpus generated is annotated so it depends on inputs from users
P. Bafana (2020) [16]	BaSa technique, classification and clustering of proses and verses	Identification of context based common tokens, Corpus of 820 proses and 710 verses was processed	Finding exact context/intentions is the challenge
S. Bharti (2017) [17]	Sentiment analysis, Opinion mining	Proposed automated sarcasm detector system for text data	Finding sarcasm can be dependent on creativity of content creator
S. Pawar (2017) [18]	SVM,Lexicon based sentiment analysis	Technical solutions available	Insufficient resources(language dictionary)
Ansari (2016) [19]	Comparative analysis of techniques available	Assessment of current status, standards and achievements	Focus on Comparative analysis than accuracy
V. Jha (2015) [20]	Hindi language subjectivity analysis(HSAS)	71.4% agreement with human annotators and approximately 80% accuracy in classification on parallel language dataset	Work done on Hindi language subjectivity
R. Naukardar (2021) [30]	Polarity checking on Marathi, Classification problem	High accuracy outcomes on positive negative and neutral words identification	Only polarity check on words

C. Research Gap

Various scholars across India tried working on the regional language majorly Hindi language text for sentiment analysis in various domains like cooking, movie reviews, tweeter based text and its polarity and e-commerce reviews to check polarity analysis. There is hardly any work done in Marathi language based text mining and analysis irrespective of the increased number of Marathi language based social media users, bloggers and authors. No proper datasets are available for any such textual analysis so we can use this as an opportunity to create a lexical based dataset and dictionary for sentiment analysis which can help many Marathi language based researchers in future. There are very few tools available to check the uniqueness of the contents created in Marathi language so this also is one of the gaps we are trying to fill up.

IV. Implementation and Outcomes

Web scraping technique was used to collect data from webpages like filmibeat.com to collect movie reviews.

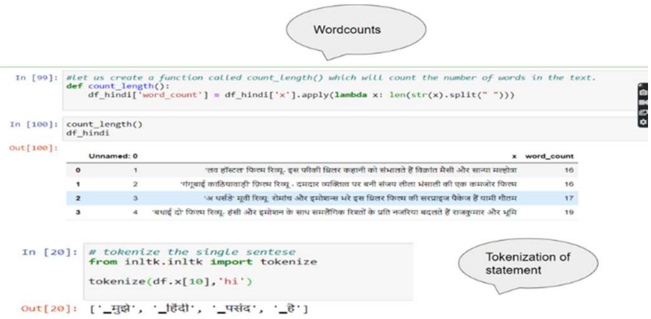


```
library('xlsx')

library("rvest")
library("dplyr")

link = "https://hindi.filmibeat.com/reviews/"

page = read_html(link)

movie_name = page %>% html_nodes('.lang-heading a') %>% html_text()

write.xlsx(movie_name,"hindiDataset.xlsx",Encoding('UTF-8'))

write.csv(movie_name,"Dataset.csv",Encoding('UTF-8'))
```

```
1 'लव हॉस्टल' फिल्म रिव्यू- इस फीकी थ्रिलर कहानी को संभालते हैं विक्रांत मैसी और सान्या मल्होत्रा
2 'गंगूबाई काठियावाड़ी' फ़िल्म रिव्यू - दमदार व्यक्तित्व पर बनी संजय लीला भंसाली की एक कमजोर फिल्म
3 'अ थर्सडे' मूवी रिव्यू: रोमांच और इमोशन्स भरे इस थ्रिलर फिल्म की सरप्राइज पैकेज हैं यामी गौतम
4 'बधाई दो' फिल्म रिव्यू- हंसी और इमोशन के साथ समलैंगिक रिश्तों के प्रति नजरिया बदलते हैं राजकुमार और भूमि
5 'गहराइयां' फिल्म रिव्यू- कॉम्प्लेक्स रिश्तों की कहानी में दीपिका और सिद्धांत की दमदार अदाकारी पर डूबता है दिल
6 लूप लपेटा फिल्म रिव्यू- इस रीमेक को अपने अभिनय से बांधकर रखते हैं तापसी और ताहिर
7 'अनपॉज्ड: नया सफर' रिव्यू- महामारी और लॉकडाउन के बीच प्यार, उम्मीद और भरोसे की दिल छूने वाली कहानियां
8 अतरंगी रे रिव्यू- आनंद एल राय की अनोखी प्रेम कहानी में चमकते हैं धनुष, सारा और एआर रहमान
9 83 फिल्म रिव्यू- कबीर खान, रणवीर सिंह और पूरी टीम की शानदार पारी, गर्व से भर देती है फिल्म
10 'तड़प' रिव्यू: डेब्यू के साथ अहान शेट्टी ने किया इंप्रेस, लेकिन इंप्रेस नहीं करती है ये जुनूनी प्रेम कहानी
11 मुझे हिंदी पसंद है
```

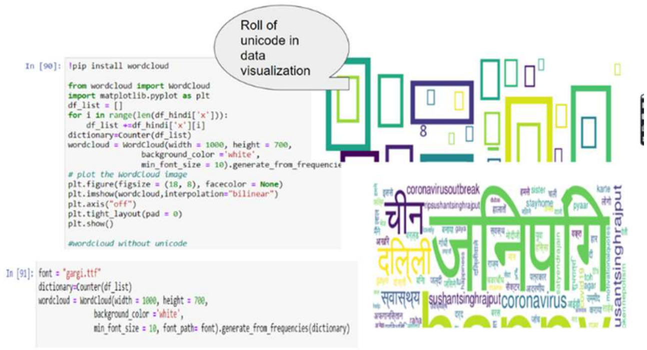
Web scraping for hindi movie review
CSV file generated using R code

Fig. 3. Web scrapping for mining regional language text



```
In [99]: #let us create a function called count_length() which will count the number of words in the text.
         def count_length():
             df_hindi['word_count'] = df_hindi['x'].apply(lambda x: len(str(x).split(" ")))

In [100]: count_length()
          df_hindi

Out[100]:
     Unnamed: 0                                                                                x  word_count
  0           1          'लव हॉस्टल' फिल्म रिव्यू- इस फीकी थ्रिलर कहानी को संभालते हैं विक्रांत मैसी और सान्या मल्होत्रा          16
  1           2         'गंगूबाई काठियावाड़ी' फ़िल्म रिव्यू - दमदार व्यक्तित्व पर बनी संजय लीला भंसाली की एक कमजोर फिल्म          16
  2           3         'अ थर्सडे' मूवी रिव्यू: रोमांच और इमोशन्स भरे इस थ्रिलर फिल्म की सरप्राइज पैकेज हैं यामी गौतम          17
  3           4  'बधाई दो' फिल्म रिव्यू- हंसी और इमोशन के साथ समलैंगिक रिश्तों के प्रति नजरिया बदलते हैं राजकुमार और भूमि          19

In [20]: # tokenize the single sentese
         from inltk.inltk import tokenize

         tokenize(df.x[10],'hi')

Out[20]: ['_मुझे', '_हिंदी', '_पसंद', '_है']
```

Fig. 4. Data cleaning and pre-processing process using python programming language

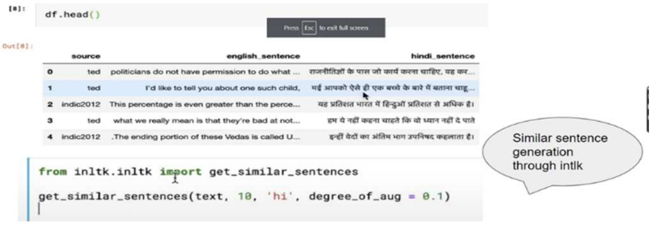


```
In [90]: !pip install wordcloud

         from wordcloud import WordCloud
         import matplotlib.pyplot as plt
         df_list = []
         for i in range(len(df_hindi['x'])):
             df_list +=df_hindi['x'][i]
         dictionary=Counter(df_list)
         wordcloud = WordCloud(width = 1000, height = 700,
                         background_color ='white',
                         min_font_size = 10).generate_from_frequencie
         # plot the WordCloud image
         plt.figure(figsize = (18, 8), facecolor = None)
         plt.imshow(wordcloud,interpolation="bilinear")
         plt.axis("off")
         plt.tight_layout(pad = 0)
         plt.show()

         #wordcloud without unicode

In [91]: font = "gargi.ttf"
         dictionary=Counter(df_list)
         wordcloud = WordCloud(width = 1000, height = 700,
                         background_color ='white',
                         min_font_size = 10, font_path= font).generate_from_frequencies(dictionary)
```

Fig. 5. Role of Unicode character for regional language data visualation

Fig. 3. Demonstrates the regional language text mining/web scraping ad Fig. 4. Shows how cleaning of such web scraped data takes place. Word clouds as shown in Fig. 6. Which was done after data cleaning and pre-processing as shown in Fig. 5. without Unicode support were shown in square like symbols whereas once Unicode support were added to the code, Hindi language words were able to represent properly in word cloud graph.

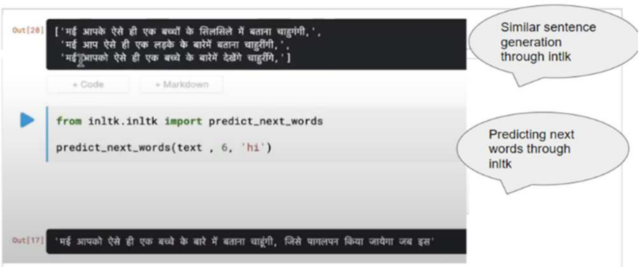
A. Functionalities supported by inltk package for Marathi language:



```
[8]: df.head()

Out[8]:
     source                             english_sentence                              hindi_sentence
  0     ted  politicians do not have permission to do what ...   राजनीतिज्ञों के पास जो कार्य करना चाहिए, वह कर...
  1     ted             I'd like to tell you about one such child,   मई आपको ऐसे ही एक बच्चे के बारे में बताना चाह...
  2  indic2012  This percentage is even greater than the perce...   यह प्रतिशत भारत में हिन्दुओं प्रतिशत से अधिक है।
  3     ted        what we really mean is that they're bad at not...   हम ये नहीं कहना चाहते कि वो ध्यान नहीं दे पाते
  4  indic2012  .The ending portion of these Vedas is called U...   इन्हीं वेदों का अंतिम भाग उपनिषद कहलाता है।

from inltk.inltk import get_similar_sentences

get_similar_sentences(text, 10, 'hi', degree_of_aug = 0.1)
```

Press Esc to exit full screen
Similar sentence generation through intlk

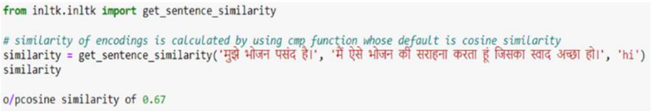
Fig. 6. Simillar sentence generation using iNLTK library



```
Out[28] ['मई आपके ऐसे ही एक बच्चों के सिलसिले में बताना चाहुंगी,',
         'मई आप ऐसे ही एक लड़के के बारेमें बताना चाहुरींगी,',
         'मई आपको ऐसे ही एक बच्चे के बारेमें देखेंगे चाहुरींगे,']

from inltk.inltk import predict_next_words

predict_next_words(text , 6, 'hi')

Out[17] 'मई आपको ऐसे ही एक बच्चे के बारे में बताना चाहूंगी, जिसे पागलपन किया जायेगा जब इस'
```

+ Code
+ Markdown
Similar sentence generation through intlk
Predicting next words through inltk

Fig. 7. Predicting next few(n) words using iNLTK library



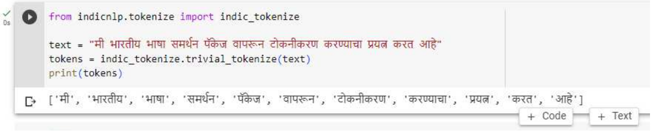
```
from inltk.inltk import get_sentence_similarity

# similarity of encodings is calculated by using cmp function whose default is cosine similarity
similarity = get_sentence_similarity('मुझे भोजन पसंद है।', 'मैं ऐसे भोजन की सराहना करता हूं जिसका स्वाद अच्छा हो।', 'hi')
similarity

o/pcosine similarity of 0.67
```

Fig. 8. Checking sentense simillarity with iNTLK

iNLTK helps you with predicting few words and similar sentences as represented in Fig. 7. And Fig. 8. Respectively.

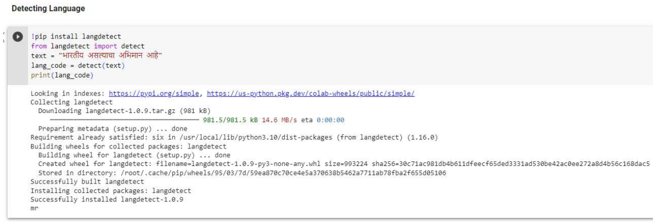
B. Functionalities supported by indicNLP package for Marathi language:

indicNLP supports proper word and sentence tokenization for Marathi regional language text as implanted and depicted in Fig. 9. And Fig. 11. Respectively. Also indicNLP can recognize and identify language code like 'mr' code for Marathi language text as shown in Fig. 10.



```
from indicnlp.tokenize import indic_tokenize

text = "मी भारतीय भाषा समर्थन पॅकेज वापरून टोकनीकरण करण्याचा प्रयत्न करत आहे"
tokens = indic_tokenize.trivial_tokenize(text)
print(tokens)

['मी', 'भारतीय', 'भाषा', 'समर्थन', 'पॅकेज', 'वापरून', 'टोकनीकरण', 'करण्याचा', 'प्रयत्न', 'करत', 'आहे']
```

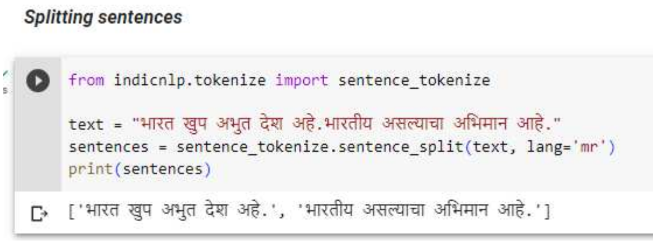
Fig.9. Tokenization of Marathi language sentence with indicNLP package



Detecting Language

```
!pip install langdetect
from langdetect import detect
text = "भारतीय असल्याचा अभिमान आहे"
lang_code = detect(text)
print(lang_code)

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting langdetect
  Downloading langdetect-1.0.9.tar.gz (981 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 981.5/981.5 kB 14.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from langdetect) (1.16.0)
Building wheels for collected packages: langdetect
  Building wheel for langdetect (setup.py) ... done
  Created wheel for langdetect: filename=langdetect-1.0.9-py3-none-any.whl size=993224 sha256=30c71ac981db4b611dfeecf65ded3331ad530be42ac0ee272a8d4b56c168dac5
  Stored in directory: /root/.cache/pip/wheels/95/03/7d/59ea870c70ce4e5a370638b5462a7711ab78fba2f655d05106
Successfully built langdetect
Installing collected packages: langdetect
Successfully installed langdetect-1.0.9
mr
```

Fig. 10. Marathi language text identification with indicNLP

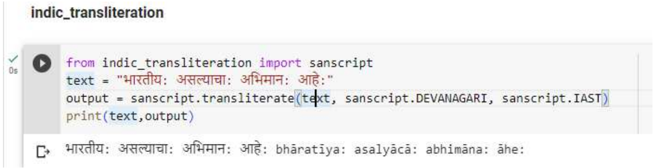


Splitting sentences

```
from indicnlp.tokenize import sentence_tokenize

text = "भारत खुप अभुत देश अहे.भारतीय असल्याचा अभिमान आहे."
sentences = sentence_tokenize.sentence_split(text, lang='mr')
print(sentences)
```

```
['भारत खुप अभुत देश अहे.', 'भारतीय असल्याचा अभिमान आहे.']
```

Fig. 11. Sentence tokenization of Marathi language paragraph with indicNLP



indic_transliteration

```
from indic_transliteration import sanscript
text = "भारतीय: असल्याचा: अभिमान: आहे:"
output = sanscript.transliterate(text, sanscript.DEVANAGARI, sanscript.IAST)
print(text,output)
```

```
भारतीय: असल्याचा: अभिमान: आहे: bhāratiya: asalyācā: abhimāna: āhe:
```

Fig. 12. Transliteration of Marathi language text with indicNLP

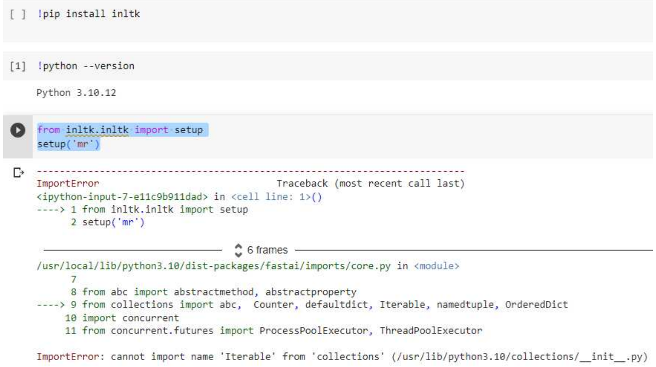
C. Functionalities supported by standfordNLP package for Marathi language:

StanfordNLP right now support hindi as Indian language but do not provide any support for Marathi as regional Indian language as depicted in Figure 15.

V. LIMITATIONS

A. Technical limitations iNLTK

Stop words limitations that is regional language stop words are not fully supported since many words which are removed by the package can add the meaning to the sentence. Punctuations removed need some more preprocessing since underscore looks like symbol which are used for tokenization was not able to be removed by the inbuilt pre-processing function. Installation of Marathi language text takes time and do not provides clear indication that language setup is complete.

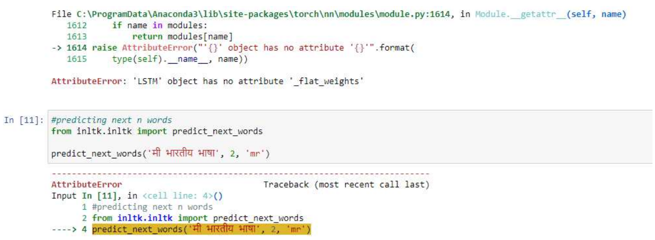


```
[ ]  !pip install inltk

[1]  !python --version

     Python 3.10.12

from inltk.inltk import setup
setup('mr')
```

```
---------------------------------------------------------------------------
ImportError                               Traceback (most recent call last)
<ipython-input-7-e11c9b911dad> in <cell line: 1>()
----> 1 from inltk.inltk import setup
      2 setup('mr')

                    6 frames
/usr/local/lib/python3.10/dist-packages/fastai/imports/core.py in <module>
      7
      8 from abc import abstractmethod, abstractproperty
----> 9 from collections import abc,  Counter, defaultdict, Iterable, namedtuple, OrderedDict
     10 import concurrent
     11 from concurrent.futures import ProcessPoolExecutor, ThreadPoolExecutor

ImportError: cannot import name 'Iterable' from 'collections' (/usr/lib/python3.10/collections/__init__.py)
```

Fig. 13. Inltk and Python 3.10 version dependancy error

Fig. 13. Shows demonstrates that it can not be installed on python version 3.10 and any futhre version whereas it works fine with 3.9 and lower versions.



```
File C:\ProgramData\Anaconda3\lib\site-packages\torch\nn\modules\module.py:1614, in Module.__getattr__(self, name)
   1612     if name in modules:
   1613         return modules[name]
-> 1614 raise AttributeError("'{}' object has no attribute '{}'".format(
   1615     type(self).__name__, name))

AttributeError: 'LSTM' object has no attribute '_flat_weights'
```

```
In [11]: #predicting next n words
         from inltk.inltk import predict_next_words

         predict_next_words('मी भारतीय भाषा', 2, 'mr')
```

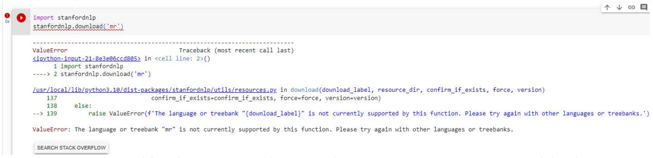
```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Input In [11], in <cell line: 4>()
      1 #predicting next n words
      2 from inltk.inltk import predict_next_words
----> 4 predict_next_words('मी भारतीय भाषा', 2, 'mr')
```

Fig. 14. Inltk package shows AttributeError

Fig. 14 depicts AttributeError for other fucntionality support for Marathi language like 'AttributeError: 'LSTM' object has no attribute '_flat_weights' for functionalities like predict nex t words or getting embedding vectors.

B. Technical limitations standfordNLP

Language support is exists but not fully functional or apt for the technical implementation of regional language preprocessing. There are lack of updates for the package support. Integration with other package of software is difficult. Do not provide similar accuracy for all the supported languages.

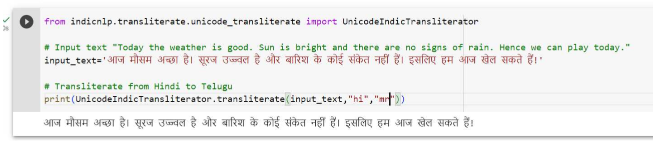


```
import stanfordnlp
stanfordnlp.download('mr')
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-21-8e3e06ccd805> in <cell line: 2>()
      1 import stanfordnlp
----> 2 stanfordnlp.download('mr')

/usr/local/lib/python3.10/dist-packages/stanfordnlp/utils/resources.py in download(download_label, resource_dir, confirm_if_exists, force, version)
    137                     confirm_if_exists=confirm_if_exists, force=force, version=version)
    138     else:
--> 139         raise ValueError(f'The language or treebank "{download_label}" is not currently supported by this function. Please try again with other languages or treebanks.')

ValueError: The language or treebank "mr" is not currently supported by this function. Please try again with other languages or treebanks.
```

Fig. 15. StandfordNLP package doesn't support Marathi language

Technical limitations indicNLP

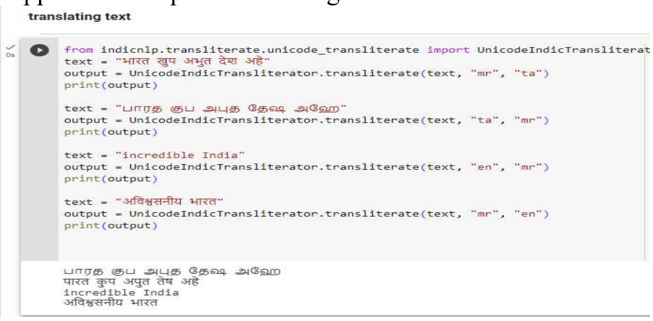


```
from indicnlp.transliterate.unicode_transliterate import UnicodeIndicTransliterator

# Input text "Today the weather is good. Sun is bright and there are no signs of rain. Hence we can play today."
input_text='आज मौसम अच्छा है। सूरज उज्ज्वल है और बारिश के कोई संकेत नहीं हैं। इसलिए हम आज खेल सकते हैं।'

# Transliterate from Hindi to Telugu
print(UnicodeIndicTransliterator.transliterate(input_text,"hi","mr"))
```

```
आज मौसम अच्छा है। सूरज उज्ज्वल है और बारिश के कोई संकेत नहीं हैं। इसलिए हम आज खेल सकते हैं।
```

Fig. 16. Transliteration from Hindi to Marathi is not done and same output for English to Marathi as well

Proper language support is not provided as shown in the Fig. 16. Transliteration from Hindi or English to regional language Marathi is not supported by indicNLP. Tranliteration from Marathi languge text to other languges like Tamil,English and from other languages to Marathi is not supported as respresented in Figure 17.



translating text

```
from indicnlp.transliterate.unicode_transliterate import UnicodeIndicTransliterat
text = "भारत खुप अभुत देश अहे"
output = UnicodeIndicTransliterator.transliterate(text, "mr", "ta")
print(output)

text = "பாரத குப அபுத தேஷ அஹே"
output = UnicodeIndicTransliterator.transliterate(text, "ta", "mr")
print(output)

text = "incredible India"
output = UnicodeIndicTransliterator.transliterate(text, "en", "mr")
print(output)

text = "अविश्वसनीय भारत"
output = UnicodeIndicTransliterator.transliterate(text, "mr", "en")
print(output)
```

```
பாரத குப அபுத தேஷ அஹே
पारत कुप अपुत तेष अहे
incredible India
अविश्वसनीय भारत
```

Fig. 17. Transliteration from Hindi to Marathi is not done and same output for English to Marathi as well.

VI. FUTURE SCOPE

The regional language datasets available in Marathi language currently cannot be considered as state-of-the-art tools for any detailed sentiment analysis hence there is scope for designing our own dataset and sentiment score dictionary. Many researchers did polarity checks and basic text classification on regional languages i.e. Marathi language dataset so we have a gap to fill in order to do the sentiment analysis. In future various predictive analysis and context based sentiment is also possible if proposed basic sentiment analysis gives high accuracy results.

VII. CONCLUSION

By comparing the available tools and technologies, we have found that a lot of work can be done in the area of regional language sentiment analysis for national and state languages (regional languages). There is a high requirement of the regional language open-source datasets for achieving high quality and precision in sentiment analysis and natural language processing domain. Scholars and researchers can contribute towards sentiment dictionaries, regional language words and its sentimental weightage for respective regional languages.

ACKNOWLEDGMENT

We acknowledge The Swami Ramanand Teerth Marathwada University and all the officials for providing us opportunity to work in the area of Natural Language Processing and Sentiment Analysis as a part of PhD work. Special acknowledgement to the College of Computer Science and Information Technology (COCSIT), for providing us with all the necessary infrastructure and motivation required to conduct the research work.

REFERENCES

[1] “Digital news/regional language internet thrives amid challenges” exchange4media www.exchange4media.com/digital-news/regional-language-internet-thrives-amid-challenges-110787.html (accessed May. 11, 2023)

[2] S. Tammina, "A Hybrid Learning approach for Sentiment Classification in Telugu Language," 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), 2020, pp. 1-6, doi: 10.1109/AISP48273.2020.9073109.

[3] M. K. Patil, N. Chaudhari, B. V. Pawar and R. Bhavsar, "Exploring various emotion-shades for Marathi Sentiment Analysis," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544961.

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

[5] C. Nanda, M. Dua and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," 2018 International Conference on Communication and Signal Processing (ICCSP), 2018, pp. 1069-1072, doi: 10.1109/ICCSP.2018.8524223.

[6] V. Yadav, P. Verma and V. Katiyar, "E-Commerce Product Reviews Using Aspect Based Hindi Sentiment Analysis," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-8, doi: 10.1109/ICCCI50826.2021.9402365.

[7] K. Yadav, A. Lamba, D. Gupta, A. Gupta, P. Karmakar and S. Saini, "Bi-LSTM and Ensemble based Bilingual Sentiment Analysis for a Code-mixed Hindi-English Social Media Text," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342241.

[8] A. K. Goel and K. Batra, "A Deep Learning Classification Approach for Short Messages Sentiment Analysis," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1-3, doi: 10.1109/ICSCAN49426.2020.9262430.

[9] A. Madan and U. Ghose, "Sentiment Analysis for Twitter Data in the Hindi Language," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 784-789, doi: 10.1109/Confluence51648.2021.9377142.

[10] AlBadani, B.; Shi, R.; Dong, J. A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. Appl. Syst. Innov. 2022, 5, 13. doi:10.3390/asi5010013

[11] K. Thapar, R. Singh, Z. M. Fadlullah, M. M. Fouda, N. Nasser and A. Ali, "A Hybrid AI Model for Improving COVID-19 Sentiment Analysis in Social Networks," ICC 2022 - IEEE International Conference on Communications, 2022, pp. 1752-1757, doi: 10.1109/ICC45855.2022.9839198.

[12] Divate, M.S. Sentiment analysis of Marathi news using LSTM. Int. j. inf. tecnol. 13, 2069–2074 (2021). doi:10.1007/s41870-021-00702-1

[13] S. C, S. Adak and S. C. Tigga, "Opinionated Text Classification For Hindi Tweets Using Deep Learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1217-1222, doi: 10.1109/ICCMC51019.2021.9418361.

[14] S. Tammina, "A Hybrid Learning approach for Sentiment Classification in Telugu Language," 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), 2020, pp. 1-6, doi: 10.1109/AISP48273.2020.9073109.

[15] A. Prasad and N. Sharma, "Syntactically and Semantically Annotated Hindi Corpus for an Opinion Mining System," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 36-42, doi: 10.1109/ICACCCN51052.2020.9362761.

[16] P. B. Bafna and J. R. Saini, "BaSa: A Technique to Identify Context based Common Tokens for Hindi Verses and Proses," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154124.

[17] S. K. Bharti, K. S. Babu and R. Raman, "Context-based Sarcasm Detection in Hindi Tweets," 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), 2017, pp. 1-6, doi: 10.1109/ICAPR.2017.8593198.

[18] Snehal V. Pawar., & Prof. Swati Mali.(2017), “Sentiment Analysis in Marathi Language”, International Journal on recent and innovative trends in computing and communications (ITCC), Aug 2017, volume 5 issue 8, ISSN: 2321-8169-21-25

[19] Ansari, M. M. A., & Govilkar, S. (2016). Sentiment Analysis of Transliterated Hindi and Marathi Script. In Sixth International Conference on Computational Intelligence and Information (pp. 142-149).

[20] Vandana Jha, Manjunath N, P. D. Shenoy and Venugopal K R, "HSAS: Hindi Subjectivity Analysis System," 2015 Annual IEEE India Conference (INDICON), 2015, pp. 1-6, doi: 10.1109/INDICON.2015.7443824.

[21] V. Rohini, M. Thomas and C. A. Latha, "Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016, pp. 503-507, doi: 10.1109/RTEICT.2016.7807872.

[22] R. Naidu, S. K. Bharti, K. S. Babu and R. K. Mohapatra, "Sentiment analysis using Telugu SentiWordNet," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 666-670, doi: 10.1109/WiSPNET.2017.8299844.

[23] AlBadani, B.; Shi, R.; Dong, J. A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. Appl. Syst. Innov. 2022, 5, 13. doi.org/10.3390/asi5010013

[24] E. D. Madyatmadja, B. N. Yahya and C. Wijaya, "Contextual Text Analytics Framework for Citizen Report Classification: A Case Study Using the Indonesian Language," in IEEE Access, vol. 10, pp. 31432-31444, 2022, doi: 10.1109/ACCESS.2022.3158940.

[25] J. Wang, L. -C. Yu, K. R. Lai and X. Zhang, "Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 581-591, 2020, doi: 10.1109/TASLP.2019.2959251.

[26] U. Naqvi, A. Majid and S. A. Abbas, "UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods," in IEEE Access, vol. 9, pp. 114085-114094, 2021, doi: 10.1109/ACCESS.2021.3104308.

[27] A. Anggraini, E. M. Kusumaningtyas, A. R. Barakbah and M. T. Fiddin Al Islami, "Indonesian Conjunction Rule Based Sentiment Analysis For Service Complaint Regional Water Utility Company Surabaya," 2020 International Electronics Symposium (IES), 2020, pp. 541-548, doi: 10.1109/IES50839.2020.9231772.

[28] K. T. -K. Phan, D. Ngoc Hao, D. V. Thin and N. Luu-Thuy Nguyen, "Exploring Zero-shot Cross-lingual Aspect-based Sentiment Analysis using Pre-trained Multilingual Language Models," 2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2021, pp. 1-6, doi: 10.1109/MAPR53640.2021.9585242.

[29] “3-important-nlp-libraries-indian-languages-python” Analytics Vidhya (accessed May. 11, 2023).

[30] Naukarkar, R. A., & Thakare, A. N. (2021). A Design on Recognization of Sentiment Analysis of Marathi Tweets using Natural Language Processing.