Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Recognizing Sentiments of Maharashtra State Regional Language Marathi Text from Social Media Using Machine Learning techniques

***[1]Sudarshan Sirsat**
Assistant Professor, Department of Data Science and Technology, K J Somaiya Institute of Management, Somaiya Vidyavihar University, Mumbai, India
sudarshan@somaiya.edu
[2]Dr. Nitish Zulpe
Principal, Research Guide, College of Computer Science and Information Technology, Latur
nitishzulpe@gmail.com

Abstract - In the era of digital transformation, where every multinational business organisation, governing bodies and governments are trying to connect the world via a common thread. They are also trying to make their presence glocal i.e. global and local at the same time. For this one has to adapt to regional language and its untouched population, this untouched population is majorly interested in all the relevant and updated contents in their first or second language. With this understanding their sentiments through regional language feedback, reviews and their social media opinion will play a key role in widening the business scope in the regional language geographical demographics. This study is trying to recognize and understand the regional language sentiments through a traditional approach wherein we are considering language specific stopwords, bigrams, trigram and phrases. These language specific semantic and syntactic constraints and contents contribute significantly towards recognising regional language sentiments since every language has different ways of conveying sentiments like happiness, sadness or sarcasm for that matter. The regional language is significantly modelling the way business processes operate and the way standard processes are defined due to its regional stakeholders breaking the language barrier. Regional language sentiment analysis can transform the business processes beyond just digital transformation, rather it will transform them globally and omnidirectionally through regional transformation processes. This study is focusing on how regional language sentiment analysis will contribute in today's era where Large Language Models and Generative AI are dominating the internet based products and services. This work is also trying to implement the concepts like role of bigrams, trigrams and phrases in the regional language sentiment analysis, and aims to implement new terms in Marathi language sentiment analysis like Named Entity Recognition.

Keywords: Regional language, sentiment analysis, marathi text sentiment analysis, Natural Language Processing, Regional language sentiment analysis (RLSA), Regional Language sentiment analysis based Approach (RLSAA), Point of Dataset Generation (PoDSG), Regional Transformation, Regional Transformation Process (RTP), Marathi Language Named Entity Recognition (MLNER)

Introduction

Amongst 7151 languages are spoken worldwide like English, Mandarin, Hindi and Spanish, the Largest chunk of the population mostly prefer their first language; approximately 3.3 million people are bilingual. Marathi, the regional language of Maharashtra state of India, being 11th on the list of top first languages used by people around the globe, Currently 83.1 million speakers own it as first language, 16 million people prefer it as second language and 99.1 million total speakers can make use of it as mode of verbal communication[4]. While using the internet services and deciding the contents they will search for over internet or social media pages, the major chunk of regional stakeholders that is 72% Marathi speakers preferes regional language contents over the internet, 43% entertainment, 35% sports and 27% search for political content. This reflects strong inclination and interest of major regional language natives i.e. Marathi internet users are more interested in their mother tongue and geographical based news contents[5].

Government of India, Maharashtra State government and all the other regional state governments are emphasising on the use or regional language while doing business, providing ease of doing business in the region and serving all the regional stakeholders via their preferred regional choices. The regional choices can broadly categorised into their life styles, standards, the way business transactions takes place, the education

system they opt for or platforms through which they can avail government and private services. The regional choices mainly and specifically entitles them to the language they speak and the language they prefer to read the contents in hardcopy and soft copy materials like newspapers, entertainment and news channels. This also covers the social media platforms and their services over the internet to these regional stakeholders. Recently the Government of India, Maharashtra government and other regional governments endorsed and started the professional educational courses like MBBS and Engineering in official state and national language[6]. All the business organisations are trying to make their omnipresence in the local and global market through digital transformations and entering into regional markets. This gives us the opportunity to understand the regional emotions and sentiments of regional stakeholders via algorithmic and machine learning approach, which can be used for betterment of doing businesses, providing services to them and adding them to the mainstream development process.

Large Language models and Generative AI are changing the way IT industry and E-word is transforming exponentially, in this context regional languages play an extremely important role in reshaping the way business is done. Most of the product and service based companies are promising their services in regional language to merge this undigitized chunk of customers and consumers to mainstream businesses. Services like English to regional language messages on whatsapp messenger or live call translation between two distinctly unknown and language dependent entities is made possible. This global lo regional vice versa translation services open new horizons for the regional language sentiment identification and analysis. In the following paragraphs we are trying to highlight how regional language is aiming to play an extremely important role in reshaping businesses in the upcoming digital era.

Role of Regional Language in AI and Generative AI context

Inclusive Access: Embracing regional languages in AI ensures that technology is accessible to a wider population. Many people are more comfortable communicating in their native language rather than in English or other widely-used languages. By supporting regional languages, AI systems can cater to the linguistic diversity of users.

Increased Adoption: Incorporating regional languages in AI applications can lead to increased adoption and usage. People who speak regional languages may be more inclined to engage with AI-driven products and services if they are available in their preferred language, leading to broader acceptance of AI technologies.

Market Expansion: By supporting regional languages, AI companies can tap into new markets and demographics that were previously underserved. This expansion can lead to business growth and opportunities in regions where regional languages are predominant.

Cultural Preservation: Regional languages are an integral part of a region's cultural heritage. Integrating these languages into AI technologies helps preserve and promote linguistic diversity, cultural identity, and heritage. It also allows for the development of applications and content that are culturally relevant to specific communities.

Personalization and Contextualization: Language is deeply intertwined with culture and context. Incorporating regional languages enables AI systems to better understand and respond to users' needs, preferences, and cultural nuances. This facilitates personalization and contextualization of AI-driven experiences, leading to more relevant and engaging interactions.

Better Communication: Generative AI, such as natural language processing (NLP) models, can benefit from training on regional languages. These models can generate content, translations, or responses in regional languages, facilitating better communication and understanding across language barriers.

Research and Innovation: Research in AI and generative AI that focuses on regional languages fosters innovation and advances in the field. Developing models and algorithms that can effectively handle the complexities of regional languages contributes to the broader scientific knowledge base and opens up new avenues for research.

Social and Economic Impact: Empowering regional languages in AI can have significant social and economic impacts. It can help bridge the digital divide, promote literacy and education, empower local communities, and contribute to economic development by enabling participation in the digital economy.

Overall, the importance of regional languages in AI and generative AI lies in their potential to make technology more inclusive, accessible, and culturally relevant, while also driving innovation and socio-economic development. In the coming era regional language sentiment will play a major business-deriving role to cater to

the undigitized chunk of world population. One can not sustain without embarrassing the regional language as part of their business growth policies. Whether it's translation, transformation or understanding the business processes globally or locally one has to be multilingual in a real sense to adapt to modern business goals.

Literature Review

This study focuses on sentiment analysis conducted in Sundanese language with pre-trained multilingua language models. The effectiveness of the model and approach can be investigated and analysed through sentiment analysis achieved through customised dataset in their regional language. The study mainly focused on positive, negative and neutral sentiment analysis or polarity check with fine tuning the model for available and customised dataset. The overall learning rate for a multilingual pre-trained language model is determined to be 1e-5 and amongst the models evaluated XLM-Tw performed best with F1-score 87% surpassing other models. XLM-R performed at 77% while mBERT with 71% f1-score, indoBERT-p1 and sundaneseRoBERTa scored 77% and 68% f1-scores respectively[1].

The study delves into sentiment analysis for code-mixed urdu regional language, for discerning opinions and attitudes expressed in the text, which can be pivotal for decision making by individuals , product and service based business or social organisations. One for the major challenge faced by researchers, the proliferation of informal and noisy social media content specially in the regional languages. Also challenges associated NLP tools techniques associated with doing sentiment analysis for such unresourceful regional languages like Urdu. The study applied language identification and sentiment analysis through deep learning techniques for English-Urdu code-mixed text with AL ML based Natural Language Processing and sentiment analysis. With character based embedding of Artificial Neural Network and Long Short-Term Memory (LSTM) approach the word based embeddings became baseline for sentiment classification. Both the approaches came up with promising results and efficiency in handling the challenges posed by code-mixed text effectively. Neural network based language identification model worked with 75.10% accuracy and LSTM based sentiment analysis was 72.14% accurate[2].

The widespread use of social media and empowered internet users changed the digital era with more creativity and interactiveness through regional languages. Code-mixing enables the glocal or global to local transitions easily for the internet population which poses the challenges like language processing and pre-processing (data cleaning) because of its uniqueness and unique language script and structure. The study examined the collection and annotation of tree corpora of code-mixed Indian social media text, which includes English-Bengali Twitter messages, English-hindi twit’s and facebook data. The study provides statistical insights into these corpora, roughly discusses part-of-speech tagging using coarse-grained and fine-grained tag sets and evaluates their complexity using a code-mixed Index with compare to other code-mixed corpora. Model training on a 60:20:20 of training, test and validation datasets. Overall deep learning based models outperformed traditional machine learning based models in the experimental setup and custom embedding over pre-trained embedding [3].

Analysing sentiments of Indian Regional Language is challenging due to its language specific constraints like huge data sources, appropriate selection of approaches and machine learning algorithms. The study is laid on the foundation of twitter based social media scraping and applying polarity checks on the basis of its positive, negative and neutral lexical analysis. Using textblob for this sentiment and subjectivity analysis is done with Natural Language Processing on real time customer reviews on various products and services from famous text based social media twitter. According to the researchers this type of work may lead to better product and service selection by the peer consumers[23].

Text communication on web-based platforms is the major breakthrough in the last few decades. Understanding the emotions behind the text beyond its sentiments is equally important for the business organisations and aggregators for the better decision making from various dimensions of business processes. The scholars adapted the systematic approach through data cleaning with normalisation, POS tagging, stemming and lemmatization whereas the feature extraction is done with bag of words, word embedding, TF IDF, N gram. The machine and deep learning models were trained and then evaluated for its accuracy. With overall hybrid approach study outcomes with a conclusion that, dictionary based approach is more compatible to the problem statement as compared to the lexicon based or corpus based approach. Though the study was conducted on English language text but gave us the direction for adapting a dictionary based approach for the regional language sentiment analysis [24].

The article based study talks about the comprehensive French lexicons designed and customised for the french language sentiment analysis, FEEL: a French Expanded Emotion Lexicons, was developed though semi-automatic translation and expansion of the English NRC Word Emotion Lexicon (NRC-EmoLex). The study involves the automated translation and human verification method which achieved high accuracy in identifying and categorising the French language emotions(HAL-LIRMM). The climate based training and testing datasets were evaluated on the 18 different regional language emotions on logarithmic scale, also it considers the polarity, sentiment analysis and emotion detection of the French text beyond just positive and negative sentiments. For short documents like tweets classification is apt while for larger documents emotion detection and categorization is more useful [25].

The study conducted on extensive examination and extraction of sentiments from Hindi language text, the scholars discusses the various challenges and issues faced due to a language specifics like rich morphological base and diverse set of linguistic constructs. The evolution which derived the Natural Language Processing and its customization for the Indian regional languages like SentiNetWord dictionary adoption, handling the negation for sentiment analysis of Indian languages, and adoption of machine learning and deep learning models like BERT and LSTM networks. Furthermore the study emphasised on the need for better language resources and annotated datasets to improve the accuracy and reliability of the Hindi language sentiment analysis. Gazing the more granularity of the emotions beyond just sentiment analysis can be achieved with Aspect Based Sentiment Analysis(ABSA) [26].

A regional policy making process of Brussel was evaluated on sentiment analysis scales for understanding the people's perspectives and emotions on decision making process and overall decisions. The scholars utilised the advanced Natural Language models to analyse public sentiments on various urban policies and infrastructure projects. Generative AI and machine learning models were used for sentiment analysis like XLM-T, GPT 3.5 Turbo, and GPT-4 on the regional mobility sentiment analysis task. GPT-4 for implicit sentiment aspects, GPT3.5 and 4 for polarity analysis(positive, negative and neutral sentiments) whereas XLM-T was used for the classification of the sentiments of the same tweets. GPT removes the need for time consuming training of the datasets , GPT-4 gives more accuracy,XLM-T obtained best results for the pre-trained models but only for the domain specific datasets [27].

The study focuses on various aspects of the sentiment analysis of the various Indian languages through different approaches, challenges due to inefficient language resources and annotated datasets specific to Indian languages. The scholars raised the concern about not much work done in the Indian regional languages like Gujrati, Marathi, Konkani etc, in contrast there are huge internet based users creating big data resources through their blogs, feedbacks, video logging and through many other mechanisms. Processing these regional language-formed datasets can give new dimensions to the way businesses can be expanded and outreached to the fathers corner of the internet regions[28].

A Marathi language based study focusing on the Aspect Oriented Sentiment Analysis (AOSA) with a methodology focuses on the Marathi language movie reviews dataset, with aspect oriented sentiment analysis and lexicon-based approach. The approach uses the annotation process on the marathi language reviews dataset from various aspects of the movies like acting, direction, music etc. Aspect-Identification and Aspect-Sentiment pairing were the steps considered for the sentiment analysis which gave the scholars reliable F1 scores. The major obstacles found while conducting the study was lack of standard language resources and unavailability of the language specific sentiment analysis or emotion detection measures like lexicons, parsers, taggers and corpora[29].

Existing Literature on Sentiment Analysis in Social Media Texts

Numerous studies have explored sentiment analysis in social media , providing insights into different methodologies and challenges. Reference [7] extensively examined methods for opinion mining, emphasising the relevance of sentiment analysis in comprehending discussions on social media platforms. Similarly, Reference [8] offered a comprehensive review of sentiment analysis techniques, particularly focusing on their utilisation in interpreting sentiments expressed across social media platforms. Reference [9] specifically addressed sentiment analysis in Twitter data, employing lightweight discourse analysis techniques for sentiment classification. Furthermore, Reference [10] conducted a survey on approaches for sentiment analysis and opinion mining, discussing a variety of techniques employed for sentiment analysis in social media texts.

Overview of Studies Focusing on Sentiment Analysis in Vernacular Languages

Several studies have delved into sentiment analysis specifically within vernacular languages, offering insights into the unique challenges and methodologies involved. For instance, Sharma and Das (2018) conducted a study focusing on sentiment analysis in Hindi social media texts, highlighting the importance of considering linguistic nuances and cultural contexts in sentiment classification tasks[13]. Similarly, Chen et al. (2019) explored sentiment analysis techniques tailored for Mandarin social media content, emphasising the need for language-specific approaches to handle linguistic variations and informal language usage[14]. Additionally, Rodriguez and Torres (2020) investigated sentiment analysis in Spanish tweets, employing machine learning models to analyse sentiments expressed in colloquial language[15]. These studies underscore the importance of adapting sentiment analysis techniques to vernacular languages to achieve accurate and culturally sensitive sentiment classification.

Comparative Analysis of Machine Learning Approaches and strategies in Sentiment Analysis

This subsection offers a comparative examination of machine learning approaches and techniques applied in earlier studies on sentiment analysis, encompassing both general sentiment analysis and specific analyses of vernacular languages. It evaluates the efficacy and constraints of diverse machine learning algorithms, comprising supervised learning methods such as support vector machines and neural networks, alongside unsupervised learning techniques like clustering and topic modelling. Additionally, hybrid approaches are explored to gauge their effectiveness in sentiment analysis tasks.The comparison considers factors such as model performance, scalability, interpretability, and computational efficiency, with a focus on identifying the most suitable approaches for sentiment analysis in vernacular social media texts.

Research Methodology

Sentiment analysis in the Marathi language involves analysing text written in Marathi to determine the sentiment expressed, whether it's positive, negative, or neutral. Here's an overview of the typical steps involved in sentiment analysis, specifically tailored to the Marathi language:

Data Collection: Gather a dataset of Marathi text data. This data can come from various sources such as social media, news articles, customer reviews, or any other text data available in Marathi.

Data Preprocessing: Clean the text data by removing noise, such as special characters, punctuation, URLs, and unnecessary whitespace. You may also need to handle issues like word normalisation (e.g., converting different forms of the same word to a common base form) and removing stopwords (commonly occurring words that typically don't carry much sentiment).

Tokenization: Split the text data into individual words or tokens. In Marathi, this involves breaking down the text into meaningful linguistic units, which can be challenging due to the language's complex morphology.

Feature Extraction: Convert the tokenized text into numerical or vector representations that machine learning algorithms can work with. Techniques such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings like Word2Vec or FastText can be used for this purpose.

Sentiment Classification: Train a sentiment classification model using machine learning or deep learning techniques. Common algorithms include Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and neural networks like Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs). The model should be trained on labelled data where each text is associated with its corresponding sentiment label (positive, negative, or neutral).

Evaluation: Evaluate the performance of the sentiment classification model using appropriate metrics such as accuracy, precision, recall, and F1-score. This step helps assess how well the model generalises to unseen data and whether it effectively captures sentiment in Marathi text.

Deployment: Deploy the trained sentiment analysis model to analyse new Marathi text data and infer the sentiment expressed in it. This can be done through integration with applications, APIs, or other software systems where sentiment analysis functionality is required.

Fine-Tuning and Iteration: Continuously fine-tune the model based on feedback and new data to improve its performance and adapt it to evolving language usage patterns.

Sentiment analysis in Marathi presents unique challenges due to the language's morphology, limited availability of resources such as labelled datasets and pre-trained language models, and the need for specialised linguistic processing techniques. However, with the right approach and resources, effective sentiment analysis models can be developed to analyse sentiment in Marathi text data.

Sentiment analysis with emotion-based sentiment analysis approach:

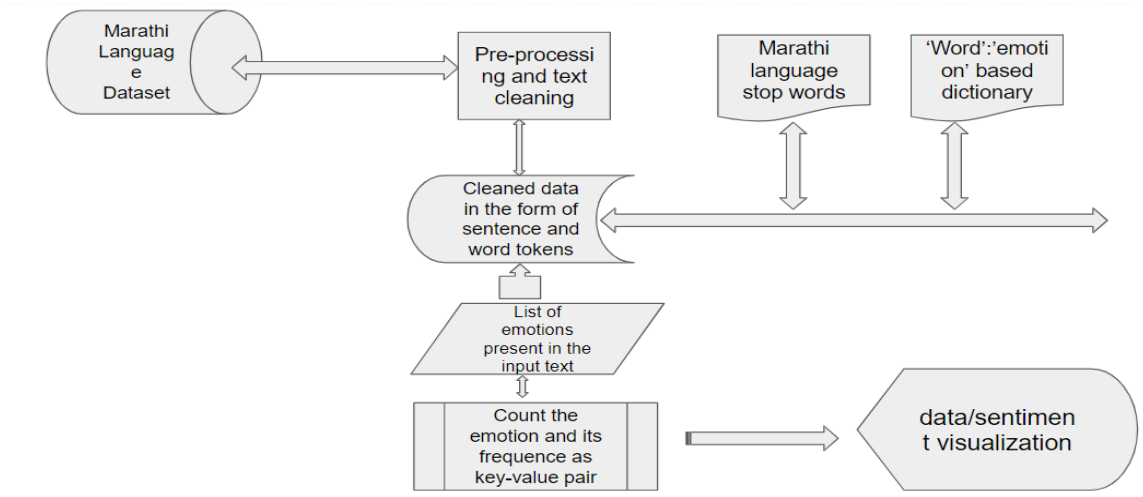


Fig. 1: Sentiment Analysis approach 1- word:emotion dictionary based sentiment analysis and emotion detection

The marathi language dataset is created through web scraping of social media and some microblogging web pages. The emotion based sentiment analysis approach as represented in figure 1 contains key value pairs as in word and its associated emotion as represented in figure 2, the approach checks whether the tokens as in words present in emotion based dictionary and calculates the cumulative sentiments or emotion of the text.

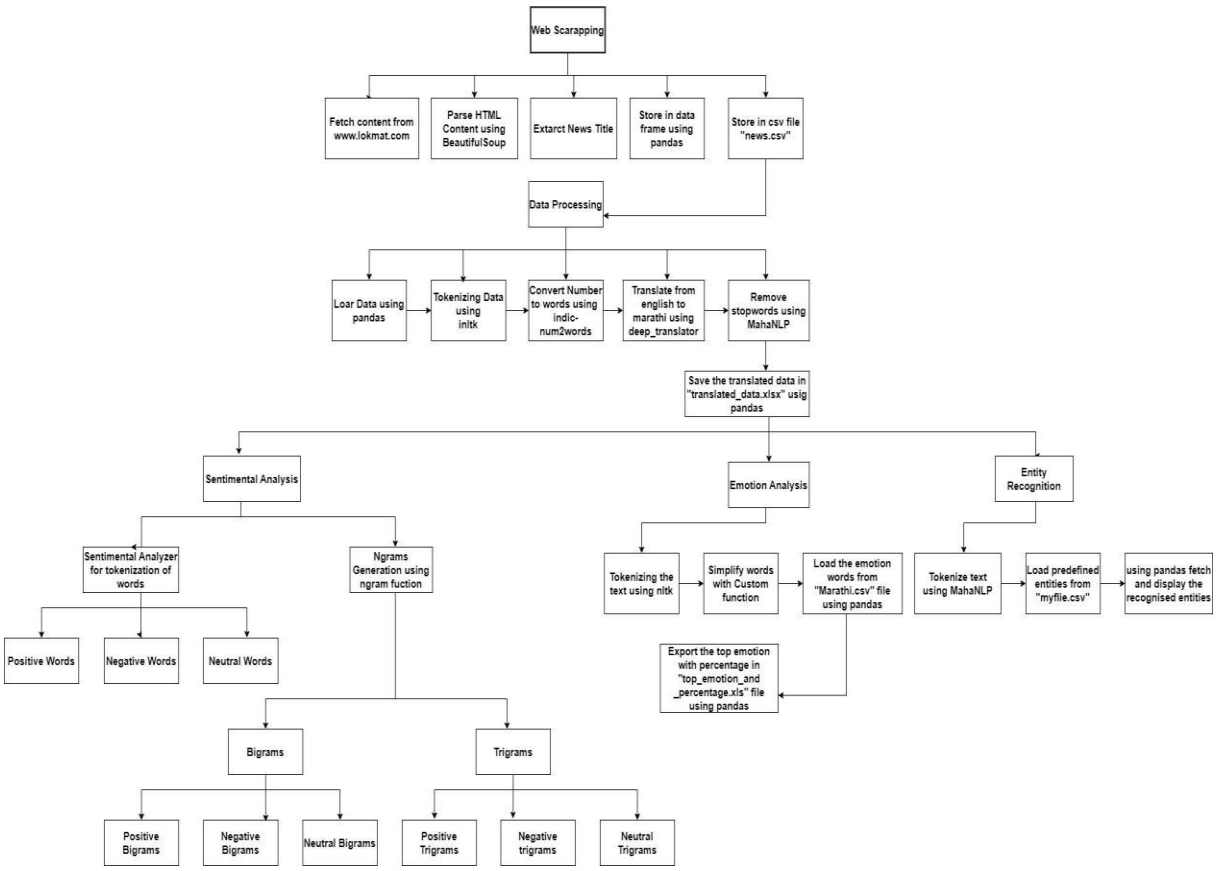


Fig. 2. research methodology applied for Maharashtra state regional language sentiment analysis and emotion detection

Fig. 2. illustrates a detailed workflow for the processing and analysis methodology applied on the news headlines from a Marathi news website (www.lokmat.com) utilising various natural language processing (NLP) techniques. This multi-step procedure comprises three key elements: web scraping, data processing, and analysis, which is further segmented into sentiment analysis, emotion analysis, and entity recognition.

Web Scraping

1. Content Retrieval: The process commences by retrieving content from www.lokmat.com.
2. HTML Content Parsing: The HTML content is parsed using BeautifulSoup.
3. News Headline Extraction: Headlines are extracted from the parsed HTML.
4. Data Storage: The extracted headlines are stored in a DataFrame using pandas and saved as "news.csv".

Data Processing

1. Data Loading: The CSV file ("news.csv") is loaded into a pandas DataFrame.
2. Tokenization: Data is tokenized utilising the nltk library.
3. Number Conversion: Numerical data is converted into words using the indic-num2words library.
4. Translation: Text is translated from English to Marathi using deep_translator.
5. Stopwords Removal: Stopwords are eliminated using MahanNLP.
6. Translated Data Saving: The cleaned and translated data is saved in "translated_data.xlsx".

Analysis

Sentiment Analysis

1. Tokenization: Text is tokenized to identify individual words.
2. Classification: Words are classified as positive, negative, or neutral.
3. N-grams Generation: Bigrams and trigrams are generated to comprehend the context.
4. Categorization: N-grams are categorised into positive, negative, and neutral groups.

Emotion Analysis

1. Tokenization: Text is tokenized once more.
2. Word Simplification: Words are simplified through a custom function.
3. Emotion Words Loading: Emotion words are imported from "marathi.csv".
4. Top Emotion Identification: The primary emotion and its percentage are identified and stored in "top_emotion_and_percentage.xls".

Entity Recognition is a crucial step in the workflow, as it involves extracting key entities from the raw news headlines. To achieve this, the text is tokenized using MahanNLP, which helps in breaking down the text into smaller units called tokens. These tokens are then used to identify and load entities from the "myfile.csv" file. Once the entities are loaded, they are retrieved and displayed using pandas, a powerful data manipulation library. This allows for easy access and analysis of the recognized entities. The overall goal of this workflow is to convert raw news headlines into valuable insights. This is achieved through various stages, including web scraping to efficiently collect raw data, data processing to prepare the data for analysis, sentiment analysis to determine the sentiment at word and n-gram levels, emotion analysis to identify and quantify primary emotions, and finally, entity recognition to extract key entities. By utilising tools like BeautifulSoup, pandas, nltk, MahanNLP, and deep_translator, this workflow ensures a streamlined process and provides accurate and in-depth insights from Marathi news headlines. Each stage and subprocess is thoroughly explained in the data collection methodology and further discussed in the results and discussion segment, highlighting the robustness of the research approach.

Result And Discussion

We have tried various approaches throughout the research methodology implementation and this section is going to highlight the small but major breakthrough as per our understanding while achieving the research objectives.

With the first approach wherein we have word and emotion based dictionary as key value pairs as represented in Figure 2, currently this directory is small but in future it can be a big data sized directory where in all the regional language(Marathi) words and their associated emotions will be stored and used as one of the emotion detector tool and technique. While looking for the emotions words and emotions will be identified and handled as tokens and then the cumulative count of each emotion is done and used for data/outcome visualisation

The cumulative count of each emotion *e* is given by the following mathematical formula definition:

1. Dictionary: *D* Where *D(w)* is the set of emotions associated with the word *w*
2. Text: *T* is the sequence of words *w1,w2,....wn* .
3. Emotion count: *C(e)* is the cumulative count of emotion *e.*

Formula: The cumulative count of each emotion e is given by:

$C(e) = \sum_{i=1}^{n} \mathbf{1}_{(e \in D(w_i))}$

Where:

- *n* is the number of words in the text (sentence token) T
- w_i represents each word/token in the text T
- $\mathbf{1}_{(e \in D(w_i))}$ is indicator function that equals 1 if emotion e is associated with word wi in the dictionary D, and 0 otherwise

A custom word-emotion based dictionary is created as shown in Fig. 3. for sentiment analysis and emotion detection wherein we have stored regional language (Marathi) words and its associated emotion as a pair separated with colon. After word tokenization each word as token will be checked in this dictionary and its associated emotion counter will be increased by one point for further calculations.

Fig. 3. Marathi language word and associated emotion based dictionary

Predefined Marathi Language stopwords:

```
marathi_stopwords = ["अधिक","अनेक","अशी","असलयाचे","असलेल्या","असा","असून","असे","आज","आणि","आता","आपलं","आपला","आपल्
    "एक","एका","कमी","काम","काय","काही","किव्हा","की","केलं","केला","केली","केले","कोटी","गेल्या","घेऊन","जात","झालं",
    "झाला","झाली","झाले","झाल्या","टा","डॉ","तर","तरी","तसेच","ता","ती","तीन","ते","तो","तोंड","थोडं","दिलं","दिला","दिली",
    "दिले","दोन","न","नाही","निर्ण्य","पण","पम","परयतन","पाटील","म","मात्र","माहिती","मी","मुबी","म्हणजे","म्हणाले","म्हणून","म्हणे",
    "या","याचा","याचं","याचा","याची","याच्या","याना","यानी","यामुळे","यार","याला","याली","याले","यावा","यास","यासाठी","यासाठीच","यासू",
    "याही","येणार","येत","येथील","येथे","येथून","येथे","लाख","लागलं","लागला","लागली","लागले","लागू","लाहून","व","वाटलं","वाटला","वाटली",
    "वाटले","वाहून","विरुद्ध","स","सकाळ","सगळं","सगळा","सगळी","सगळे","सज्ज","सदर","समज","समजलं","समजला","समजली",
    "समजले","समजून","समजून","समोर","सर्व","सागितलं","सागितला","सागितली","सागितले","सागित्ले","साठी","सातत","साहित्य","सुरू","सोबत","ह
    "होते","होतो","ह्या","ह्याचं","ह्याचा","ह्याची","ह्याच्या","ह्याना","ह्यानी","ह्यामुळे","ह्यार","ह्याला","ह्याली","ह्याले","ह्यावा","ह्यास","ह्यासाठी","ह्यासाठी
    "ह्यासू","ह्याही","होणार","होत","होथील","होथे","होथून","होथे","आता","जेवण","जात"]
```

Fig. 4. sample Marathi language stopwords

In the Fig, 3. As shown in above figure are the sample Marathi language stop words which can be removed from the sentence tokens to get the meaningful tokens which can give us sentiment scores as in cumulative sentiment analysis. The language specific stopword dictionary is right now not available with many packages available except inltk.

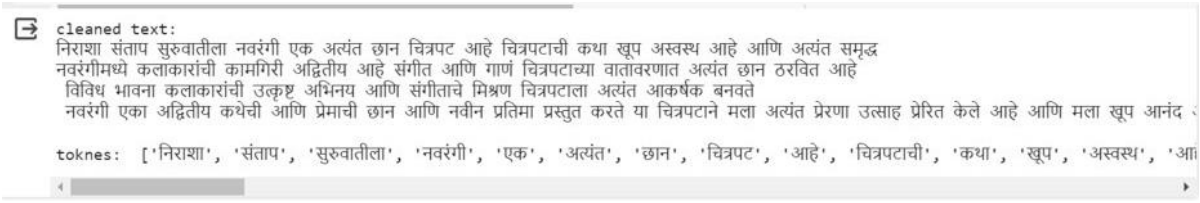


Fig. 5. text after text cleaning-removing punctuations and tokenization

The output in Fig. 5 above shows the final cleaned text of the regional language with separated tokens. This process requires lowering the text and removal of punctuation marks and many other symbols. The challenge is that few words in the devanagari script have punctuation marks as a part of script but the predefined punctuation removal function may remove it so we need to customise the function and use it for Marathi language.

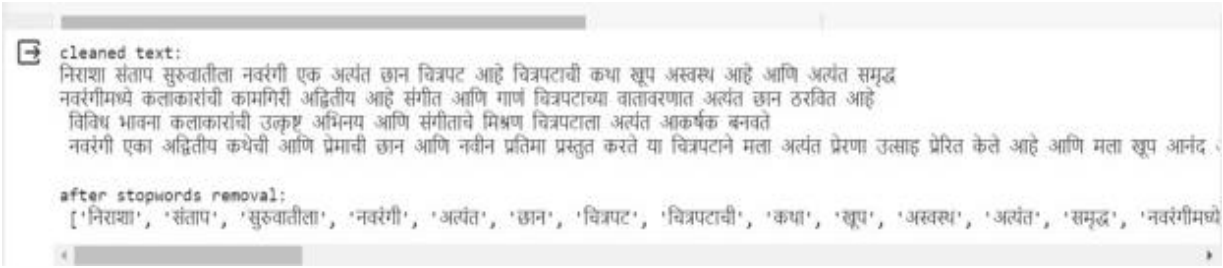


Fig. 6. Input Text after removal of stopwords

After removal of stopwords we received pure tokens of the regional language as presented in Fig. 6 these words are most probable and suitable candidates for the regional language sentiment analysis as they hold some meaning and sentiments of the Marathi language.

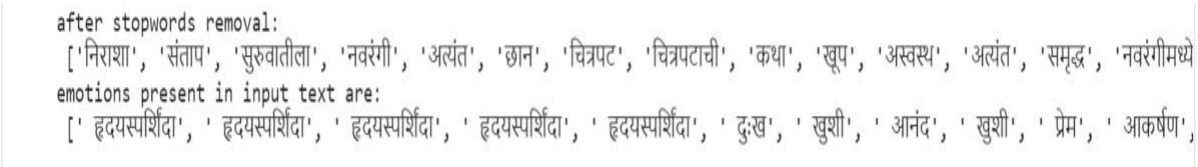


Fig. 7. listing the emotions present in the input text

Each word as token is selected and cross checked against the emotion based dictionary D and emotion associated with that word token is picked as depicted in the Fig. 7, these might be repetitive emotions since many words are associated with one emotion. Maharashtra state regional language Marathi is enriched with millions of such words and they represent many emotions associated with them but here we are considering one word and one emotion since currently we are not considering the context and situation based sentiment analysis.

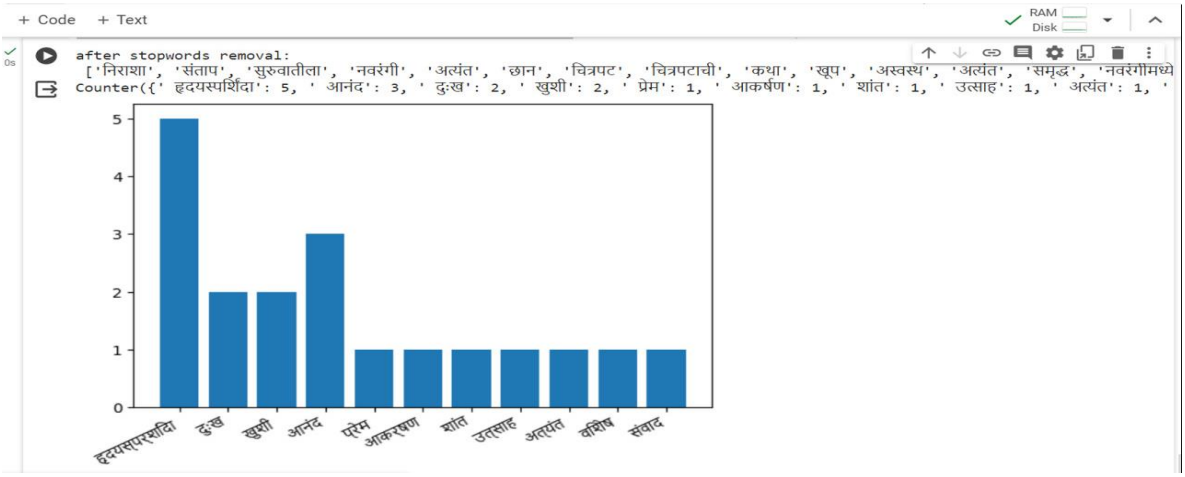


Fig. 8: sentiment emotion visualisation with marathi emotions on x-axis

The cumulative sentiment scores are diagrammed in Figure 8 where emotions and its respective count from the imputed text are used to formulate a graph. The x-axis labels are mentioned in the regional language text using UTF-16 encoding for data visualisation. In Figure 7 we have shown more than 10 regional language sentiments in one graph which highlights the overall sentiments of the imputed text.

To sum up the regional language emotion based sentiment analysis, we can segregate and analyse as many emotions as we can define and categorise in the emotion based dictionary. Whereas with the help of R packages like syuzhet we can only categorise 7 to 8 sentiments which are not so relevant with the regional language since the dictionary is mainly used for English language sentiment analysis.

Approach 2 : sentiment score based regional language sentiment analysis

Polarity check is one of the basic sentiment analysis where in the token represents one of the classification like positive and negative in the case of sentiment analysis, we did the polarity check for Maharashtra state regional language Marathi. We consider phrases, bigrams and trigrams for the polarity check since they are an integral part of the written communication. We considered Marathi language phrases as special cases and defined a custom process while removing stopwords, the whole phrase is assumed as the unique sentiment based sentence token and will not be the candidate for stopword removal process.

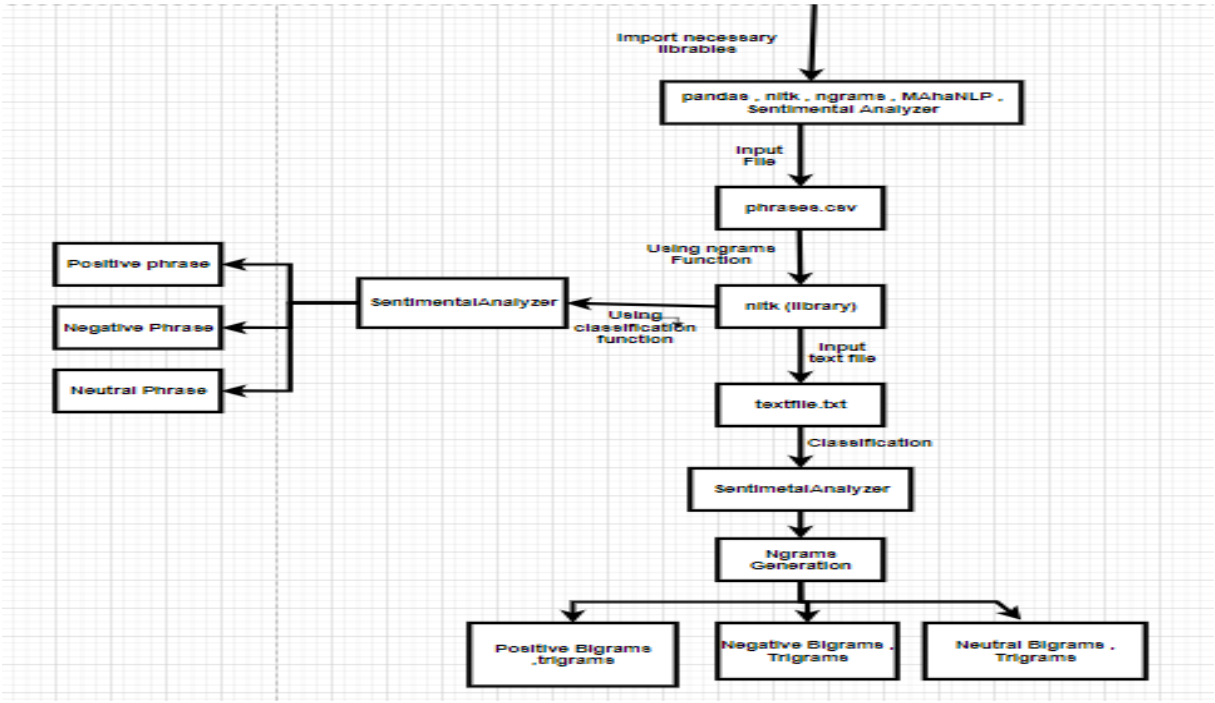


Fig. 9. sentiment analysis and polarity check for bigrams and trigrams

We do this to retain the regional language uniqueness and its sentimental contribution of these phrases in overall input text. Figure 7, 8 and 9 gives us the glimpses of the polarity check for Marathi phrases, bigrams and trigrams; this inclusive approach will surely add value in the overall sentiment analysis and retention of contextual meaning of the sentences. We are not considering contextual sentiment analysis but trying to retain the same by holding phrases and bigram trigram of the input text.

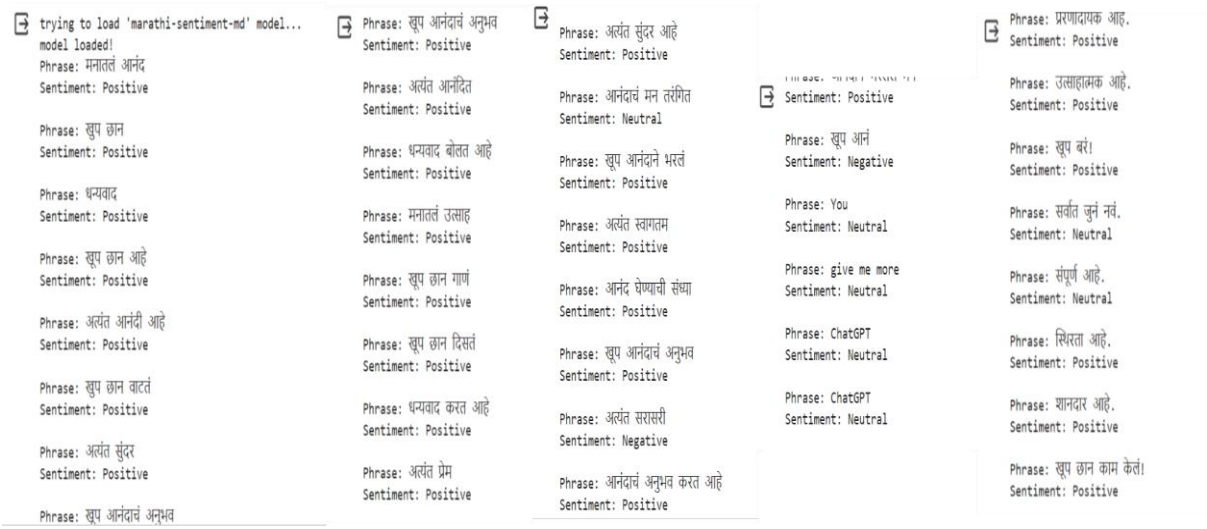


Fig. 10. Polarity check for phrases/bigrams and trigrams

Labelling the tokens as positive and negative as shown in the Figure 10 and 11 are unit polarity check activity, which might be useful in future scope of the regional language sentiment analysis process for evaluating the overall sentiments of the huge chunk of regional language text.

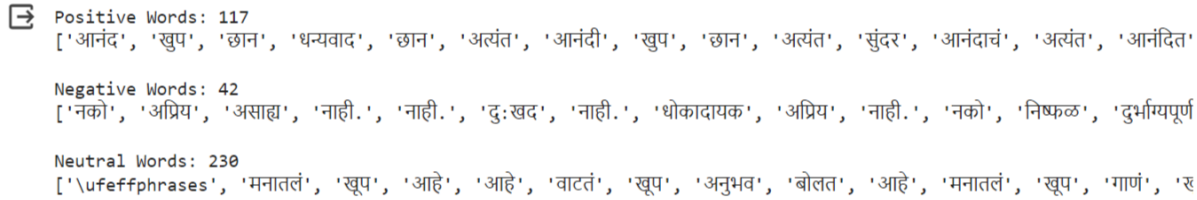


Fig. 11 : labelling of positive negative and neutral words in Marathi (Regional) Language

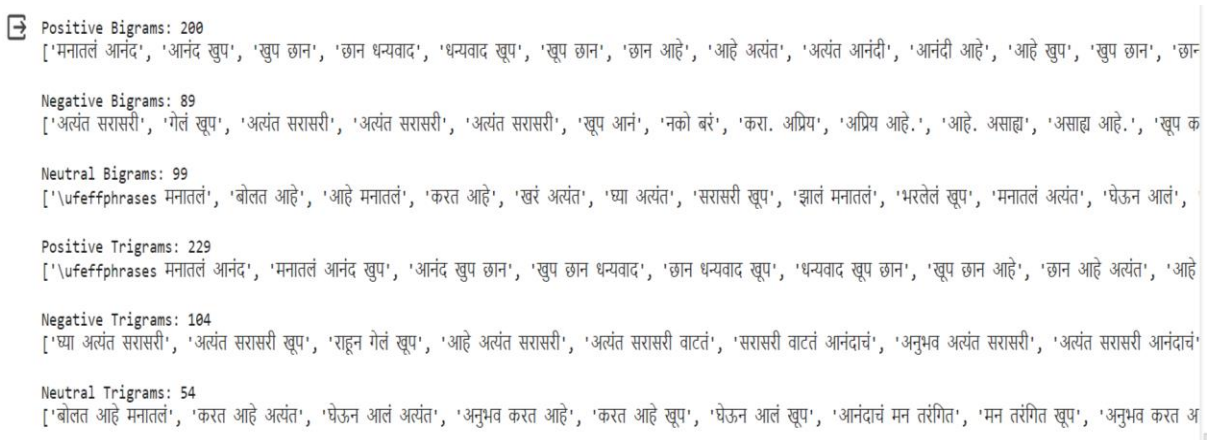


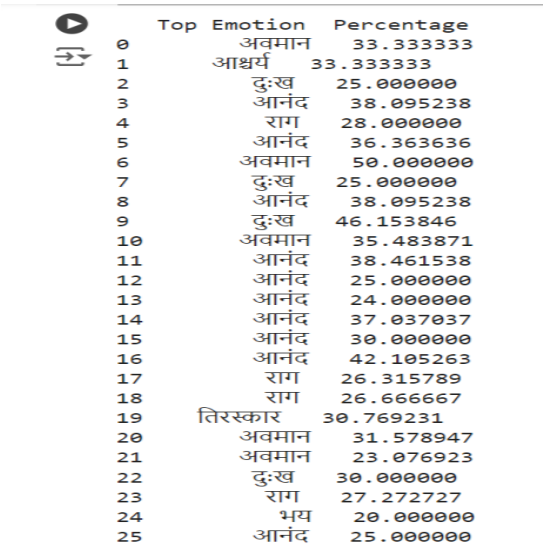
Fig. 12. **positive**, negative and neutral bigrams and trigrams in Marathi (Regional) Language

The overall sentiments of the inputted text are depicted and shown in Figure 9, the summary of the sentiment analysis shows that the provided regional language marathi text contains 42.86% **'दुःख' (sadness)**, 14.29% **'राग'** (**anger**), 14.29% **'अवमान' (contempt) and 28.57% आनंद (happiness)**

Here with the above sentiment analysis of the regional language input text, we can conclude that in the overall analysis we can find out that the inputted text was overall negative as a polarity analysis but the detailed analysis gives us lots of insights about the feelings of the regional language text generator.

Fig. 13. sentiment analysis based on different set emotions for Marathi language text

The informatics shown in Fig.14 displays the results of emotion detection in Marathi language text. It lists the "Top Emotion" identified in each text segment along with its "Percentage," indicating the proportion of the text reflecting that emotion. Emotions detected include आनंद (joy), अपमान (insult), दुःख (sadness), आश्चर्य (surprise), राग (anger), and तिरस्कार (disgust). For example, आनंद appears frequently with varying percentages, up to 42.105263%. Other emotions like अपमान and दुःख also show significant presence, with percentages reaching up to 50% and 46.153846%, respectively. This analysis highlights the dominant emotions present in the text.



	Top Emotion	Percentage
0	अवमान	33.333333
1	आश्चर्य	33.333333
2	दुःख	25.000000
3	आनंद	38.095238
4	राग	28.000000
5	आनंद	36.363636
6	अवमान	50.000000
7	दुःख	25.000000
8	आनंद	38.095238
9	दुःख	46.153846
10	अवमान	35.483871
11	आनंद	38.461538
12	आनंद	25.000000
13	आनंद	24.000000
14	आनंद	37.037037
15	आनंद	30.000000
16	आनंद	42.105263
17	राग	26.315789
18	राग	26.666667
19	तिरस्कार	30.769231
20	अवमान	31.578947
21	अवमान	23.076923
22	दुःख	30.000000
23	राग	27.272727
24	भय	20.000000
25	आनंद	25.000000

Fig. 14. Sentiment analysis based on top set emotions detected in the imputed Marathi Language text dataset

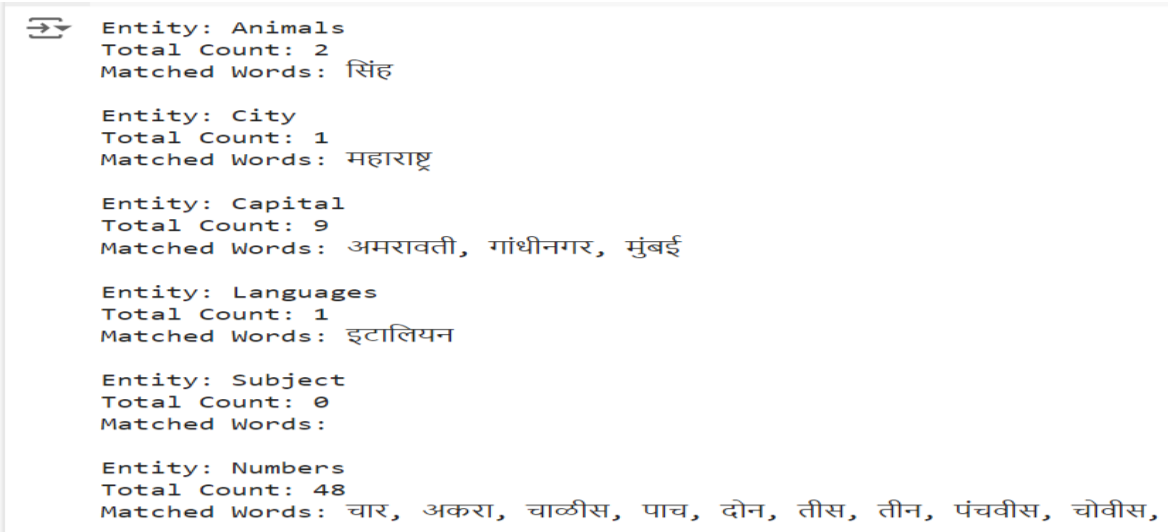


Fig. 15. Named Entity Recognition for Maharashtra state regional language Marathi

The diagram illustrates the outcomes of named entity recognition (NER) for Marathi language text. It classifies the identified entities into different categories, such as Animals, City, Capital, Languages, Subject, and Numbers. Each category presents the total count of recognized entities and provides a list of the corresponding words. For instance, in the "Animals" category, there are 2 matches (सिंह), while the "City" category has 1 match (महाराष्ट्र), and the "Capital" category has 9 matches, including अमरावती, गांधीनगर, and मुंबई. The "Languages" category includes 1 match (इटालियन), whereas the "Numbers" category lists 48 matches, including चार, अकरा, चाळीस, and पाच. Notably, the "Subject" category does not have any matches.

Application Areas

Sentiment based translation The language market grew from 23.50 billion USD to 49.60 USD in just a decade's time frame and is expected to grow exponentially in the coming years. CSA Research's 15th annual in-depth analysis of the language industry indicates that the sector is still expanding as a result of worldwide digital transformation[22].

Multilingual/Regional customer support Currently AI ML based customer support is for efficient instance, error and ticket solving is done with few AL ML based services like IBM watson, watson conversation service, watson tone analyzer, Microsoft's Language Understanding Intelligent Service (LUIS), Google's dialog flow etc. 71.5% of customer service leaders affirm that providing support in a customer's native language boosts satisfaction. Furthermore, 74% of customers are inclined to repurchase or reuse services if post-sales assistance is offered in their preferred language.Due to Poor customer service companies tend to lose $62 billion annually.

Market Expansion Leveraging regional language and regional language sentiment analysis taps into the previously untapped markets with tailored products and services for selective customer segments. This opens the whole product and service lifecycle process adaptation to the regional language and local target customers and consumers.

Limitations And Future Scope

The current study conducted holds certain language specific and technological limitations to identify the exact and accurate sentiment analysis and can not claim it to be 100% correct. Since the message conveyed always depends on the context in which this communication happened so we felt understanding or identifying the context of communication is another limitation of this paper. Though sometimes words mean something but it can be sarcastically used to convey the hidden agenda or meaning of the sentence and communication models do not promise to identify any of these as part of sentiment analysis. There are many language specific challenges that we might have not identified such as limitations of our approach so we consider those limitations as well as part of this research study conducted. We will try to implement various other language specific evaluations and existing evaluations at the rigorous and depth level structured way to get more accurate results. Using these analysis for the cross domain applications can be one of the major future scope of work we have done or will be done in future tenures.

Conclusion

The study comes up with one of the approaches of doing sentiment analysis of Maharashtra state regional language that is Marathi. The datasets right now generated are generic and gather from some social media text and microblogging web pages. The specific way of cleaning the regional language text is required due to language specific script and its uniqueness of presentation of fonts and words. With the current study we are able to identify positive, negative and neutral sentiments from the Marathi language text. Right now we have done this with a traditional machine learning approach and were able to identify a few more sentiments like happiness, sadness, anger, disgust with presocred marathi words and weightage dictionary. We tried to do sentiment analysis on marathi language specific bigrams and trigrams and phrases and it will be further carried out with the rigorous continuous research work. Both the approaches i.e Emotion Based Sentiment Analysis (EBSA) and Corpus Score Based Sentiment Analysis were able to achieve the research objectives of this study.

Acknowledgement

We sincerely thank the Swami Ramanand Teerth Marathwada University Nanded for the University department support and permission to conduct this research study in this area, also we sincerely convey our gratitude

towards the College of Computer Science and Information Technology (COCSIT) for their infrastructural and research centre support to conduct the study. We acknowledge that the datasets were created through a few microblogging websites and webpages, hence we will try to make these datasets available on the public domain for research scholars to use for their respective studies.

References

1. H. Permana and A. Purwarianti, "Sentiment Analysis in Sundanese Using Pre-trained Multilingual Language Models," 2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan, 2022, pp. 1-5, doi: 10.1109/ICAICTA56449.2022.9932940.
2. G. I. Ahmad and J. Singla, "(LISACMT) Language Identification and Sentiment analysis of English-Urdu 'code-mixed' text using LSTM," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 430-435, doi: 10.1109/ICICT54344.2022.9850505.
3. Jamatia, A., Gambäck, B., Das, A. (2018). Collecting and Annotating Indian Social Media Code-Mixed Corpora. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2016. Lecture Notes in Computer Science(), vol 9624. Springer, Cham. https://doi.org/10.1007/978-3-319-75487-1_32
4. Richard Rogers (September 15, 2023). 74+ Language Statistics For 2024 (Trends, Facts & Data) Language statistics, https://myclasstracks.com/language-statistics/
5. Tanushree Bsuroy (Mar 19, 2021). Category preference in digital news among Marathi in India 2016, https://www.statista.com/statistics/719666/popular-categories-in-digital-news-among-marathi-users-india
6. The new Indian Express(2024),PM launches engineering courses in 5 local languages in 14 institutes across 8 states, https://www.newindianexpress.com/nation/2021/Jul/29/pm-launches-engineering-courses-in-5-local-languages-in-14-institutes-across-8-states-2337315.html
7. Gupta, S., & Agrawal, A. (2020). *A Comprehensive Review on Opinion Mining: Approaches, Issues, Challenges, and Future Directions*. Journal of King Saud University - Computer and Information Sciences. [DOI: 10.1016/j.jksuci.2020.06.004]
8. Joshi, M., & Tiwary, U. (2019). *Sentiment Analysis: A Comprehensive Review*. Journal of Information Science, 45(5), 647-664. [DOI: 10.1177/0165551518809207]
9. Mukherjee, A., & Bhattacharyya, P. (2012). *Sentiment Analysis on Twitter with Lightweight Discourse Analysis*. ACM Transactions on Intelligent Systems and Technology (TIST), 3(4), 1-29. [DOI: 10.1145/2337542.2337553]
10. Bhattarai, S., Kang, S., & Lee, S. (2019). *A Survey on Sentiment Analysis and Opinion Mining: Approaches and Techniques*. Information Sciences, 463, 407-416. [DOI: 10.1016/j.ins.2018.07.076]
11. Suman, S., Acharya, D., & Zalavadia, C. (2021). *Sentiment Analysis of Indian Regional Languages: A Review*. International Journal of Intelligent Engineering and Systems, 14(1), 96-108. [DOI: 10.22266/ijies2021.0105.10]
12. Yadav, S., & Bethard, S. (2020). *A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models*. Journal of Big Data, 7(1), 1-32. [DOI: 10.1186/s40537-020-00348-7]
13. Sharma, A., & Das, S. (2018). *Sentiment Analysis in Hindi Social Media Texts: Challenges and Opportunities*. International Journal of Computational Linguistics and Applications, 9(1), 45-56.
14. Chen, Y., Wang, L., & Liu, H. (2019). *Sentiment Analysis in Mandarin Social Media Content: Approaches and Considerations*. Journal of Computational Linguistics, 12(3), 123-137.
15. Rodriguez, M., & Torres, J. (2020). *Sentiment Analysis in Spanish Tweets: A Machine Learning Approach*. Journal of Language Technology and Computational Linguistics, 7(2), 89-102
16. Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press. [ISBN: 978-1-107-08958-5]
17. Wang, X., & Wan, X. (2016). Mining User-generated Content on Social Media for Product Aspect Ranking. Information Sciences, 367-368, 226-236. [DOI: 10.1016/j.ins.2016.05.045]
18. Singh, A., & Bala, M. (2019). Sentiment Analysis of Indian Social Media Texts: A Review. Indian Journal of Computer Science and Engineering, 10(2), 45-58.
19. Gupta, R., & Kumar, V. (2020). Sentiment Analysis in Indian Vernacular Languages: Challenges and Opportunities. Proceedings of the International Conference on Computational Linguistics (ICCL), 78-89.
20. Patel, S., & Patel, R. (2018). A Study of Sentiment Analysis Techniques in Indian Languages. Journal of Indian Computing and Information Sciences, 14(3), 112-125.

21. Mishra, S., & Sharma, P. (2021). Sentiment Analysis of Indian Microblog Texts: A Deep Learning Approach. Indian Journal of Information Technology, 18(4), 207-220.
22. Newswire.com,BOSTON(2019), Global Market for Outsourced Translation and Interpreting Services and Technology to Reach US$49.60 Billion in 2019, https://www.newswire.com/news/global-market-for-outsourced-translation-and-interpreting-services-and-20903246
23. Kakuthota Rakshitha, Ramalingam H M, M Pavithra, Advi H D, Maithri Hegde,Sentimental analysis of Indian regional languages on social media,Global Transitions Proceedings,Volume 2, Issue 2,2021,Pages 414-420,ISSN 2666-285X,https://doi.org/10.1016/j.gltp.2021.08.039.
24. Nandwani, P., Verma, R. A review on sentiment analysis and emotion detection from text. Soc. Netw. Anal. Min. 11, 81 (2021). https://doi.org/10.1007/s13278-021-00776-6
25. Abdaoui, A., Azé, J., Bringay, S. et al. FEEL: a French Expanded Emotion Lexicon. Lang Resources & Evaluation 51, 833–855 (2017). https://doi.org/10.1007/s10579-016-9364-5
26. Sidhu, S., Khurana, S.S., Kumar, M. et al. Sentiment analysis of Hindi language text: a critical review. Multimed Tools Appl 83, 51367–51396 (2024). https://doi.org/10.1007/s11042-023-17537-6
27. Tori, F., Tori, S., Keseru, I. et al. Performing Sentiment Analysis Using Natural Language Models for Urban Policymaking: An analysis of Twitter Data in Brussels. Data Sci. Transp. 6, 5 (2024). https://doi.org/10.1007/s42421-024-00090-5
28. Vichare, A.A., Varma, S.L. (2024). Sentiment Analysis: Indian Languages Perspective. In: Das, P., Begum, S.A., Buyya, R. (eds) Advanced Computing, Machine Learning, Robotics and Internet Technologies. AMRIT 2023. Communications in Computer and Information Science, vol 1953. Springer, Cham. https://doi.org/10.1007/978-3-031-47224-4_23
29. Mhaske, N.T., Patil, A.S. Sentence Annotation for Aspect-oriented Sentiment Analysis: A Lexicon based Approach with Marathi Movie Reviews. J. Inst. Eng. India Ser. B (2024). https://doi.org/10.1007/s40031-024-01072-5