Efficient Dataset Preparation Techniques for Regional/Marathi Language Analysis: Creating Customized Dataset for Regional Language/Marathi Language Text Analysis

Sudashan Sirsat,
Department of Data Science and Technology,
K J Somaiya Insitute of Management,
Somaiya Vidyavihar University
Mumbai, India
sudarshan@somaiya.edu
ORCID: 0000-0001-6043-9814

Nitish Zulpe,
College of Computer Science and Information Technology, Latur, India
nitishzulpe@gmail.com

2023 Somaiya International Conference on Technology and Information Management (SICTIM) | 979-8-3503-3329-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/SICTIM56495.2023.10104666
2023 Somaiya International Conference on Technology and Information Management (SICTIM) | 979-8-3503-3329-9/23/$31.00 ©2023 IEEE |

***Abstract*—Regional language contents are the key to globalization of any successful internet based business model. Looking at the huge population interested in accessing the internet using their mother tongue or regional language is the new normal. This regional language contents on social media and word wide web pages fetched the attention of a large chunk of business analysts, data scientists and social reformists to understand the regional language sentiments through this humongous amount of regional language opinionated text. Regional Language Sentiment Analysis or Marathi language sentiment Analysis will be possible if one can create a dataset which can face text analytics language challenges like uniformity, syntactic and semantic challenges of regional language. This study is a small attempt to create a basic dataset capable of facing future Regional Language Sentiment Analysis or Marathi Language Sentiment Analysis based on NLP and SA based algorithmic approaches. This study will try to generate a Marathi language dataset from social media opinionated text and web scraping of a Marathi language webpage. All the technical issues associated with generating regional language or Marathi language dataset will be recorded, rectified and relatively refined through rigorous iterations to make the dataset future ready Marathi language sentiment analysis. This study will try to understand the needs of Regional Sentiment analysis requirements in terms of dataset, the best suitable file structure and efficient way of creating and customizing the Marathi text dataset in order to make it Natural Language Processing (NLP) and Sentiment Analysis SA ready for future studies in continuation.**

Keywords—Regional Language Sentiment Analysis (RLSA), Regional language Natural Language Processing (RNLP), Natural Language Understanding(NLU), Natural Language Generation (NLG)

I. INTRODUCTION

Regional language is one of the measure factors engaging global and local end users from non-digitized chunk of users to growing social media user networks. Use of local language for reading and writing the social media contents by rural users who are not very well versed with English language, and due increasing to curiosity of updating themselves in their respective interest domain is growing linearly day by day due to language convenience factor.

According to a study 43% of the population i.e. around 10.72 million rural people can be reached provided the regional language content over the internet [1]. India is home of 120 different languages and over 430 million people are availing the internet services yet only 10% of internet using Indian population is well versed with English language and over 70% of Indian population prefer regional/native language content over internet and social media networks[2].

This huge chunk of opinionated text is attracting the major research scholars to perform Natural Language Processing and Sentiment Analysis techniques to understand the various dimensions of regional language sentiments with respect to research. This regional language generated opinionated text is huge unstructured data which can be used for data analysis purposes if performed some data processing. There can be various barriers associated with data pre-processing and cleaning processes due to lexical and semantic challenges of respective regional languages. This study is going to understand the challenges in working with Maharashtra state language i.e. Marathi to perform web scraping and social media scraping. Challenges associated with the cleaning process of regional language, language related complexities like punctuations, similarities between two different languages use the same scripting language example both Hindi and Marathi use Devanagari script to write and read language text.

II. LITERATURE REVIEW

A. Literature Review: English and other Foreign languages

Chinese-Korean Weibo Sentiment Classification Based on Pre-trained Language Model and Transfer Learning, the study proposed the use of pre-trained language model and transfer learning techniques for Chinese-Korean language text from a popular regional social media network Weibo and text posted by Chinese-Korean people on this social media webpages. By web crawling and then labeling the Chinese-Korean text, a dataset was prepared named Chinese-Korean Weibo sentiment Analysis Dataset, Bidirectional Encoder Representations from Transformers (BERT) and other pre-trained language models were applied to do the polarity check [3]. Chinese Paraphrase Dataset and Detection, to handle language diversity as the biggest challenge in the area of Natural Language Processing, the study conducted for identification of paraphrases from Chinese language dataset. It involved trying various word similarity and sentence similarity algorithms; they have concluded that BERT and few other models can help solve the language diversity problem. After using some pre-training language models on Chinese language dataset they have found

that depending on size of dataset performance may vary to some extent [4]. **T**he study focused on stopwords identification and removal techniques for Text Classification (TC) and Information Retrieval (IR) applications. Generic and subjective stopwords are major categories which can be identified with either Rules Based or method based approaches and these stopwords can be removed with various techniques like classic methods and Deterministic Finite Automata (DFA) [13].

B. Literature Review: Indian languages Hindi and other than Marathi

Cross-Lingual Sentiment Analysis for Indian Regional Languages, the study used BRAE and modified BRAE model for cross-lingual polarity check based sentiment analysis and found that modified BRAE model using embedding and cross training performed better compared to plain model. The study was performed on various Indian state languages like Kannada and Marathi. The researchers used HindiSentiNet dictionary structure and transformations techniques to create other regional language dictionaries [5].

The study focused on the regional language Kannada direct dataset, performing regional language sentiment analysis for a particular domain. Research scholars used direct language dataset and machine translation of the English language for analyzing attitude, opinions for domain based sentiment analysis. The study conducted sentiment classification, sentiment calculation and then did comparison of regional language with machine translated language [6]. Modh (2020) et. al. ., the study used the Machine Translation system (MTS) to translate Gujarati (official state language of Gujarat state) language bigram and trigrams into English language. This token sequence of any language creates a different meaning in the context they have used than the literal meaning of the words. One can identify the frequency of these bigram and trigrams in the document and understand the weightage and added meaning of these words/ tokens into the sentiment analysis [16]. Y. Sharma (2015), et. al., the study discusses the previous approaches taken by fellow researchers for sentiment analysis of Hindi Language text like subjective lexicon, N-gram modeling and machine learning approaches. The study mainly performed polarity analysis in terms of positive and negative opinions based on Twitter social media, for which scholars used bi-lingual dictionary, machine translation and use of WordNet scholars performed word count based dominant polarity check [11]. K. V. V. Girish (2016 et. al., the study talks about use of multiple distance bigram features to identify between five Indian languages. Scholars use K-medoids clustering method for identifying the less satisfying outliers from the multiple distance bigram clusters. Clustering is used to identify particular sound sequences [15**]**. R. Naidu (2018), et. al., the paper focused on the unavailability of the proper annotated dataset to perform the sentiment analysis of major Indian regional languages. The study did phrase extraction, sentiment annotation and built the SentiPhraseNet for overcome its sentiment classification problem and also performed the accuracy comparison of various machine learning techniques available for the same. The study found that SPNet worked better than other available techniques in the technical domain [12].

TABLE 1. COMPARATIVE STUDY OF DIFFERENT METHODS USED BY VARIOUS SCHOLARS FOR THEIR DATASET PREPARATION.

Year	Author	Techniques used	Accuracy
2022	Hengxuan Wang	pre-trained language model and transfer learning	Achieved higher accuracy over traditional methods (Top-l accuracy)
2022	N. Rathod	XLM-R Base XLM- R Large	82.5 % 83.82 %
2021	Bo An	Chinese paraphrase dataset and detection (PPD)	Achieved 0.999 accuracy in all three metrics: Accuracy, F1-value and Matthews correlation
2020	D.J. Ladani	text classification (TC) information retrieval (IR)	Average 98% as per the comparative study conducted by authors
2020	J. C. Modh	bigram and trigram translation using machine translation system	Machine Translation System is preferred over google translator
2018	R. Naidu	phrase extraction, sentiment annotation	74% and 81% for subjectivity and sentiment classification respectively
2017	Impana P.	HindiSentiNet dictionary Bilingually Constrained Recursive Auto-encoder (BRAE) model	BRAE gives higher accuracy and removes the necessity of machine translation system
2016	K. V. V. Girish	K-medoids clustering method multiple distance bigram feature, one-vs-one classification	Accuracy varies from 84.29 for Marathi to 95.23 for English and various other languages
2016	Govilkar S	Designed Morphological Analyzer	Up to 96%.
2016	V. Rohini	direct language dataset for Kannada and TF-IDF, decision tree classifier, machine translation	Precision 0.78 for Kannada and 0.86 for English
2015	Y. Sharma	subjective lexicon and N-gram modeling	Accuracy of 73.53 and precision of 0.93

The details of year wise comparative study of various methods used and accuracy achieved by scholars across for regional/Marathi or Hindi language text analysis or sentiment analysis is elaborated in Table 1.

C. Literature Review: Marathi Language

N. Rathod (2022). et. al., the authors did the social media opinion mining with XLM-RoBERTa (XLM-R) models without any machine translation by using the publicly available dataset initially named the publicly available L3CubeMahaSent. The study achieved higher accuracy without using machine translations for small datasets [17]. Govilkar (2016) et. al., the study focuses on identification, extraction and removal of root words from the Devanagari script text and still sustaining the meaning of sentences for performing NLP and sentiment analysis. Stem word and root words were extracted on the basis of filtration of documents and then morphological analysis of Devanagari text. The study concluded that the performance of the proposed morphological analyzer will be depending on the strict rules generation and implementation for removing inflections from the word [14].

III. DATASET PREPARATION APPROACHES

A. Approach 1: social media mining - using social media opinionated text like tweets from twitter

This approach used Tweepy package of python and twitter developer account for this research study purpose and sample dataset was created in csv file format

Algorithmic steps to represent Research Methodology used for technical implementation:

1. Marathi language tweets - using Tweepy and pandas python packages
2. Pre-processing: Remove @username mentions, other #tags, Remove URL's, \n, emoticons and smileys. Fig 1 demonstrates removing of emoji as part of data preprocessing for regional language and Fig 2. Demonstrates removal of URL's if any to enhance the meaning of sentences.

```
def remove_emojis(data):
    emoj = re.compile("["
        u"\U00002700-\U000027BF"  # Dingbats
        u"\U0001F600-\U0001F64F"  # Emoticons
        u"\U00002600-\U000026FF"  # Miscellaneous Symbols
        u"\U0001F300-\U0001F5FF"  # Miscellaneous Symbols And Pictographs
        u"\U0001F900-\U0001F9FF"  # Supplemental Symbols and Pictographs
        u"\U0001FA70-\U0001FAFF"  # Symbols and Pictographs Extended-A
        u"\U0001F680-\U0001F6FF"  # Transport and Map Symbols
```

Fig. 1. Removing emoji's as a part of pre-processing applied on fetched tweets.

```
for tweet in tweets:
    parsed_tweet = {}
    parsed_tweet['author'] = tweet.user.name
    partial_tweet = re.sub(r'@\w+', '', tweet.full_text) #removing mentions
    partial_tweet = remove_urls(partial_tweet) # remove urls
    #remove punctuations
    partial_tweet= partial_tweet.translate(str.maketrans('', '', string.punctuation))
    partial_tweet = re.sub("#[A-Za-z0-9_]+","", partial_tweet) # remove hashtags
    partial_tweet = re.sub(r'"', '', partial_tweet)
    partial_tweet = re.sub(r'“', '', partial_tweet)
    partial_tweet = re.sub(r'\n', ' ', partial_tweet)
    partial_tweet = re.sub(r"'", ' ', partial_tweet)
    partial_tweet= translator.translate(partial_tweet)
    parsed_tweet['text'] = remove_emojis(partial_tweet)
    all_tweets.append(parsed_tweet)
df = pd.DataFrame(all_tweets)
```

Fig. 2. Removing URL mentions and unnecessary hashtags as a part of preprocessing applied on fetched tweets

3. Sentence and word tokenization
4. POS Tagging for regional language
5. Check for Bigrams, trigram and n-gram for Marathi language
6. **.csv files/ data frames with UTF-8 text**
7. Dataset eligible for regional language sentiment analysis

Fig 3. Represents the process of social media mining and data file preparation in .csv file format along with the necessary data cleaning and data preprocessing implementations with the help of programming techniques. UTF-16 or UTF-8 encoding will be enforced to read and write the Marathi language text in preferred file format.

Approach 1: Social Media Mining - Research methodology flowchart

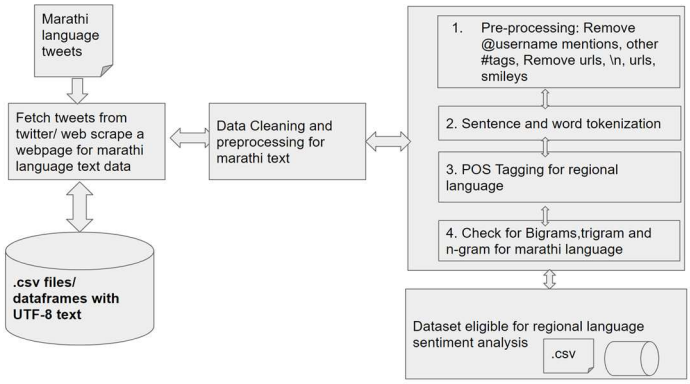

Fig. 3. Research Methodology flowchart for approach 1 - social media mining.

*B. **Approach 2:** web scraping regional/Marathi language website and web page*

This approach will select the random Marathi Language content based website and web page to scrape the Marathi language web contents.

1. select source webpage
   a. here its https://www.sumanasa.com/loksatta/topstories
      Webpage for collecting top news stories
2. use beautiful soup to get Document Object Model(DOM)
3. select html tag from where text can be fetched from

a. Fig 4. Show the use of beautiful soup package to web scrape the div HTML tag from the source URL. Here

its HTML div tag with class “newsheadline col-xs-8 col-sm-9

```
# web source url
req = requests.get('https://www.sumanasa.com/loksatta/topstories').text
# here we get DOM object
soup = BeautifulSoup(req,'html.parser')
# fetch marathi text from div html tag
lp = soup.find_all('div',{ "class" : "newsheadline col-xs-8 col-sm-9" })
```

Fig. 4. Use of beautiful soup for web scraping

4. Using google translator to translate English words to Marathi words. Fig 5 & 6 depicts the differential outputs about before and after using google translator for translating English words to Marathi to enhance meaning of sentences.

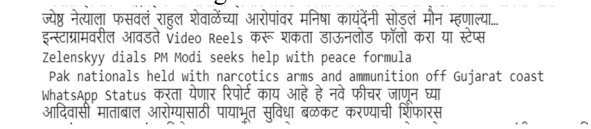

Fig. 5. Before translating English words to Marathi.

Fig. 6. Before translating English words to Marathi.

5. Fig 7 demonstrate the basic and core data pre-processing techniques and two the clean data looks like which is ready for further processing. Below as 3 phases mentioned for the same:
    a. removing punctuations
    b. removing symbols
    c. removing numbers

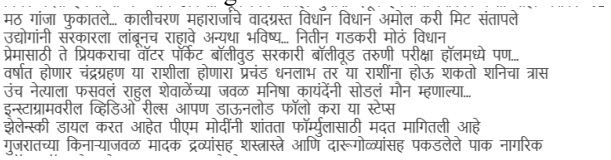

Fig. 7. Preprocessed data after removing impurities.

6. writing csv file of scrapped data
7. Fig 8 depicts the tokenization process and tokenized Marathi language words

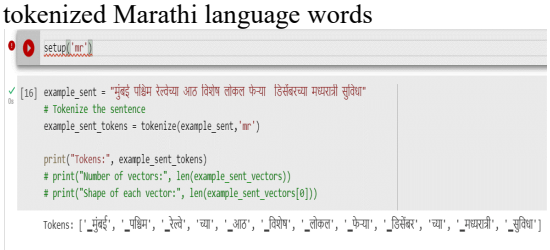

Fig. 8. Tokenization of random Marathi headline web scraped.

9. Bigram and trigram modeling is used to check the sequence of N-tokens of words more frequently used together with each other. These language constructs like Bigram and trigram used to add special meaning to sentences. Bigram are two words, and trigrams are three words used together. Fig 9 is used to represent the output scenario wherein bigram i.e. two words used together frequently to enhance meaning of sentence. And fig 10 demonstrates trigram identification from the same text, i.e. three words used together more frequently that means they have combined contextual meaning together.

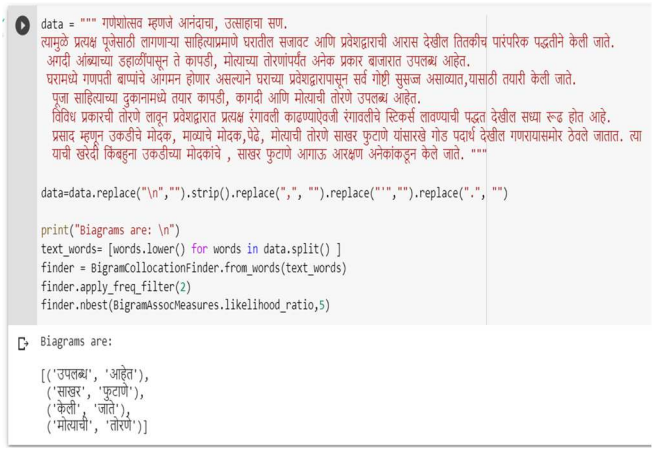

Fig. 9. Pre-processing - finding bigrams from Marathi data.

Fig. 10. Pre-processing - finding trigrams from Marathi data.

Approach 2- Web Scraping - Research methodology flowchart

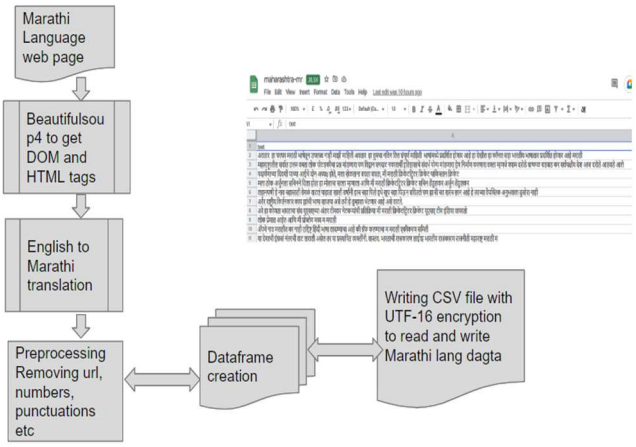

Fig. 11. Research Methodology flowchart for approach 1 - social media mining.

To demonstrate the process of web scrapping approach for dataset preparation the study have created diagrammatic representation which is depicted in Fig 11. This approach too generates the dataset in csv file format with Unicode 17 encoding techniques to easy read and write operations on Marathi language dataset.

Proposal for few Additional Pre-processing Techniques:

With literature review and practical implementation of regional language dataset preparation, our paper we found that traditional pre-processing techniques might not effectively clean and create a dataset ready for further text

classification and text processing like Natural Language Processing. Removing the words assuming that they are meaningless can be a bit risky and may reduce the extent of sentence meaning or intensity of opinion.

The proposed pre-processing techniques involves:
1. Multipoint checked Stopwords**:** to make sure that we are removing only words which are not adding any value to the sentences from the text and the removing only those words that are not hampering the meaning of sentences. The multipoint checking will be involved.

1. The creation of a reinforcement learning based stopwords dictionary.
2. Check against inbuilt Named Entity Recognition based word dictionaries and assigning neutral values to words matching the file contents.
3. Check against stopwords custom dictionary and Instead of removing them we can neutralize the words which hold minimum or no meaning in the sentence.
4. If words are used as bigram or trigram then don't remove them

2. Emoticon to text conversion: for enhanced sentiment analysis of regional language.

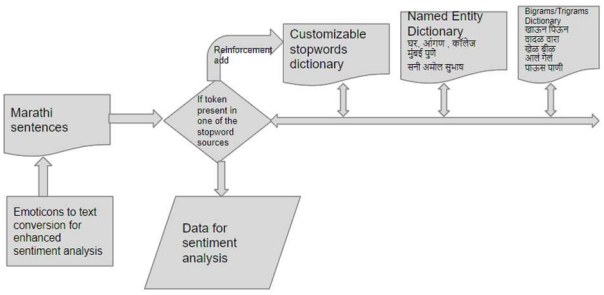

Fig. 12. Proposed efficient data pre-processing approach for enhanced sentiment analysis of Marathi language text.

In future scope of this study, we will create the customized corpus or data dictionary of stopwords along with Named Entity Recognition and pre-defined/inbuilt bigrams and trigrams as data source. Fig 12 represents the process diagram which will be referred for customized dataset preparation for regional/ Marathi language sentiment analysis.

IV. Outcomes And Limitations

To get the most suitable dataset for regional/Marathi language opinionated text, one can either use available resources and frameworks like Tweepy or can go with preparing customized dataset for more accurate results with Regional language Marathi. Data generated from Tweepy will need an authorized twitter developer account and will provide you opinionated text using API keys and access Keys. Challenges faced during the dataset preparations were, removal of punctuations, removal of mentions and emoji's. One of the major challenges still faced by the study is to translate English words to Marathi text to maintain and hold the meaning of sentences in order to perform future sentiment analysis. The most outstanding features of the dataset created using Tweepy and twitter is that it provides you metadata about these tweets like username, place, date time and other details like likes, shares etc.

Data generated with the help of web scraping using internet web pages of regional language/Marathi is data easily available to all the researchers without any authorization key or developer account but comparatively will provide lesser dimensions of data compared to twitter tweets. This data is generated and maintained by the expert content generators and maintained well to give best user experience on their webpages.

A. Limitations

1) Privacy and security: Not all web pages are web scrape enables

2) Few symbols and punctuations are difficult to remove

3) Unavailability of complete stop word dictionary for Marathi language text

a) Create own dictionary: contextual list of words, which have no meaning in the sentences can be removed as stopwords.

b) Use available dataset

c) Combined/Hybrid approach: use both own literature knowledge and available dataset to create stopwords dictionary

B. Future scope

A proper dataset efficient to be used for regional language sentiment analysis will be prepared and analyzed. A rigorous process/algorithm will be developed to generate standard Marathi language dataset. Word and meaning based directory will be formed to do the basic sentiment analysis on these dataset, which will be base for paragraph wise or file sentiment analysis. A customized dataset with inbuilt bigram, trigrams, inbuilt NER would be preferably prepared for Marathi language sentiment analysis.

V. Conclusion

For regional language analysis dataset preparation, with respect to NLP and Sentiment analysis. One can fetch data from two different types of data sources i.e. either from social media webpages like Twitter or from open end regional language web pages like websites of news channels, webpages with microblogs. Both the data sources will need rigorous data cleaning and data pre-processing in order to make it usable for any future text analysis or NLP/SA algorithmic approaches. Writing this data to a file like csv file structure will require UTF-16 encoding to keep it in human readable form. The study also conclude that both the techniques i.e. social media mining and web scraping are equally important and useful for future regional language sentiment analysis, depending on the objective of study for sentiment analysis of regional language Marathi.

ACKNOWLEDGMENT

We acknowledge the twitter developer account services for making text mining easier for researchers, students and scholars. We also acknowledge the web pages that we have scrapped that the text is used for research purposes and we have maintained is a completely unbiased process.

REFERENCES

[1] Impact of regional languages in Social Media, https://www.opendesignsin.com/blog/impact-regional-languages-social-media/

[2] Why Creating Content in Regional Languages is Necessary? https://blog.tagmango.com/why-creating-content-in-regional-languages-is-necessay/

[3] H. Wang, Z. Zhang, X. Cui and R. Cui, "Chinese-Korean Weibo Sentiment Classification Based on Pre-trained Language Model and Transfer Learning," 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI), 2022, pp. 49-54, doi: 10.1109/CCAI55564.2022.9807755.

[4] B. An, "Chinese Paraphrase Dataset and Detection," 2021 International Conference on Asian Language Processing (IALP), 2021, pp. 235-240, doi: 10.1109/IALP54817.2021.9675232.

[5] P. Impana and J. S. Kallimani, "Cross-lingual sentiment analysis for Indian regional languages," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017, pp. 1-6, doi: 10.1109/ICEECCOT.2017.8284625.

[6] 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.

[7] R. G, S. W, A. M. Bharadwaj and C. H N, "Kannada ImageNet: A Dataset for Image Classification in Kannada," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-4, doi: 10.1109/ICCCI50826.2021.9402356.

[8] P. Awatramani, R. Daware, H. Chouhan, A. Vaswani and S. Khedkar, "Sentiment Analysis of Mixed-Case Language using Natural Language Processing," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 651-658, doi: 10.1109/ICIRCA51532.2021.9544554.

[9] J. Patel, K. Makvana and P. Shah, "Cross-lingual Information Retrieval: application and Challenges for Indian Languages," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033563.

[10] Marathi language text web scrapped from https://www.sumanasa.com/loksatta/topstories

[11] Y. Sharma, V. Mangat and M. Kaur, "A practical approach to Sentiment Analysis of hindi tweets," 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 2015, pp. 677-680, doi: 10.1109/NGCT.2015.7375207.

[12] R. Naidu, S. K. Bharti and K. Sathya Babu, "Building SentiPhraseNet for Sentiment Analysis in Telugu," 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 2018, pp. 1-6, doi: 10.1109/INDICON45594.2018.8987162.

[13] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 466-472, doi: 10.1109/ICACCS48705.2020.9074166.

[14] Govilkar, Sharvari & Bakal, J. & Kulkarni, Sagar. (2016). Extraction of Root Words using Morphological Analyzer for Devanagari Script. International Journal of Information Technology and Computer Science. 8. 33-39. 10.5815/ijitcs.2016.01.04.

[15] K. V. V. Girish, V. Vijai and A. G. Ramakrishnan, "Relationship between spoken Indian languages by clustering of long distance bigram features of speech," 2016 IEEE Annual India Conference (INDICON), Bangalore, India, 2016, pp. 1-6, doi: 10.1109/INDICON.2016.7839074.

[16] J. C. Modh and J. R. Saini, "Context Based MTS for Translating Gujarati Trigram and Bigram Idioms to English," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154112.

[17] N. Rathod, N. Mistry, D. Talati, M. Parikh, A. Kore and P. Kanani, "Marathi Social Media Opinion Mining using XLM-R," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 730-736, doi: 10.1109/ICAAIC53929.2022.9793308.