IT350 - Stance Detection in Telugu-English Code-Mixed Social Media Data

G R Revanth Varma
Information Technology
NIT Karnataka
Surathkal, India 575025
geddamrevanth.201it122@nitk.edu.in

E Vennela
Information Technology
NIT Karnataka
Surathkal, India 575025
evennela.201it218@nitk.edu.in

Saride Madesh
Information Technology
NIT Karnataka
Surathkal, India 575025
saridemadesh.201it254@nitk.edu.in

***Abstract*—Stance detection involves identifying the speaker's opinion towards a particular topic. This task has considerable attention in recent years due to its potential applications in various domains. While there has been significant research on stance detection in English, very little has been done in other languages, including Telugu. In this study, we focus on local, national, and international issues in the Andhra Pradesh and Telangana geographic regions, such as RRR Oscar, Vandebharat Express, Vishakapatnam as Capital, YS Jagan Mohan Reddy as CM, and Ujjwala Yojana.Our stance detection algorithm has been specifically designed to utilize a three-category classification system, which categorizes user comments as either in favor, against, or neutral. We perform stance detection using various machine learning models and compare and analyze the results.**

***Index Terms*—Stance detection, Telugu-English code-mixed language, Social media data, Natural language processing, Machine learning, Three-label system, Performance evaluation**

I. Introduction

Stance detection is a critical task in natural language processing that aims to identify the speaker's attitude towards a particular topic. It has gained attention in recent years due to its applications in various domains, including politics, social media analysis, and market research. Stance detection involves classifying a given text into one of three categories: favor, against, or neutral, based on the speaker's stance towards a given topic.

While much research has been conducted on stance detection in English, there has been limited research on other languages, including Telugu, a language spoken by millions of people in India. Telugu is the most spoken common language in the southern state of Andhra Pradesh and Telangana. With the rise of digital platforms and social media, more and more Telugu-speaking users are expressing their opinions on various issues like politics, entertainment, sports, new technology, etc.

In this context, the task of stance detection in Telugu is becoming increasingly important. However, due to the lack of annotated data and resources for Telugu language processing, stance detection in Telugu poses significant challenges. Therefore, the development of effective approaches to detect the stance of Telugu texts towards a given topic is of great significance for various applications. This paper aims to address thechallenge of stance detection in Telugu and proposes a deep learning-based approach to automatically detect the stance of Telugu texts towards various trending issues in Andhra Pradesh and Telangana regions.

Stance detection systems are widely used in information retrieval, opinion mining, text entailment, and other natural language processing applications. By identifying the speaker's position on a particular topic or issue, stance detection techniques can provide valuable insights into user attitudes and opinions, which can be used to inform various applications in different domains.public reaction in social media can help the government evaluate their new election policies. Stance detection techniques can be used to analyze user opinions towards political candidates, policies, and issues, which can help political parties tailor their campaigns to appeal to different voter segments.

Target: RRR Oscar
Comment: *Prapanchavyaptanga bhāratīya calanacitra parisramalo atyanta saktivantamaina citranni manam chustunnam*
[English: We are experiencing the most powerful film in Indian Film Industry all over the world]

Fig. 1. Target

The above example shows a target-comment pair for target entity RRR Oscar. RRR Oscar is a new trending topic in entertainment. Here one of the twitter users expressed his stance towards the target unit. Clearly, from above example, this user is expressing favor towards target entity RRR Oscar. Compared to sentiment detection, stance detection is more meaningful for analyzing the user's opinion towards a target entity.Stance refers to the speaker's perspective on a particular issue or topic.Sentiment, on the other hand, refers to the speaker's emotional response to a particular topic or issue. It involves classifying the speaker's opinion as positive, negative, or neutral. It involves classifying the speaker's position as favor, against, or neutral. While both approaches involve identifying the speaker's attitude or opinion, they are used in different contexts and have different goals.

II. Objective

Modeling of an efficient stance detection system for telugu language. To develop a system that can accurately identify the speaker's attitude towards a given topic or issue. The system should be able to analyze user comments or posts in Telugu language and classify them into one of three categories: favor, against, or neutral.

The primary goal of stance detection is to provide insights into user opinions and attitudes towards specific topics or entities. The Telugu language is spoken by millions of people in India, and understanding user attitudes towards different issues can be valuable for various applications.

1. Collecting a data set of Telugu language comments or posts related to different issues or entities.

2. Pre-processing the Telugu language data to prepare it for analysis.

3. Implementing various machine learning models to classify the Telugu language data into one of three categories: favor, against, or neutral.

4. Evaluating the performance of the stance detection system

III. Literature Survey

"Sentiment Extraction from English-Telugu Code Mixed Tweets Using Lexicon Based and Machine Learning Approaches"[2]in this research work, This research study involves the collection and analysis of political tweets related to Andhra Pradesh. The accuracy of a machine learning approach em- ploying the SVM algorithm is compared with that of a lexicon- based approach utilizing the Sentiment Extraction from Telugu Code Mixed Tweets (STCMT) algorithm.

The objective is to provide a comparative analysis of the effectiveness of these two approaches in accurately detecting the sentiment of political tweets in the Telugu language. This aims to contribute to the advancement of sentiment analysis techniques in the field of natural language processing.

Research paper " A Tutorial on Stance Detection"[3] This research paper aims to explore the fundamental concepts and associated research problems concerning stance detection. This provides an overview of the historical and contemporary approaches employed in the field of stance detection, including the shared tasks and tools utilized.

Additionally, the paper discusses related datasets that can be used in the development of stance detection techniques. The study also covers open research directions and potential application areas of stance detection, providing insights into the advancement of natural language processing techniques. The purpose is to provide a comprehensive understanding of stance detec- tion and its significance in the analysis of social media data.

Another paper "A systematic review of machine learning techniques for stance detection and its applications"[7] This research paper presents a comprehensive analysis of existing studies related to stance detection. The authors have classified the analyzed studies based on a taxonomy of six dimensions, which include approaches, target dependency, applications, modeling, language, and resources. The study examines the- various techniques employed in each dimension and provides a detailed analysis of their respective strengths and weaknesses. By analyzing the studies from each dimension's perspective, the research paper aims to provide a better understanding of stance detection and contribute to the development of more effective techniques. The objective of this is to provide insights into the current state of stance detection and identify areas that require further research to improve the accuracy and effectiveness of the techniques employed.

We present stance detection system for telugu language. The dataset contains telugu language code-mix text comments for five target entities. Stance detection system is implemented three labels.Different machine Learning models are implemented and analysed.

IV. Creating Dataset to Identify Stance

In this section, we will illustrate the various steps involved in dataset creation. The first step is to collect data for the targets included, followed by preprocessing which includes filtering out the tweets, and comments for the target entity and removing other types of scripts. Finally, stance annotation is carried out.

A. Target Selection

The stance detection system is implemented for five present trending local and national, international current issues in AndhraPradesh, Telangana geographic regions like RRR Oscar, Vandebharat Express, Vishakapatnam as capital,YS Jagan Mohan Reddy As CM and Ujjwala Yojna. Targets are selected based on popularity level of local and national,International news.

Above five targets had more comments count compare to other social issues. Table I gives the brief description of each target.

TABLE I
DESCRIPTION ABOUT TARGETS

Target	Description
1) RRR Oscar	SS Rajamouli's RRR has created history by becoming the first Indian feature film to win an Oscar for the best original song 'Naatu Naatu'.
2) Vande Bharat Trains	The Vande Bharat Express trains, formerly known as Train 18, are India's first indigenous semi-high-speed trains.
3) Visakhapatnam as capital	On 31 January 2023, it was announced that the city will become the capital of Andhra Pradesh.
4) Y S Jagan Mohan Reddy	He is the founder,president of the Indian political party, YSR Congress Party (YSRCP). He has been ruling Andhra Pradesh since 2019.
5)Ujjwala Yojna	This is an Indian government scheme that provides free LPG connections to below-poverty-line households.

B. Collection of Data

The dataset for stance detection system is collected from Twitter,YouTube using "Twitter data scraping" or "Twitter data mining". It involves using the Twitter API (Application Programming Interface) to access the Twitter platform and collect

data such as tweets, user profiles, and other information. To access the Twitter API, developers need to obtain API keys, which are unique codes that allow them to interact with the API and collect data.

Once the API keys are obtained,we can use Python to write scripts that connect to the Twitter API and retrieve data.Similarly from YouTube.Twitter is popular social media site for users to express an opinion on target entity. Compare to Facebook, Twitter offers easy tools for the user to express his stance towards target interest. Twitter offers API to extract contents like posts, comments, user name, created time,etc. The code is written in Python language.Tweets and Comments related to five targets are extracted.

c. Telugu Social Media Text Properties

Telugu uses a unique script called Telugu script, which is derived from the Brahmi script. The script consists of 56 characters, including vowels, consonants, and vowel-consonant combinations.Telugu is an agglutinative language, which means that words are formed by adding prefixes, suffixes, and infixes to root words. This results in a large number of possible word forms and a complex morphology.

Text data from social media sites can be challenging to process due to the informal language, non-standard grammar, and the use of multiple scripts. Social media users often mix different scripts, including Roman script, native script, mixed script, and code-mixed script when communicating. This presents a unique challenge for natural language processing (NLP) tasks.

Language class	Script type	Comment Text
telugu	Roman Script (code-mix)	Desham kosam darmam kosam ticket retu kuda chaala ekkuva vunnai
	Native Script	దేశం కోసం, ధర్మం కోసం టికెట్స్ రేట్లు కూడా చాలా ఎక్కువ ఉన్నాయి
Mixed Language	Mixed Script Telugu-English	Train to Busan సినిమా లో కూడా రైలు ఇలానే ఉంటుంది.
	Code-Mix Telugu-English	Very nice. వందేభరత్ కి స్పీడ్ కి తట్టుకునే ట్రాక్స్ ఉండలి కద ఉంటే అ జర్నీ యే వేరు

Fig. 2. Various Different Types Of Scripts

D. Preprocessing and Cleaning

In order to convert the data to a favorable format for performing stance detection task, several changes were made to the collected tweets. Preprocessing steps such as removal of emojis, hashtags and URL links were implemented. In the tweets, URL links do not add any information to the data and were hence, removed. Additionally, hashtags were removed so that the models would classify effectively without any bias towards the tags used for filtering the tweets while scraping.

E. Annotating tweet for Stance

Annotating tweets,comments for stance in Telugu language involves identifying the attitude or perspective of the commenter towards a particular topic or issue. This can be done by analyzing the language used in the tweet,comment and identifying the specific words, phrases, or grammatical constructions that express the commenter's stance.

To annotate comments for stance in Telugu language, the annotator would need to be proficient in Telugu language and have knowledge of the specific topic or issue being discussed. They would need to read and analyze the comment carefully to identify the stance being expressed and mark it using appropriate annotation tags that is favour(FV),against(AG),neutral(NR).

Overall, annotating tweets,comments for stance in Telugu language requires a combination of language proficiency, subject matter expertise, and familiarity with annotation tools and techniques. It is an important task for getting accurate results.Figure 3 illustrate annotation task.

Target	Comment	Annotation Class
RRR Oscar	నాటు నాటు పాట చాలా బాగుంది. ఈ పాటకు ఆస్కార్ అవార్డు రావడం భారతీయులందరికీ గర్వకారణం. (The song" naatu naatu" is very good. Getting Oscar award for this song is a proud moment for all Indians)	Favour (FV)
RRR Oscar	దీని కంటే చాలా ఇతర పాటలు అర్థవంతంగా ఉన్నాయి (There are many meaningful songs compared this)	Against (AG)
RRR Oscar	ఇది తెలుగు పాట (This is a telugu song)	Neutral (NR)

Fig. 3. Stance Annotation Procedure

V. Statistics of stance dataset

This section presents the statistics of the stance dataset. The annotated datasets follow an 80-20 rule, where 80% of the dataset is used for training and 20% for testing.

Figure 4 shows the statistics of the dataset for three different class labels, which correspond to the target entities. The FV label represents comments that express a favorable stance, AG label represents comments that express an unfavorable stance, and NR label represents comments that are neutral in stance.

Target	Statistics of dataset						
	Train			Test			Total
	Favour	Against	Neutral	favour	Against	Neutral	
RRR Oscar	447	102	251	97	44	59	1000
Vande Bharat Trains	379	75	345	85	49	66	1000
Visakhapatnam as capital	287	366	287	65	72	63	1000
Y S Jagan Mohan Reddy	550	140	110	107	39	54	1000
Ujjwala Yojna	320	192	288	88	67	88	1000

Fig. 4. Dataset Statistics

VI. Methodology

A. Source For Data Collection

Nowadays social media platforms like Twitter and YouTube have emerged as valuable sources for data collection in the field of stance detection in Telugu. These platforms provide a vast amount of user-generated content in Telugu, which can be used to identify the sentiment and stance of users on various topics.

The use of Twitter and YouTube data for stance detection has become increasingly popular due to their accessibility, popularity, and ability to capture real-time opinions and sentiments. Furthermore, these platforms have a diverse user base, providing a broad range of viewpoints and opinions. Therefore, researchers can leverage the vast amount of data available on these platforms to train machine learning models and improve the accuracy of stance detection in Telugu.

Overall, Twitter and YouTube are valuable sources for data collection for stance detection in Telugu, providing researchers with a wealth of information to better understand the opinions and stances of Telugu speakers on a variety of topics.

B. Data Collection and Removing Noise

1) Twitter Data Collection: We used Python script that uses the Tweepy library to extract tweets from Twitter. Here is an explanation of the steps involved:

1.Authentication: The script sets up authentication credentials using the consumer key, consumer secret, access token, and access token secret provided by Twitter.

2.API object: The script creates an API object using the Tweepy library and the authentication credentials.

3.Search query: The script defines the search query as "Target" and specifies that it should only retrieve tweets in the Telugu language.

4.Retrieving tweets: The script uses the API object to retrieve a specified number of tweets that match the search query. It iterates over each tweet and extracts the text of the tweet, which is then added to a list of tuples.

2) YouTube Data Collection: We used Python script that extracts comments from a list of YouTube videos using the YouTube API.

1.The script uses the googleapiclient library to build an authenticated connection to the YouTube API, and specifies the api-service-name and api-version to be used. The DEVELOPERKEY variable stores the API key that is used for authentication.

2.The script then loops through a list of videoids, which correspond to the YouTube videos from which comments are to be extracted. For each video, the script sends a commentThreads().list() request to the API, passing in the video ID as a parameter, and specifying that both the snippet and replies parts of the comment thread should be included in the response. The maxResults parameter is set to 50 to limit the number of comments returned in each API response.

3.The script then loops through the items in the API response, extracting the text of each comment and any replies to that comment, and adding them to a comment-list list. If there are more comments than can be returned in a single API response, the script sends additional requests to the API using the page-Token parameter to retrieve the next page of results.

3) Removing noise: Remove Noise from the tweets or comments.These noises can include URLs,emojis,Special characters,extra spaces,typos, misspellings, grammatical errors, punctuation errors, formatting errors, repeated words, filler words (like "um" and "ah"), false starts, and other extraneous

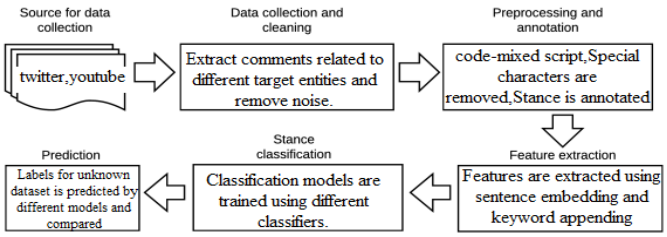

Fig. 5. Proposed Methodology

content that does not add meaning to the text. Removing these noises from text can improve the readability and clarity of the content and make it easier to understand.

C. Preprocessing and Annotating

The script uses Telugu regex patterns to filter out any non-Telugu comments, as well as regex patterns to remove URLs from the comments. It then loops through each comment in the dataset and checks if it contains any Telugu letters. If the comment does contain English Language the code drops the row.If the comment does contain Telugu letters,Roman Script (code-mix),Native Script,Mixed Script(Telugu-English) it applies the following preprocessing steps to clean the text:Remove URLs,emojis,non-Telugu characters,extra spaces.

Training dataset is manually annotated for three class label favor, against and neutral.Before the application of classification algorithm, the first step is feature extraction after preprocessing and tag annotation.

D. Feature Extraction

It takes the train data set with Telugu tweets and their corresponding stances. Then uses a pre-trained sentence embedding model (Paraphrase-XLM-R) to encode the tweets into dense vector representations.Next, it extracts keywords from each tweet using a TF-IDF vectorizer. The keywords are also represented as dense vectors.Figure 6 shows the extracted keywords.

The sentence embeddings and keyword vectors are concatenated to create a feature vector for each tweet.The stances are encoded using label encoding (mapping each unique value to a numerical label).Finally, the sentence embeddings and keyword features are combined to create the final feature matrix, which will be used as input to a machine learning model.

E. Stance Classification

stance is classified using different Machine Learning Models.Those are

1) Bagging Classifier: Boosting is a popular ensemble learning technique that combines several weak learners to create a strong learner. In the context of stance detection in Telugu language, Boosting Classifier models can be used to classify the stance of a given tweet as either in favor of, against, or neutral towards a specific topic. Boosting works by iteratively

training a sequence of weak classifiers, where each subsequent classifier focuses on the samples that the previous classifiers misclassified. This approach allows the Boosting Classifier to learn a more accurate and robust decision boundary, resulting in better classification performance.

2) Gradient Boosting Classifier: Gradient Boosting is a ML algorithm that can be used for stance detection in Telugu language. It works by building a sequence of decision trees, where each subsequent tree is constructed to correct the errors of the previous tree. The algorithm starts by fitting a single decision tree to the data and making predictions. The errors made by the first tree are then used to train a second decision tree, and this process continues iteratively until a pre-defined stopping criterion is met. The final prediction is obtained by summing the predictions from all the decision trees.

Gradient Boosting Classifier can effectively handle non-linear relationships between the features and the target variable, and it can also handle missing data. In addition, it can automatically perform feature selection and feature scaling, which can improve the accuracy of the model.

3) AdaBoost Classifier: AdaBoost (Adaptive Boosting) is a popular boosting algorithm that is used to improve the performance of weak classifiers. In the context of stance detection in Telugu language, AdaBoost can be used to classify a tweet's stance as either supporting, opposing or neutral towards a given topic. AdaBoost works by combining multiple weak classifiers to form a strong classifier that can make accurate predictions.

The algorithm starts by assigning equal weights to all training examples, then trains a weak classifier on a subset of the training data. After each iteration, the algorithm adjusts the weights of the misclassified training examples to give them more importance in the next iteration.

This process continues till a maximum number of iterations or a desired level of accuracy is reached. The final model is a weighted combination of all the weak classifiers, with more weight given to the most accurate classifiers.

AdaBoost can be effective in handling imbalanced datasets, where there are more examples of one class than the others, and has been used successfully in natural language processing tasks, including sentiment analysis and text classification

4) Multi-layer Perceptron Classifier: The Multi-layer Perceptron (MLP) Classifier is a type of neural network that can be used for stance detection in Telugu language. The MLP Classifier consists of several layers of interconnected nodes, and each node performs a simple mathematical operation on its input, which is then passed on to the next layer. The MLP Classifier can learn complex patterns in data by altering the weights of the connections between nodes during training. This allows it to classify Telugu language tweets into different stances based on their features. The MLP Classifier can be trained using labeled Telugu language data and can achieve high accuracy in stance detection tasks.

5) Gaussian Naive Bayes Classifier: The Gaussian Naive Bayes Classifier is a probabilistic machine learning algorithm that uses Bayes' theorem to predict the likelihood of a given data point which belongs to a certain class based on its feature values.

Target	Class Label Type		
	Favour	Against	Neutral
RRR Oscar	గొప్ప (Great)	బాలేదు (Bad)	బానే ఉంది (Average)
	బాగుంది (Good)	చెండాలం (Worst)	చేసారు (Have Done)
	ప్రోత్సహనం (encouragement)	సమస్య (Problem)	సందేహస్పదమైన (Doubtful)
Vande Bharat Trains	బాగుంది (Good)	అవసరం లేదు (do not want)	ఎందుకు (why)
	మంచిది (Nice)	సమస్య (Problem)	చెడ్డది కాదు (Not bad)
	గొప్ప (Great)	ప్రమాదకరం (Dangerous)	ఏం జరిగింది (What Happened)
Visakhapatnam as capital	మంచిది (Nice)	వద్దు (not needed)	ఎందుకు (why)
	గొప్ప (Great)	చెండాలం (Worst)	ఎలా (how)
	బావుంది (Good)	అవసరం లేదు (do not want)	తటస్థ (Neutral)
Y S Jagan Mohan Reddy	బాగా (well)	బాలేదు (Bad)	సందేహస్పదమైన (Doubtful)
	మంచిది (Nice)	సమస్య (Problem)	చెడ్డది కాదు (Not bad)
	బావుంది (Good)	ప్రమాదకరం (Dangerous)	చెప్పలేము (Cannot predict)
Ujjwala Yojna	ఆదర్శం (Ideal)	చెండాలం (Worst)	ప్రభావం లేని (No effect)
	ప్రోత్సహనం (encouragement)	ప్రమాదకరం (Dangerous)	సమంగా (Neutral)
	అనుకూలం (favourable)	వద్దు (not needed)	ప్రభావం లేకుండా (Without influence)

Fig. 6. Keyword Feature For Each Target

In the context of stance detection in Telugu language, the classifier would use the training data to learn the relationship between the input features (such as bag-of-words or TF-IDF vectors) and the corresponding stance labels. During prediction, it would compute the conditional probability of the data point belonging to each possible stance label, and choose the label with the highest probability as the prediction.

Gaussian Naive Bayes assumes that the input features are independent and normally distributed, which may not always hold true in practice, but it can still be a useful and relatively simple algorithm for stance detection tasks.

6) Random Forest Classifier: Random Forest Classifier is an ensemble-based machine learning model used for classification tasks. In the context of stance detection in Telugu language, it can be trained on a labeled dataset of Telugu tweets to classify them into different stance categories such as "for", "against", "neutral", etc. The model works by building a forest of decision trees where each tree is learned on a different subset of the data and a random subset of the features. This process of creating diverse decision trees helps to reduce overfitting and increase the model's generalizability. During inference, each tree in the forest makes a prediction, and the final prediction is determined by taking a majority vote among all the trees. Overall, Random Forest Classifier is a powerful and versatile model that can be used for various classification tasks, including stance detection in Telugu language.

F. Prediction

The stance for the test data is predicted using the above mentioned machine learning algorithms(Bagging classifier as 'Bg',Gradient Boosting Classifier as 'GB',Ada Boost Classifier as 'AB',Multi-layer Perceptron Classifier as 'MlP',Gaussian Naive Bayes Classifier 'GNB',Random Forest Classifier 'RF'),Accuracy measures such as precision, recall, support and the F1 score is measured to test the accuracy of the implemented system.

VII. Experiment Results

This section describes various result measurements of implemented stance classification system.Various result measures like Precision, Recall,F1-Score,Support is measured to evaluate the performance of the stance classification system.

Table II is the results of accuracy measure for six different classifiers.According to the table we got highest Accuracy as 90% for Gradient Boosting Classifier particularly for vishakapatnam as capital Dataset.Among all Classifiers Gradient Boosting Classifier,Bagging Classifier is predicting well.

TABLE II
Accuracy for different classifiers

Targets	Classifiers	Accuracy (%)
RRR-Oscar	Bagging Classifier	73.75
	Gradient Boosting Classifier	**76.25**
	AdaBoost Classifier	75.10
	Multi-layer Perceptron Classifier	73.62
	Gaussian Naive Bayes Classifier	74.62
	Random Forest Classifier	73.50
Vandebharat	Bagging Classifier	**81.30**
	Gradient Boosting Classifier	79.30
	AdaBoost Classifier	73.70
	Multi-layer Perceptron Classifier	75.80
	Gaussian Naive Bayes Classifier	62.50
	Random Forest Classifier	80.80
Vishakapatnam as capital	Bagging Classifier	61.60
	Gradient Boosting Classifier	**89.8**
	AdaBoost Classifier	43.90
	Multi-layer Perceptron Classifier	57.50
	Gaussian Naive Bayes Classifier	49.50
	Random Forest Classifier	56.12
CM Y S Jagan	Bagging Classifier	**73.87**
	Gradient Boosting Classifier	68.12
	AdaBoost Classifier	71.87
	Multi-layer Perceptron Classifier	71.87
	Gaussian Naive Bayes Classifier	68.75
	Random Forest Classifier	73.12
Ujjwala yojana	Bagging Classifier	45.00
	Gradient Boosting Classifier	**88.30**
	AdaBoost Classifier	51.80
	Multi-layer Perceptron Classifier	56.30
	Gaussian Naive Bayes Classifier	45.00
	Random Forest Classifier	48.90

The overall accuracy of RRR-Oscar is 76% which we got but using Gradient Boosting Classifier, the accuracy of Vandebharat is 81% which we got by using Bagging, the accuracy of Vishakapatnam as capital is 90% Gradient Boosting Classifier, the accuracy of CM Y S Jagan is 74% by using Bagging Classifier and accuracy of Ujjwala Yojana is 88% by using Gradient Boosting Classifier.

A. Visualization of Accuracy

Accuracy of all the classifiers plotted and visualized. We have compared by considering **25 keywords** and **50 keywords**.

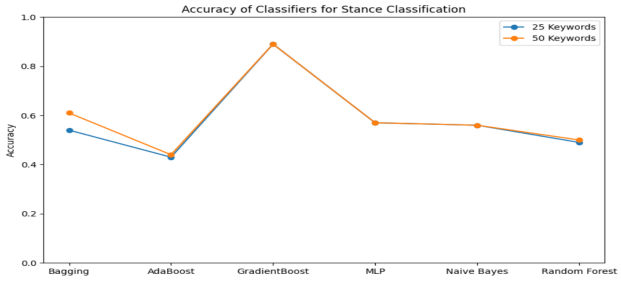

Fig. 7. Accuracy of different classifiers for **25 keywords** and **50 keywords**

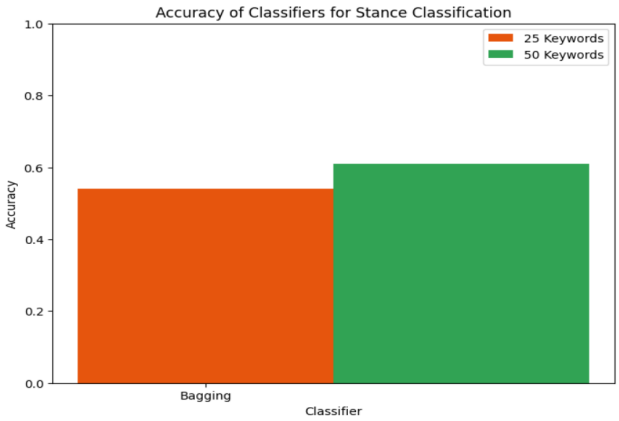

Fig. 8. Accuracy of bagging classifier **25 keywords** and **50 keywords**

As we can see there is a difference in the accuracy of the bagging classifier as with 50 keywords it's accuracy is 0.61 and with 25 keywords it is 0.54. the texts you are working with are relatively short and have a limited vocabulary, then reducing the keywords to 25 may still be sufficient to capture the most important words in the text. On the other hand, if the texts are longer and more complex then reducing keywords may result in a loss of important information and lower accuracy.

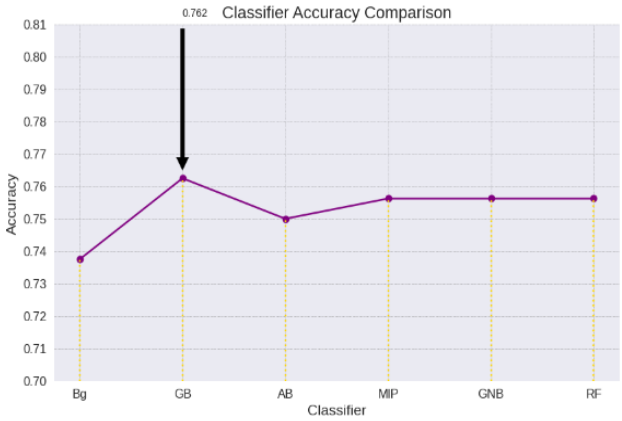

Fig. 9. Accuracy of different classifiers for RRR-Oscar Dataset

Accuracy of all above mentioned algorithms are considered and plotted for better visualization of accuracy of different models.

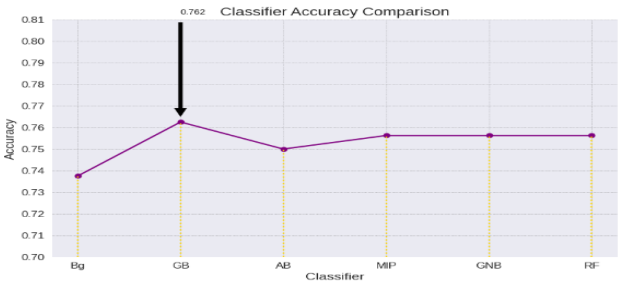

Fig. 10. Accuracy of different classifiers for RRR-Oscar Dataset

We took X-axis plane with classifiers as Bagging classifier as 'Bg',Gradient Boosting Classifier as 'GB',Ada Boost Classifier as 'AB',Multi-layer Perceptron Classifier as 'MlP',Gaussian Naive Bayes Classifier 'GNB',Random Forest Classifier 'RF'.And Y-axis contains the Accuracy.

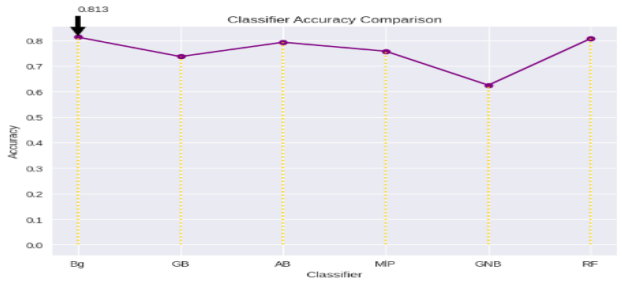

Fig. 11. Accuracy of different classifiers for VandeBharat Dataset

In Fig[10][12][14] we can see that Gradient Boosting Algorithm is providing best Accuracy.Almost all the remaining algorithms are providing almost similar accuracies.

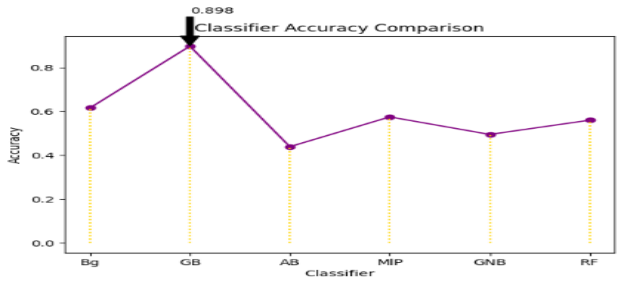

Fig. 12. Accuracy of different classifiers for Vishakapatnam as capital Dataset

In Fig[11][13] Bagging Classifier is providing best Accuracy,by the visualizations we can clearly see the difference in accuracy for different algorithms and compare them.

A confusion matrix is a useful ml method that allows to measure recall, precision, accuracy.It is a systematic way to allocate the predictions to the original classes to which the data originally belonged.Fig[14] shows the confusion matrix for the best classifier that is Gradient Boosting.Similarly All confusion matrix for each classifier is visualized for better analysis.

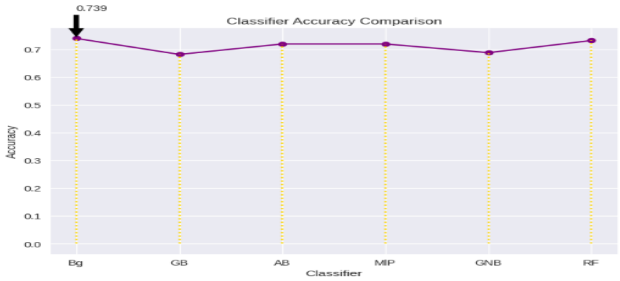

Fig. 13. Accuracy of different classifiers for CM Jagan Dataset

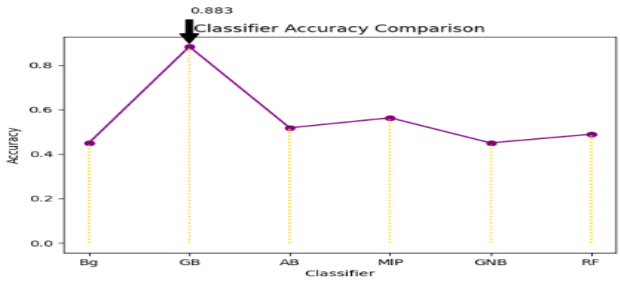

Fig. 14. Accuracy of different classifiers for Ujjwala Yojana Dataset

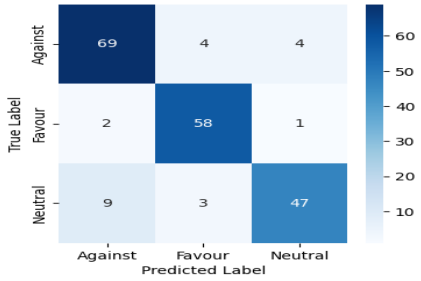

Fig. 15. Confusion matrix for Vishakapatnam Dataset

B. *Visualization of Support*

The support is the number of occurrences of each particular class in the true responses it is calculated by summing the rows of the confusion matrix.

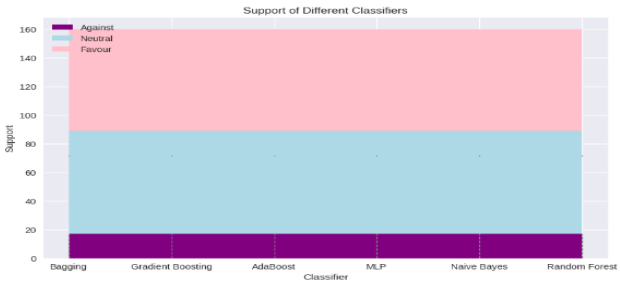

Fig. 16. Support Values for RRR-Oscar

In Figure[16] Support value for RRR-Oscar target is shown which explains that most of the people are in favour with getting Oscar,where some of them are against claiming that most of the other songs are good.

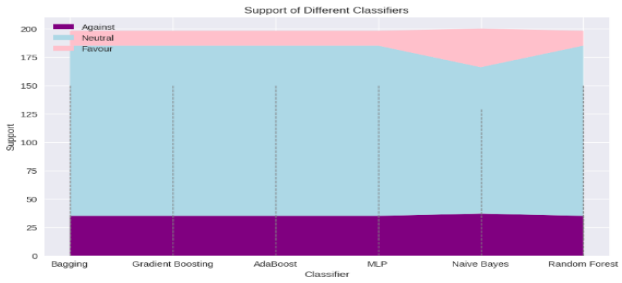

Fig. 17. Support Values for VandeBharat

In Figure[17] Support value for VandeBharat Express target is shown which explains that most of the people are in favour with the concept of speed trains,where some of them are against.

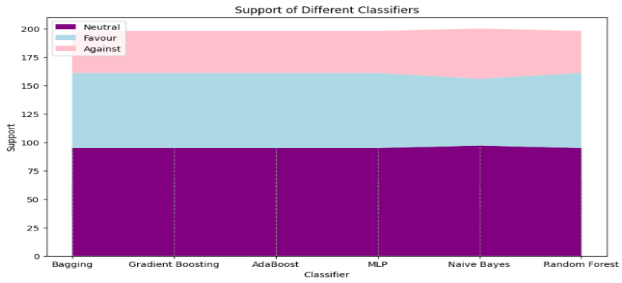

Fig. 18. Support Values for Vishakapatnam as Capital

In Figure[18] Support value for Vishakapatnam as Capital target is shown which explains that most of the people are against with this proposal,where some of them are in favour.

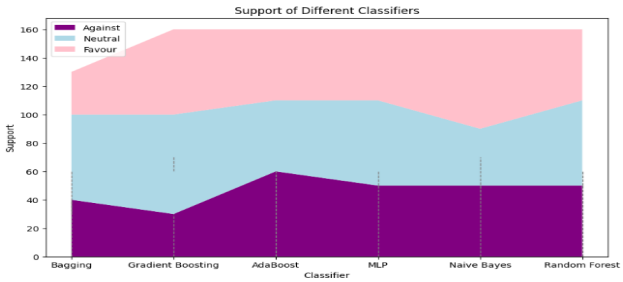

Fig. 19. Support Values for Y S Jagan as CM

In Figure[19] Support value for Y S Jagan as CM target is shown which explains that most of the people are in favour with the ruling of Y S Jagan Mohan Reddy,where some of them are against.

In Figure[20] Support value for Ujjwala Yojana target is shown which explains that most of the people are in favour with the concept of Ujjwala Yojana,where some of them are against.

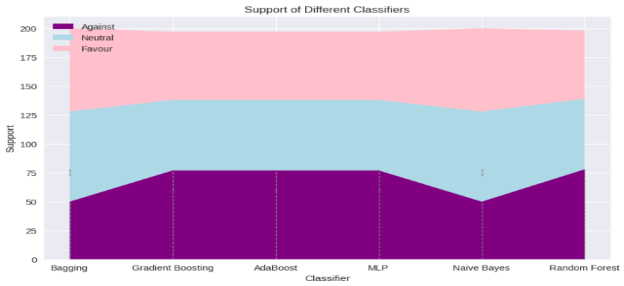

Fig. 20. Support Values for Ujjwala Yojana

VIII. Conclusion and Future Work

In conclusion, Stance Detection in Telugu-English Code-Mixed Social Media Data is a difficult task due to the complexities of the Telugu language and the unique features of code-mixed data. Twitter and YouTube are user review platforms that have vast amounts of user-generated content in Telugu-English code-mixed language.A manually annotated dataset consisting of 5000 Telugu texts was created to determine stance, with labels including ”favour”, ”against”, and ”neutral”. Six machine learning methods were employed to analyze the dataset, and it was determined that the Gradient Boosting Algorithm provided the most accurate results.The results of the analysis showed that people generally held a favourable stance towards all targets, except for the proposal to make Visakhapatnam the capital.These findings were then visualized to facilitate a clearer understanding of the results

In future work, we can try to implement the more advanced feature extraction algorithm like RNN based word embedding technique to increase the accuracy of the stance detection system.Deep learning-based methods CNN , LSTM may show promising results for stance detection in code-mixed data. Further research is needed to improve the performance of these methods and adapt them to different code-mixed languages.

References

[1] Skanda, V. Kumar, M. Kp, Soman. (2017). Detecting stance in kannada social media code-mixed text using sentence embedding. 964-969. 10.1109/ICACCI.2017.8125966.

[2] Kodirekka, A., Srinagesh, A. (2022). Sentiment Extraction from English-Telugu Code Mixed Tweets Using Lexicon Based and Machine Learning Approaches. In: Satyanarayana, C., Gao, XZ., Ting, CY., Muppalaneni, N.B. (eds) Machine Learning and Internet of Things for Societal Issues. Advanced Technologies and Societal Change. Springer, Singapore.

[3] Küçük, Dilek, and Fazli Can. ”A tutorial on stance detection.” Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022.

[4] Mohammad, Saif M., Parinaz Sobhani, and Svetlana Kiritchenko. ”Stance and sentiment in tweets.” ACM Transactions on Internet Technology(TOIT) 17, no. 3 (2017): 26

[5] Küçük, Dilek, and Fazli Can. ”Stance Detection and Open Research Avenues.” arXiv preprint arXiv:2210.12383 (2022).

[6] Patel, Shaswat, Prince Bansal, and Preeti Kaur. ”Rumour detection using graph neural network and oversampling in benchmark Twitter dataset.” arXiv preprint arXiv:2212.10080 (2022).

[7] Alturayeif, Nora, Hamzah Luqman, and Moataz Ahmed. ”A systematic review of machine learning techniques for stance detection and its applications.” Neural Computing and Applications (2023): 1-32.

[8] He, Zihao, Negar Mokhberian, and Kristina Lerman. ”Infusing Knowledge from Wikipedia to Enhance Stance Detection.” arXiv preprint arXiv:2204.03839 (2022).

[9] Nababan, Arif Hamied, Rahmad Mahendra, and Indra Budi. "Twitter stance detection towards job creation bill." Procedia Computer Science 197 (2022): 76-81.
[10] Martínez, Rubén Yáñez, Guillermo Blanco, and Anália Lourenço. "Spanish Corpora of tweets about COVID-19 vaccination for automatic stance detection." Information Processing Management 60.3 (2023): 103294.
[11] Khandagale, Kanhaiyya, and Hetal Gandhi. "Sarcasm Detection in Hindi-English Code-Mixed Tweets Using Machine Learning Algorithms." Applied Computational Technologies: Proceedings of ICCET 2022. Singapore: Springer Nature Singapore, 2022. 221-229.
[12] Khandagale, Kanhaiyya, and Hetal Gandhi. "Sarcasm Detection in Hindi-English Code-Mixed Tweets Using Machine Learning Algorithms." Applied Computational Technologies: Proceedings of ICCET 2022. Singapore: Springer Nature Singapore, 2022. 221-229.
[13] Ng, Lynnette Hui Xian, and Kathleen M. Carley. "Is my stance the same as your stance? A cross validation study of stance detection datasets." Information Processing Management 59.6 (2022): 103070.