# Survey on Sentiment Analysis of Marathi Speech and Script

**Prof. Vina M. Lomte[1], Pratik Jadhav[2], Onkar Kalshetti[3], Sonali Deshmukh[4], Arjun Jadhav[5]**

[1-5]*Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, India*

---

**Abstract –** *Sentiment Analysis on Voice and textual data is being carried out widely. Sentiment analysis has many applications ranging from business, Government and consumer perspective in this data era. According to Google, India today has a population of 1380 Million. Out of which 825 million users are active on the internet. These huge user bases are distributed across various languages. Out of which Marathi stands to be 3rd most popular language in India. Even then it is the least focused language when talking about sentiment analysis. This is due to absence of proper dataset and lack of research focusing on Marathi text and speech. So, in this study we have studied different methodologies of sentiment analysis across various languages. In this survey paper we have taken comparative study of different techniques and approaches of sentiment analysis on voice data, tweets and textual data.*

## 1. INTRODUCTION

In this modern world, a large number of people have access to mobile phones. People use mobile phones to call each other, text each other and use social Media platforms. This in turn generates a huge amount of data. The data generated is mainly in 2 forms i.e., Speech and text. If a person wants to buy any product online or wants to use any service, then he/she will first look upon its reviews online. Before making any decision, he/she might as well discuss it with other consumers who have already used the product or service. Hence the data generated by the users in form of reviews is sentiment rich and valuable for the companies to gain insights about their product. But the data produced is unstructured and can be in the form of text, emoticons, URLs, punctuation, etc. too vast to analyze manually. Therefore, Natural Language Process (NLP) is done. One of the main applications of NLP is Text mining. Sentiment Analysis (SA) uses text mining and NLP to process this user created content. Similarly whenever any user buys any product online, the ecommerce sellers call the user asking for feedback. This feedback is in the form of voice data. This data can be then analyzed using NLP. Primary goal in this process is to collect data in the form of Speech or text, pre-process this raw data using various techniques and then use this data to get useful insights. In this paper, we are going to perform a detailed analysis of various Sentiment analysis methodologies across various languages.

In today's Scenario, we can see that Sentiment analysis of the English language is being done the most and has a lot of outcomes each with a unique efficiency and result. But as we know, with technological advancements people are able to learn more than 2 languages. In short, people are able to speak multiple languages, so it becomes natural that even the machines that are learning should learn all the languages. There has been some great development even in other languages like Chinese, Bengali, etc. and all with successful results. There are still certain areas which are not yet entirely focused or successfully researched in the field of sentiment analysis; and those fields are of the local languages, India is a very big country which speaks a lot of languages. There are 22 official languages in India and creating a successful sentiment analysis model for all of them is nearly impossible – mostly due to the fact that not enough data is available on the unscheduled languages and that not a lot of people really use the language. A large number of people in India use English words or sentences in their day to day lives. Marathi is a language spoken by the state of Maharashtra and hence is the local language of the state. Hence, that makes Marathi one of the most important and spoken languages of the country. Hence, we have decided to use this as a local language for our study and research.

## 2. Literature Survey

**Table -1:** Deep Literature Survey of Current Technologies

| SR. No | Publication Details | Authors | Tech Used | Accuracy | Dataset | Research Gap |
|---|---|---|---|---|---|---|
| 1. | Applying natural language processing to analyse customer satisfaction *IEEE* | Armin Alibasic and Tomo Popovic, *Senior Member* | Scraping, Beautiful Soap library **Pre-processing:** Tokenization, Removing Stop Words, Stemming - cutting the words to the root of that word flying - fly, seats - seat, etc. **Post-processing:** Python NLTK library | 95% | 50,000 airline reviews from TripAdvisor | sentiment score is calculated using the frequency of appearing positive and negative words. Dataset contained reviews by Native English Speakers. |
| 2. | A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals IEEE International Conference ICOSEC 2020 | Deepak Bharti (Author) Poonam Kukana (Co-Author) | **Feature Extraction:** (Ant Lion Optimization)ALO, GFCC **Post-processing:** Classification (MSVM) | 97% | RAVDESS data set (Emotional Speech containing 7356 Files) t the | Dataset used contains Male: Female Voice Ratio as 4: 1 |
| 3. | Deep Learning Based Language Identification System From Speech | Athira N P Poorna S S | **Pre-processing:** DC and Noise Removal, Pre-emphasis, Spectrograms **Post-processing:** precision, recall, F1 score and confusion matrix | 74% | Speech database of seven Indian languages from IIIT-Indic | Dataset can be Specifically Designed for Marathi Speech data to achieve better accuracy |
| 4. | Speech Recognition using HTK Toolkit for Marathi Language | Supriya S.Chavan Dr.S.M.Handore | **Pre-processing:** Speech Enhancement, Linear Predictive Coefficients, Speech Segmentation, **Feature Extraction:** Hidden Markov Modelling Tool Kit 3.4, MFCC, LPC **Post-processing:** **Speech** | 80% | 910 Marathi sentences, recorded by 3 male and 3 female speakers. | Speech Recognition accuracy can be Improved as Current accuracy is 80%. Dataset Size & variety can be Increased. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | **Recognition:** ANN | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5. | Multiclass Classification and Class based Sentiment Analysis for Hindi Language | Prof.Sumitra Pundlik, Prachi Kasbekar, Gajanan Gaikwad, Prasad Dasare, Akshay Gawade, Purushottam Pundlik | HindiSentiWordNet(HSWN) Language Mode(LM) Classifier | | Speeches from leaders across various Topics converted to text files. | Self-Learning ontology not Implemented. Works only with text documents with Word limitations |
| 6. | Isolated Spoken Marathi Words Recognition using HMM | Sai Sawant and Mangesh Deshpande | **Pre-processing:** Training and Testing the model **Post-Processing:** Hidden Markov Models (HMMs) | 86.5% | 2,000 Marathi words spoken by ten native speakers | To increase The recognition performance and speaker independence of this system, training dataset needs to be increased by using data obtained from increased number of speakers. |
| 7. | Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing | Mrs. Rushali Dhumal (Deshmukh) and Dr. Arvind Kiwelekar | **Pre-processing:** Early Dropping and Drop out Techniques are used. **Post–Pre-processing:** Python NLTK, Scikit-learn | 85% for the deep learning model and 97% for the Bi-LSTM model. | Pos - tagged corpus containing 1500 sentences with 10115 words has been created using Unified Parts of Speech from Marathi Newspapers. | |
| 8. | Urdu Sentiment Analysis With Deep Learning Methods | Lal Khan, Ammar Amjad, Noman Ashraf, HSIEN-TSUN | **Pre-processing:** Stop Words, Normalization **Post-** | 82.05% | The data of Movie and electronic reviews. Urdu text from Urdu news | One of the limitations of this study is that it includes only positive and negative classes; |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | G CHANG | **Processing:** SVM, NB, RF, AdaBoost, MLP, LR, 1D-CNN, and LSTM | | websites. | |
| 9. | A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter | MONDHER BOUAZIZI, TOMOAKI OHTSUKI | **Pre-processing-** By removing the URLs, tags, Tokenization **Post-Processing**: Random Forest, Binary Classification, Ternary Classification | 60.2% | **Set 1**: 21000 tweets **Set 2:** 19740 tweets | The proposed paper hasn't used the results obtained through Ternary Classification |
| 10. | Emotional voice conversion using multitask learning with Text-To-Speech. Information and Electronics Research Institute, KAIST, Daejeon, Republic of Korea | Tae-Ho Kim , Sungjae Cho, Shinkook Choi, Sejik Park and Soo-Young Lee | **Pre-processing:** Voice activity detection algorithm **Post-Processing :** Attention,decoder | | male-Korean-Emotional-Text-to-speech (mKETTS) dataset. | In future need to focus on improvement of TTS by VC As pronunciation differs also focus on explicit loss to keep down the difference between TTS and VC. |
| 11. | CubeMaha Sent: A Marathi Tweet-based Sentiment Analysis Dataset | Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi | Algorithms Used: CNN, BiLSTM+Global MaxPool, ULMFiT, BERT | 3 class accuracy – 84.13, 2 class accuracy – 92.93 | Marathi Sentiment Analysis Dataset - L3CubeMahaSent | The dataset does not retain any context of the tweets such as the tweeting profile, time of posting, region, etc. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12. | Spontaneous Emotion Recognition for Marathi Spoken Words International Conference on Communication and Signal Processing, | Vaibhav V. Kamble , Bharatratna P.Gaikwad , Deepak M.Rana | Principal Component analysis, Linear Discriminate Analysis, Mel-frequency scale analysis, cepstrum, Dynamic feature extractions | 83.33% | Berlin Emotion database | 1.The feature vectors sample speech signal are never exactly the same as those provided during the enrollment or training phase for the same user.<br><br>2.If the inferior database is used as an input to the system then incorrect conclusion may be drawn |
| 13. | Vachantar–Lokbhasha: A Speech to Text Conversion for Marathi | Archana. V. Chechare | Mel Frequency Cepstral Coefficient (MFCC), Artificial Neural Networks (ANN) | 87.76% | | As in this paper recognition is in Marathi it has been performed in Offline and Runtime modes. The accuracy of the system in offline mode is more as compared to runtime mode. |
| 14. | Multi-Class Sentiment Analysis in Twitter: What if Classification is not the Answer | Mondher Bouazizi, Tomoaki Ohtsuk | **Pre-processing:** Features Extraction<br><br>**Post-processing:** Naïve Bayes Classifier, Random forest classifier, Iterative Dichotomiser 3 classifier | 77.4% | Dataset is made of tweets collected using Twitter API | Need to address the case of tweets with sentiments belonging to different polarities and try to find possible ways to identify these sentiments. |
| 15. | Natural Language Processing based Rule Based Discourse Analysis of Marathi Text | Kalpana Khandale, C Namrata Mahendar | **Pre-processing:** Discourse Analysis,<br><br>**Segmentation:** POS tagger<br><br>**Post-processing:** Anaphora Resolution | 85.43% | Marathi Balbharati book from standard 1st to 4th class | This paper highlighted to the issues and rule based structure to resolve the discourse anaphora based on Marathi Grammer |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16. | Novel Technique for Script Translation using NLP: Performance Evaluation | Darshana Patil, S.B. Chaudhari, Sharmila Shinde | **Pre-processing:** E-M Cross conversion system,which is based on query translation approach **Post–processing:** Hybrid translation | 91% | Marathi WordNet | Can implement best hybrid system for multi-language translation using NLP in future |
| 17. | A Spell-checker Integrated Machine Learning Based Solution for Speech to Text Conversion | H.M Mahmudul Hasan, Md.Toufiqe Hasan, Md.Adnaul Islam, Md.Araf Hasan | **Pre-processing:** TensorFlow **Segmentation:** Deep Speech **Post–processing:** Peter Nerving's correct spelling suggestion algorithm | Before applying spell corrector: 30-54%<br><br>After applying spell corrector: 45-67% | Bengali Dataset: "Bengali.ai" & English Dataset: "Mozilla Common voice" | Can be focused on developing a larger and more effective dataset for the Bengali language so that the machine can process data better and faster |
| 18. | Kannada Speech to Text Conversion Using CMU Sphinx | Shivakumar K.M, Aravind K.G, Anoop T.V, Deepa Gupta | **Pre-processing:** CMU Sphinx **Feature Extraction:** Nfilt parameter (with value 52), **Post-processing:** Context Dependent Model(Single User Speech) | Context Dependent Model: 95%<br><br>Context Independent Model: 80% | Speech corpora for Kannada created by IIIT Hyderabad | These system holds accuracy for Kannada sentences with only four to ten word length. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 19. | Cognitive Devanagari (Marathi) Text to Speech System | Shaikh Shadab Shakil, Manjare Chandrapra bha Anil | **Pre-processing:** Mapper & Combiner for input text **Feature Extraction:** Optical Character Recognition( OCR) system **Post-processing:** Text-to-Speech (TTS) Synthesizer | | Database creation can be done in MS excel. | In future can be focused on the signals which we get here as the output can be thus reverse engineered and can be useful as the base for a TTS engine itself. |
| 20. | Hindi Sentence Classification for Expressive Storytelling Systems | Shivli Agrawal, Yukti Kirtani, Yuki Girdhar and Swati Agrawal | **Pre-processing:** RNN-LSTM **Feature Extraction:** Word2Vec **Post-processing:** CNN-SVM | Overall accuracy : 88.1% | Hindi language stories from texts like “Panchatantr a” & “Akbar Birbal” | In future can be focused on improving by incorporating various other discourse modes and taking their subclasses into account. |
| 21. | Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning | LI YANG, YING LI , JIN WANG, R.SIMON SHERRATT | **Pre-processing:** Sentiment lexicon **Feature Extraction:** CNN model & GRU model **Post-processing:** SLCABG model | weighte d word vector: 93.5% unweigh t ed word vector: 92.8% | The data of book reviews collected from “Dangdang” using web crawler technology | Can be focused to study the Sentiment fineness classification of text. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 22. | UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods | UZMA NAQVI, ABDUL MAJID, AND SYED ALI ABBAS | **Pre-processing:** Data Cleaning & Tokenization **Feature Extraction:** Word Embedding Models **Post-processing:** DL Models(LST, BiLSTM-ATT, CNN, and C-LSTM) | 77.9% | Urdu blogs and news websites such as BBC, DW, Express, Dunya and humsub | In future can be focused on DL models with more embeddings that provide better contextual information along with increased and balanced dataset will be explored. |
| 23. | A Research Work on English to Marathi Hybrid Translation System | Pramod Salunkhe, Mrunal Bewoor, Dr.Suhas Patil | **Pre-processing:** Statistical Translation **Post-Processing:** Hybrid Translation | Average Precision : 87.5% | statistical system built on corpus downloaded from IITB produced by Pushpak Bhattacharya | Further research can be made in better translation with application of Marathi wordnet or development of trans-lingual word net for all. |
| 24. | SENTIMENT ANALYSIS OF MARATHI LANGUAGE | Sujata Deshmukh, Nileema Patil, Surabhi Rotiwar, Jason Nunes | **Pre-processing:** Segregation and lemmatization **Post-Processing:** Attribute Polarity | With Yandex translator: 60-70% | English SentiwordNet | Further real time update of dictionary can be future research direction in the field of sentiment analysis of Marathi language. |

### 3. Algorithmic Survey

**Table -2:** Algorithmic Survey of Research Studies

| SR. No | Publication Details | Authors | Tech Used | Accuracy | Dataset | Research Gap | Advantages |
|---|---|---|---|---|---|---|---|
| 1. | Applying natural language processing to analyse customer satisfaction | Armin Alibasic and Tomo Popovic, *Senior Member* | Scraping, Beautiful Soap library **Pre-processing:** Tokenization, Removing Stop Words, Stemming - cutting the words to the root of that word flying - fly, seats - seat, etc. **Post-processing:** Python NLTK library | 95% | 50,000 airline reviews from TripAdvisor | sentiment score is calculated using the frequency of appearing positive and negative words. Dataset contained reviews by Native English Speakers. . | Analyzed Sentiments using Bigrams and Trigrams |
| 2. | A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals IEEE International Conference | Deepak Bharti (Author) Poonam Kukana (Co-Author) | **Feature Extraction:** (Ant Lion Optimization)ALO, GFCC **Post-processing:** Classification (MSVM) | 97% | RAVDESS data set (Emotional Speech containing 7356 Files) t the | Dataset used contains Male: Female Voice Ratio as 4: 1 | Proposed System increased the quality of the speech signal and reduce the noise level in the uploaded samples. |
| 3. | Deep Learning Based Language Identification System From Speech | Athira N P Poorna S S | Pre-processing: DC and Noise Removal, Pre-emphasis, Spectrograms Post-processing: precision, recall, F1 score and confusion matrix | 74% | Speech database of seven Indian languages from IIIT-Indic | Dataset can be Specifically Designed for Marathi Speech data to achieve better accuracy | 99.5% accuracy is achieved for the Bengali-<br><br>Kannada and 99% accuracy for Kannada-Malayalam. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4. | Speech Recognition using HTK Toolkit for Marathi Language | Supriya S.Chavan Dr.S.M.Handore | Pre-processing: Speech Enhancement, Linear Predictive Coefficients, Speech Segmentation, Feature Extraction: Hidden Markov Modelling Tool Kit 3.4, MFCC, LPC Post-processing: Speech Recognition: ANN | 80% | 910 Marathi sentences, recorded by 3 male and 3 female speakers | Speech Recognition accuracy can be Improved as Current accuracy is 80%. Dataset Size & variety can be Increased. | Higher accuracy achieved using simple Techniques |
| 5. | Multiclass Classification and Class based Sentiment Analysis for Hindi Language | Prof.Sumitra Pundlik, Prachi Kasbekar, Gajanan Gaikwad, Prasad Dasare, Akshay Gawade, Purushottam Pundlik | HindiSentiWordNe t(HSWN) Language Mode(LM) Classifier | | Speeches from leaders across various Topics converted to text files. | Self Learning ontology not Implemented. Works only with text documents with Word limitations | Can be used by creating new Ontologies. Gives better accuracy with Combined use of LMClassifier & HSWN |
| 6. | Isolated Spoken Marathi Words Recognition using HMM | Sai Sawant and Mangesh Deshpande | **Pre-processing:** Training and Testing the model **Post-Processing:** Hidden Markov Models (HMMs) | 86.5% | 2,000 Marathi words spoken by ten native speakers | To increase the recognition performance and speaker independence of this system, training dataset needs to be increased by using data obtained from increased number of speakers. | In this work, isolated Marathi word recognition is performed with limited number of phonemes. It is observed that good recognition results are obtained when both speakers and test data are known. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7. | Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing ICMIA | Mrs. Rushali Dhumal (Deshmukh) and Dr. Arvind Kiwelekar | **Pre-processing:** Early Dropping and Drop out Techniques are used.<br><br>**Post–Pre-processing:** Python NLTK, Scikit-learn | 85% for the deep learning model<br><br>and 97% for the Bi-LSTM model. | Pos - tagged corpus containing 1500 sentences with 10115 words has been created using Unified Parts of Speech from Marathi Newspapers. | | Pos - tagged corpus containing 1500 sentences with 10115 words has been created using Unified Parts of Speech from Marathi Newspapers. |
| 8. | Urdu Sentiment Ana lysis With Deep Learning Methods | Lal Khan, Ammar Amjad, Noman Ashraf, HSIEN-TSUN G CHANG | **Pre-processing:** Stop Words, Normalization<br><br>**Post-Processing:** SVM, NB, RF, AdaBoost, MLP, LR, 1D-CNN, and LSTM | 82.05% | The data of Movie and electronic reviews. Urdu text from Urdu news websites. | One of the limitations of this study is that it includes only positive and negative classes; | Paper achieve the highest F1 score of 82.05% using LR with combination of features. The SVM classifier is the second highest performer for this task and its average performance is better than all other classifiers. |
| 9. | A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter | MONDHER BOUAZIZI, TOMOAKI OHTSUKI | **Pre-processing-** By removing the URLs, tags, Tokenization **Post-Processing**: Random Forest, Binary Classification, Ternary Classification | 60.2% | **Set 1**: 21000 tweets **Set 2:** 19740 tweets | The proposed paper haven't used the results obtained through Ternary Classification | The obtained results show some potential: the accuracy obtained for multi-class sentiment analysis in the data set used was 60.2%. |
| 10. | Emotional voice conversion using multitask learning with Text-To-Speech. Information and Electronics Research Institute, KAIST, Daejeon, | Tae-Ho Kim , Sungjae Cho, Shinkook Choi, Sejik Park and Soo-Young Lee | **Pre-processing:** Voice activity detection algorithm<br><br>**Post-Processing :** Attention,decoder | | male-Korean-Emotional-Text-to-speech (mKETTS) dataset. | In future need to focus on improvement of TTS by VC as pronunciation differs also focus on explicit loss to | In this work, higher accuracy is achieved using multitask learning also using this it has |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Republic of Korea | | | | | keep down the difference between TTS and VC. | minimized WER. |
| 11. | CubeMaha Sent: A Marathi Tweet-based Sentiment Analysis Dataset | Atharva Kulkarni, Meet Mandhane , Manali Likhitkar , Gayatri Kshirsagar, and Raviraj Joshi | Algorithms Used: CNN, BiLSTM+Global MaxPool , ULMFiT , BERT | 3 class accuracy – 84.13 , 2 class accuracy – 92.93 | Marathi Sentiment Analysis Dataset - L3CubeMahaSent | The dataset does not retain any context of the tweets such as the tweeting profile, time of posting, region, etc. | 1. It is the first major publicly available dataset for Marathi Sentiment Analysis.<br><br>They have performed 2-class and 3-class sentiment classification to provide a benchmark for future studies<br><br>2. The use of pre-trained embeddings significantly reduces overfitting |
| 12. | Spontaneous Emotion Recognition for Marathi Spoken Words International Conference on Communication and Signal Processing | Vaibhav V. Kamble, Bharatratna P.Gaikwad, Deepak M.Rana | Principal Component analysis, Linear Discriminate Analysis, Mel-frequency scale analysis, cepstrum, Dynamic feature extractions | 83.33% | Berlin Emotion database | 1.The feature vectors sample speech signal are never exactly the same as those provided during the enrollment or training phase for the same user.<br><br>2.If the inferior database is used as an input to the system then incorrect conclusion | This proposed work of (ERFMSW) system based on feature extraction got more accuracy.<br><br>2.The database that is used in this is collected from the real life situations |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | may be drawn | |
| 13. | Vachantar–Lokbhasha: A Speech to Text Conversion for Marathi | Archana. V. Chechare | Mel Frequency Cepstral Coefficient (MFCC), Artificial Neural Networks (ANN) | 87.76% | | As in this paper recognition is in Marathi it has been performed in Offline and Runtime modes. The accuracy of the system in offline mode is more as compared to runtime mode. | 1.In this the text will is saved to a file and that file can be further used.<br><br>2.Automatic speaker recognition is used to extract and recognize the information about speaker's speech |
| 14. | Multi-Class Sentiment Analysis in Twitter: What if Classification is not the Answer | Mondher Bouazizi, Tomoaki Ohtsuk | **Pre-processing:** Features Extraction<br><br>**Post-processing:** Naïve Bayes Classifier , Random forest classifier , Iterative Dichotomiser 3 classifier | 77.4% | Dataset is made of tweets collected using Twitter API | Need to address the case of tweets with sentiments belonging to different polarities and try to find possible ways to identify these sentiments. | Using the current version of SENTA we have found best fitting in the context of multi-class sentiment analysis(mainlyRandom Forest). |
| 15. | Natural Language Processing based Rule Based Discourse Analysis of Marathi Text | Kalpana Khandale, C Namrata Mahendar | **Pre-processing:** Discourse Analysis,<br><br>**Segmentation:** POS tagger<br><br>**Post-processing:** Anaphora Resolution | 85.43% | Marathi Balbharati book from standard 1st to 4th class | This paper highlighted to the issues and rule based structure to resolve the discourse anaphora based on Marathi Grammer | Total 515 sentences from which 440 discourse sentences are resolved properly. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 16. | Novel Technique for Script Translation using NLP: Performance Evaluation | Darshana Patil, S.B. Chaudhari, Sharmila Shinde | **Pre-processing:** E-M Cross conversion system,which is based on query translation approach<br><br>**Post–processing:** Hybrid translation | 91% | Marathi WordNet | Can implement best hybrid system for multi language translation using NLP in future | It fully supports the use of an automated technology-based translation system to assist in cross-domain data collection to solve the challenge of several persons. |
| 17. | A Spell-checker Integrated Machine Learning Based Solution for Speech to Text Conversion | H.M Mahmudul Hasan, Md.Toufiqe Hasan, Md.Adnaul Islam, Md.Araf Hasan | **Pre-processing:** TensorFlow<br><br>**Segmentation:** Deep Speech<br><br>**Post–processing:** Peter Nerving's correct spelling suggestion algorithm | Before applying spell corrector: 30-54%<br><br>After applying spell corrector: 45-67% | Bengali Dataset: "Bengali.ai" & English Dataset: "Mozilla Common voice" | Can be focused on developing a larger and more effective dataset for the Bengali language so that the machine can process data better and faster | |
| 18. | Kannada Speech to Text Conversion Using CMU Sphinx | Shivakumar K.M, Aravind K.G, Anoop T.V, Deepa Gupta | **Pre-processing:** CMU Sphinx **Feature Extraction:** Nfilt parameter (with value 52), **Post-processing:** Context Dependent Model(Single User Speech) | Context Dependent Model: 95%<br><br>Context Independent Model: 80% | Speech corpora for Kannada created by IIIT Hyderabad | These system holds accuracy for Kannada sentences with only four to ten word length. | This system investigates extensibility of recognizing all letters and morphological variants of spoken Kannada words. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 19. | Cognitive Devanagari (Marathi) Text to Speech System | Shaikh Shadab Shakil, Manjare Chandraprabha Anil | **Pre-processing:** Mapper & Combiner for input text<br><br>**Feature Extraction:** Optical Character Recognition(OCR) system<br><br>**Post-processing:** Text-to-Speech (TTS) Synthesizer | | Database creation can be done in MS excel. | In future can be focused on the signals which we get here as the output can be thus reverse engineered and can be useful as the base for a TTS engine itself. | We can derive or show graph outputs for sentences or phrases. |
| 20. | Hindi Sentence Classification for Expressive Storytelling Systems | Shivli Agrawal, Yukti Kirtani, Yuki Girdhar and Swati Agrawal | **Pre-processing:** RNN-LSTM<br><br>**Feature Extraction:** Word2Vec<br><br>**Post-processing:** CNN-SVM | Overall accuracy: 88.1% | Hindi language stories from texts like "Panchatantra" & "Akbar Birbal" | In future can be focused on improving by incorporating various other discourse modes and taking their subclasses into account. | Comparing to present Hindi Story Classification using CNN-RNN provides better user experience than the existing CNN-SVM system |
| 21. | Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning | LI YANG, YING LI , JIN WANG, R.SIMON SHERRATT | **Pre-processing:** Sentiment lexicon<br><br>**Feature Extraction:** CNN model & GRU model<br><br>**Post-processing:** SLCABG model | weighted word vector: 93.5%<br><br>unweighted word vector: 92.8% | The data of book reviews collected from "Dangdang" using web crawler technology | Can be focused to study the sentiment fineness classification of text. | By analyzing the experimental results, it can be found that the SLCABG model has better classification performance than other sentiment analysis models. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 22. | UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods | UZMA NAQVI, ABDUL MAJID, AND SYED ALI ABBAS | **Pre-processing:** Data Cleaning & Tokenization **Feature Extraction:** Word Embedding Models **Post-processing:** DL Models(LSTM, BiLSTM-ATT, CNN, and C-LSTM) | 77.9% | Urdu blogs and news websites such as BBC, DW, Express, Dunya and humsub | In future can be focused on DL models with more embeddings that provide better contextual information along with increased and balanced dataset will be explored | It has been observed that by using the Samar embedding model, DL models performance improved. It also highlights the importance of embedding on classification task |
| 23. | A Research Work on English to Marathi Hybrid Translation System | Pramod Salunkhe, Mrunal Bewoor, Dr.Suhas Patil | **Pre-processing:** Statistical Translation **Post-Processing:** Hybrid Translation | Average Precision : 87.5% | statistical system built on corpus downloaded from IITB produced by Pushpak Bhattacharya | Further research can be made in better translation with application of Marathi wordnet or development of trans-lingual word net for all | In Hybrid approach,we can achieve better result output of statistical system is corrected in comparison to Rule based system. |
| 24. | SENTIMENT ANALYSIS OF MARATHI LANGUAGE | Sujata Deshmukh,Nileema Patil, Surabhi Rotiwar, Jason Nunes | **Pre-processing:** Segregation and lemmatization **Post-Processing:** Attribute Polarity | With Yandex translator: 60-70% | English SentiwordNet | Further real time update of dictionary can be future research direction in the field of sentiment analysis of Marathi language. | Can be used to calculate the cumulative polarity of the text and rank the sentence as positive, negative or neutral on a set scale standard. |

## 4. CONCLUSION

Sentiment Analysis has been popular and has led to building of better products & services, understanding user's Sentiment and opinion, and for taking data driven business decisions. With rapidly increasing technology, people tend to rely more on reviews, and opinion of other people towards certain product or service. The rise in user generated data for Marathi language across various domains like news, culture, arts, sports etc has opened the data to be explored and mined effectively. The lack of datasets and proper model is one of the biggest challenges while dealing with sentiment analysis for Marathi language. In this paper we surveyed different techniques used for analysing speech and text data of various languages while comparing it with Marathi language. We also studied its usefulness in analysing a large number of human generated text to drive valuable insights.

## REFERENCES

[1] A. Alibasic and T. Popovic, "Applying natural language processing to analyze customer satisfaction," 2021 25th International Conference on Information Technology (IT), 2021, pp. 1-4, doi: 10.1109/IT51528.2021.9390111.

[2] D. Bharti and P. Kukana, "A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 491-496, doi: 10.1109/ICOSEC49089.2020.9215376.

[3] A. N..P and P. S.S., "Deep Learning Based Language Identification System From Speech," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1094-1097, doi: 10.1109/ICCS45141.2019.9065370.

[4] S. Supriya and S. M. Handore, "Speech recognition using HTK toolkit for Marathi language," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 1591-1597, doi: 10.1109/ICPCSI.2017.8391979.

[5] 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.

[6] S. Sawant and M. Deshpande, "Isolated Spoken Marathi Words Recognition Using HMM," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697457.

[7] R. Dhumal Deshmukh and A. Kiwelekar, "Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 76-81, doi: 10.1109/ICIMIA48430.2020.9074941.

[8] L. Khan, A. Amjad, N. Ashraf, H. -T. Chang and A. Gelbukh, "Urdu Sentiment Analysis With Deep Learning Methods," in IEEE Access, vol. 9, pp. 97803-97812, 2021, doi: 10.1109/ACCESS.2021.3093078.

[9] M. Bouazizi and T. Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter," in IEEE Access, vol. 5, pp. 20617-20639, 2017, doi: 10.1109/ACCESS.2017.2740982.

[10] T. -H. Kim, S. Cho, S. Choi, S. Park and S. -Y. Lee, "Emotional Voice Conversion Using Multitask Learning with Text-To-Speech," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7774-7778, doi:10.1109/ICASSP40776.2020.9053255.

[11] Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Raviraj Joshi, "L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset," arXiv:2103.11408v2 [cs.CL] 22 Apr 2021

[12] V. V. Kamble, B. P. Gaikwad and D. M. Rana, "Spontaneous emotion recognition for Marathi Spoken Words," 2014 International Conference on Communication and Signal Processing, 2014, pp. 1984-1990, doi: 10.1109/ICCSP.2014.6950191.

[13] Archana. V. Chechare, "Vachantar –Lokbhasha: A Speech to Text Conversion for Marathi," in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 7, July 2016

[14] M. Bouazizi and T. Ohtsuki, "Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer," in IEEE Access, vol. 6, pp. 64486-64502, 2018, doi: 10.1109/ACCESS.2018.2876674.

[15] K. B. Khandale and C. N. Mahender, "Natural Language Processing based Rule Based Discourse Analysis of Marathi Text," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 356-362, doi: 10.1109/ICESC48915.2020.9155653.

[16] D. Patil, S. B. Chaudhari and S. Shinde, "Novel Technique for Script Translation using NLP: Performance Evaluation," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 728-732, doi: 10.1109/ESCI50559.2021.9396969.

[17] H. M. M. Hasan, M. A. Islam, M. T. Hasan, M. A. Hasan, S. I. Rumman and M. N. Shakib, "A Spell-checker Integrated

Machine Learning Based Solution for Speech to Text Conversion," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1124-1130, doi: 10.1109/ICSSIT48917.2020.9214205.

[18] K. M. Shivakumar, K. G. Aravind, T. V. Anoop and D. Gupta, "Kannada speech to text conversion using CMU Sphinx," 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, pp. 1-6, doi: 10.1109/INVENTIVE.2016.7830119.

[19] S. S. Shakil and M. C. Anil, "Cognitive Devanagari (Marathi) Text-to-Speech System," 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 758-762, doi: 10.1109/ICCUBEA.2015.151.

[20] S. Agrawal, Y. Kirtani, Y. Girdhar and S. Aggarwal, "Hindi Sentence Classification for Expressive Storytelling Systems," 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 2019-2025, doi: 10.1109/SSCI.2018.8628858.

[21] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[22] 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.

[23] Pramod Salunkhe, Mrunal Bewoor, Dr.Suhas Patil. "A Research Work on English to Marathi Hybrid Translation System," in International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2557-2560

[24] Sujata Deshmukh,Nileema Patil, Surabhi Rotiwar, Jason Nunes, "SENTIMENT ANALYSIS OF MARATHI LANGUAGE," in International Journal of Research Publications in Engineering and Technology, VOLUME 3, ISSUE 6, Jun.-2017

## BIOGRAPHIES

**Prof. Vina M. Lomte**
Project Guide and Head of Computer Engineering Department at RMD Sinhgad School of Engineering, SPPU, Pune.

**Mr. Pratik Jadhav**
Project Team Lead and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.

**Mr. Onkar Kalshetti**
Project Research Fellow and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.

**Ms. Sonali Deshmukh**
Project Research Fellow and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.

**Mr. Arjun Jadhav**
Project Research Fellow and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.