**A Project report on**

# TELUGU DATA CLASSIFICATION

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

## Bachelor of Technology

## in

## Computer Science and Engineering

Submitted by


R.NARASIMHA

(21H55A0519)

V.DURGA BHAVANI

(20H51A05D4)

V.KEERTHANA

(21H55A0524)


Under the esteemed guidance of

Ms. KOMAL PARASHAR
(Assistant Professor)



**Department of Computer Science and Engineering**

**CMR COLLEGE OF ENGINEERING & TECHNOLOGY**

(UGC Autonomous)

*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with $A^+$ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.


**2020- 2024**

# CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING


## CERTIFICATE

This is to certify that the Major Project Phase I report entitled **"Telugu Data Classification"** being submitted by R.NARASIMHA(21H55A0519),V.DURGA BHAVANI(20H51A05D4),V.KEERTHANA(21H55A0524) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

**Ms. Komal Parashar**
**Assistant Professor**
**Dept. of CSE**

**Dr. Siva Skandha Sanagala**
**Associate Professor and HOD**
**Dept. of CSE**

# ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Ms. Komal Parashar,** Assistant Professor Department of Computer Science and Engineering for her valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving force to complete my project work successfully.

We are very grateful to **Dr. Devadas**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their cooperation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

R. Narasimha (21H55A0519)
V. Durga Bhavani (20H51A05D4)
V.Keerthana (21H55A0524)

# TABLE OF CONTENTS


| CHAPTER NO. | TITLE | PAGE NO. |
|---|---|---|
| | LIST OF FIGURES | ii |
| | ABSTRACT | iii |
| **1** | **INTRODUCTION** | |
| | 1.1 Problem Statement | 1 |
| | 1.2 Research Objective | 2 |
| | 1.3 Project Scope and Limitations | 3 |
| **2** | **BACKGROUND WORK** | |
| | 2.1. Rule-Based Approach | 4 |
| | 2.1.1. Introduction | 4 |
| | 2.1.2. Merits, Demerits | 5 |
| | 2.1.3. Implementation | 5 |
| | 2.2. Supervised Learning | 6 |
| | 2.2.1. Introduction | 6 |
| | 2.2.2. Merits, Demerits | 7 |
| | 2.2.3. Implementation | 8 |
| | 2.3. Support Vector Machine(SVM) | 9 |
| | 2.3.1. Introduction | 9 |
| | 2.3.2. Implementation | 10 |
| **3** | **RESULTS AND DISCUSSION** | |
| | 3.1. Result | 11 |
| **4** | **CONCLUSION** | 12 |
| | **REFERENCES** | |

# LIST OF FIGURES


| FIGURE NO. | TITLE | PAGE NO. |
|---|---|---|
| 2.3.2 | shows some of the instances of Telugu news dataset | 9 |
| 2.3.2 | Shows Telugu news heading converted into English | 10 |

# ABSTRACT

Telugu is one of most difficult language which is morphologically rich Dravidian languages. There are many Telugu documents available on the Internet, it is important to organize the data by automatically by assigning a collection of text with predefined categories (business, science, sports, medical, entertainment...etc) based on their content using modern techniques. So, we will be doing a Telugu text data classification where the Telugu language data will be classified by Support Vector Machine (SVM) using Natural Language Processing (NLP) of Machine Learning. By using these techniques we will be generating the output as the category the information that is provided it belongs to. KEYWORDS: SVM, supervised learning, NLP.

# CHAPTER 1
# INTRODUCTION

# CHAPTER 1

# INTRODUCTION

## 1.1 Problem Statement

Natural Language Processing (NLP) is a crucial subfield of Artificial Intelligence that focuses on understanding and facilitating communication between computers and human languages. With the advent of online multimedia platforms, such as blogs, e-commerce review websites, social media sites like Facebook, Twitter, WhatsApp, Instagram, and LinkedIn, people from diverse linguistic backgrounds share their opinions and perspectives on various subjects, encompassing topics, products, events, companies, individuals, services, and properties. However, the majority of research and sentiment analysis has been conducted in English, with limited attention to Indian regional languages. Telugu, a widely spoken language in Andhra Pradesh and Telangana states of India, boasts approximately 93 million native speakers. The growing presence of Telugu content on websites, blogs, and social media platforms necessitates the analysis of sentiments expressed in Telugu-language news.

## 1.2 Research Objective

**Language-Specific Sentiment Analysis**: The primary objective of this research is to address the gap in sentiment analysis for Indian regional languages, with a specific focus on the Telugu language. The goal is to develop specialized sentiment analysis models and techniques that can accurately classify the sentiments expressed in Telugu news content.

**Translation and Preprocessing:** This research intends to leverage machine translation tools, such as Google Translator, to translate Telugu news data into English. Post-translation, data preprocessing techniques will be applied to clean and prepare the text for sentiment analysis.

**Sentiment Classification**: The research aims to classify the translated Telugu news content into various sentiment polarities, including positive, negative, and potentially neutral sentiments. Different machine learning classifiers, such as Naive Bayes, Random Forest, Passive Aggressive Classifier, Perceptron, and SVM, will be utilized for this purpose.

**Multi-Class Classification:** In addition to binary sentiment classification (positive or negative), this research seeks to perform multi-class sentiment classification, categorizing sentiments into distinct categories such as business, editorial, entertainment, nation, and sports. This approach allows for a deeper understanding of the nuances in sentiment expressed in Telugu news content.

**Performance Evaluation:** The study will evaluate the performance of the sentiment analysis models on test data using appropriate performance metrics, such as accuracy, precision, recall, and F1-score. The goal is to determine the effectiveness of the developed models in accurately classifying sentiments in Telugu news data.

## 1.3 Project Scope

The scope of this project encompasses the following key areas:

**Data Translation and Preprocessing:** The project includes the translation of Telugu news data into English using the Google Translator library in Python. This involves handling linguistic and structural nuances to ensure the accurate representation of the original content.

**Sentiment Analysis:** The project focuses on the application of various machine learning classifiers for sentiment analysis, including binary and multi-class classifications. It aims to determine the polarity of sentiments expressed in Telugu news.

**Performance Evaluation:** The project includes a rigorous assessment of the sentiment analysis models' performance using well-established evaluation metrics. This is essential for understanding the models' accuracy and effectiveness.

**Language Specificity:** The project acknowledges the language-specific nature of the research and its implications for understanding sentiment in the Telugu language. The research contributes to the field of NLP for Indian regional languages.

**Application to Real-World Scenarios:** The research explores the practical utility of sentiment analysis in the context of news content in Telugu, with potential applications in media, marketing, and beyond. and decision-making processes.

# CHAPTER-2

# BACKGROUND WORK

# CHAPTER 2

# BACKGROUND WORK

## EXISTING MODELS

### 2.1 Rule-Based Approach

#### 2.1.1 Introduction

Deepu S. Nair and their team proposed a rule-based approach for sentiment analysis of Malayalam movie reviews. In a rule-based approach, linguistic rules are predefined and applied to analyze text and classify sentiments as positive, negative, or neutral based on these rules.

#### 2.1.2 Merits and Demerits

**Merits:**

- S**implicity:** Rule-based approaches are straightforward to understand and implement. Linguistic rules are explicit, making it easy to interpret how the sentiment classification is performed.
- **Customization**: The rules can be customized for specific languages or domains. This adaptability allows the approach to be fine-tuned to handle nuances and idiosyncrasies specific to a particular language or context**.**

**Demerits :**

- **Limited Generalization**: Rule-based approaches may struggle with generalization. The predefined rules may not be suitable for different languages or complex linguistic variations. These approaches might perform well in one context but poorly in another**.**
- **Manual Effort:** Creating and maintaining linguistic rules is a manual and resource-intensive process. It requires language experts to design, update, and refine rules as languages evolve and usage patterns change. Moreover, handling negations, sarcasm, or ambiguous expressions in text can be challenging with rule-based systems.

### 2.1.3 Implementation

The rule-based approach involves the following steps:

- **Rule Creation:** Linguistic rules are manually created based on language-specific characteristics and patterns that are indicative of sentiment. For example, rules might specify that positive sentiment is associated with the presence of certain words or phrases.
- **Text Analysis:** The text to be analyzed, in this case, Malayalam movie reviews, is processed. The linguistic rules are applied to this text.
- **Sentiment Classification:** Based on the results of rule application, the system classifies the text as having positive, negative, or neutral sentiment**.**

## 2.2 Supervised Learning(Sahu et al.)

### 2.2.1 Introduction

Sahu et al. conducted a research study aimed at classifying Odia movie reviews using supervised classification techniques. The fundamental concept behind supervised learning is that it is a type of machine learning where an algorithm learns from labeled training data. These algorithms make predictions or decisions without human intervention, as they are capable of generalizing patterns from the training data. In the context of the study, Odia movie reviews were the subject of analysis.

### 2.2.2 Merits and Demerits

**Merits:**

- **High Accuracy:** Supervised learning models, when provided with sufficient high-quality labeled training data, can achieve high accuracy in classifying and predicting outcomes. This makes them suitable for sentiment analysis tasks, as correctly identifying positive and negative sentiments is crucial.
- **Generalization:** Supervised models can be trained on diverse datasets and can effectively generalize patterns from the training data to classify data from different languages or domains. This adaptability is a valuable feature when dealing with sentiment analysis in various contexts.

**Demerits:**

- **Data Dependency:** Supervised learning heavily relies on high-quality labeled training data. Acquiring and maintaining such data can be a challenging and time-consuming task. For sentiment analysis in Odia or any other language, obtaining a substantial volume of labeled data can be especially resource-intensive.
- **Computational Complexity:** The training and optimization of supervised learning models can be computationally complex and resource-intensive. Some algorithms, especially when dealing with large datasets, require significant computational resources, which can lead to longer training times and increased computational costs.

### 2.2.3 Implementation

In their study, Sahu et al. implemented supervised learning classifiers for sentiment analysis of Odia movie reviews. The classifiers they used included:

- **Logistic Regression:** Logistic regression is a statistical model that is commonly used for binary classification problems. It is employed to predict a binary outcome, in this case, classifying movie reviews as positive or negative sentiments. It models the probability of the binary outcome based on a set of independent variables.
- **Naive Bayes:** Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between the features. It is well-suited for text classification tasks like sentiment analysis. Naive Bayes models the conditional probability of each class given the input features and selects the class with the highest probability.
- **Support Vector Machine (SVM):** Support Vector Machines are powerful classifiers commonly used in binary and multiclass classification problems. SVM aims to find a hyperplane that best separates data points into different classes. In sentiment analysis, it seeks to find the hyperplane that separates positive and negative sentiments effectively.

## 2.3 Support Vector Machine(SVM)

### 2.3.1 Introduction

In this study, the authors address sentiment analysis in Indian languages, with a particular focus on Telugu. They use a dataset of Telugu news articles and aim to classify them into different categories such as Business, Editorial, Entertainment, Nation, and Sports. The primary goal is to build a machine learning-based classification system that can effectively categorize Telugu news.

### 2.3.2 Implementation

**Dataset Description:** News acts as a vital aspect in exhibiting reality and takes a strong influence on social practices and a lot of news in Telugu generated has less attention within the Sentiment Analysis community. These reasons motivated to select news dataset. The Telugu News dataset was collected from Kaggle (SRK, 2020). This dataset contains total of 17312 document sentence heading in Telugu and the news statements belongs to five interesting zones as are taken class labels: business, editorial, entertainment, nation and sports.

| Heading | Topic |
|---|---|
| యాక్సిస్, ఐడీబీఐ బ్యాంకులకు ఆర్‌బీఐ భారీ షాక్! | Business |
| చిన్న వ్యాపారాలను ఇంట్లో నుంచే నిర్వహించుకోవచ్చు. | Business |
| క్రికెట్ లో భారత్ కు స్వర్ణం | Sports |
| భారత ప్రజలకు విజయదశమి శుభాకాంక్షలు తెలిపిన ప్రధాని | Nation |
| గోవాలో అల్లు అర్జున్ తన కుమారుడితో కలిసి స్విమ్మింగ్ | Entertainment |

**Figure 1.** shows some of the instances of Telugu news dataset

**Translation:** The news statements in Telugu language translated into English by using GoogleTranslators API and removing the duplicate instances, unnecessary columns in the data frame which is needed to efficiently classifies this data.

| Heading | Tidy_Tweets |
|---|---|
| ఐడీబీఐ వైపు కన్నేసిన ఆర్బీఐ | RBI looks at IDBI |
| నేడు బ్యాంకింగ్ అధిపతులతో జైట్లీ భేటీ | Jailtey met with banking chiefs today |
| జడేజాకు కీలక వికెట్ | Jadeja takes key wicket |
| పాక్ మరో రెచ్చగొట్టే వ్యాఖ్యలు | Another provocative by Pakistan |
| కొడుకుతో కలిసి గోవాలో స్విమ్మింగ్ చేస్తున్న అల్లు అర్జున్ | Allu Arjun swimming with his son in Goa |

**Figure 2.** Shows Telugu news heading converted into English

**Pre-Processing**: Pre-processing is an essential stage to remove noise data and to increase consistency so that the cleaned data can be applied to text mining or opinion mining task efficiently (E.Haddi et al., 2013). Above represented translated data was cleansed by removing extra spaces, extra new lines, quotation marks, and other garbage values. Later Word Segmentation is done where this data was split into individual words.

**Feature Extractor:** Two feature extraction methods are employed, namely Count Vectorizing (One-Hot Encoding) and Tf-Idf Vectorizing. These methods convert words into numerical values suitable for machine learning algorithms.

**Classifiers:** The study utilizes several machine learning classifiers, including Multinomial Naïve Bayes, Random Forest, Passive Aggressive Classifier, Perceptron, and Support Vector Machines (SVM). Each classifier is trained on the pre-processed and feature-extracted data.

# CHAPTER 3

# RESULTS AND DISCUSSION

# CHAPTER 3
# RESULTS AND DISCUSSION

## 3.1 Result

The results of the study for Telugu news data classification using machine learning models are presented in Table. These results include accuracy, precision, recall, and F1-score for different classifiers applied to the dataset. The results vary across different classifiers and feature extraction methods, highlighting the importance of selecting the right combination for the best performance. When using the Tf-Idf Vectorizer, the Passive Aggressive Classifier achieved the highest accuracy of 80%.

| Parameters | Multinomial Naive Bayes | Random Forest | Passive Aggressive | Perceptron | SVM |
|---|---|---|---|---|---|
| **Accuracy** | 76 | 69 | 78 | 78 | 79 |
| **Precision** | 0.77 | 0.69 | 0.77 | 0.78 | 0.78 |
| **Recall** | 0.76 | 0.69 | 0.78 | 0.78 | 0.79 |
| **F1-Score** | 0.74 | 0.68 | 0.79 | 0.76 | 0.78 |

# CHAPTER 4
# CONCLUSION

# CHAPTER 4
# CONCLUSION

The study demonstrates the feasibility of sentiment analysis for Telugu news data using machine learning techniques. Different classifiers and feature extraction methods impact the performance of sentiment analysis models. The Passive Aggressive Classifier and SVM performed well, achieving high accuracy on the dataset. The choice of feature extraction method, Count Vectorizer or Tf-Idf Vectorizer, also had a significant impact on the results, with Tf-Idf Vectorizer outperforming Count Vectorizer in some cases. These findings can guide further research into sentiment analysis for Indian languages and news categorization. The choice of classifier and feature extraction method should be carefully considered for optimal results.

# REFERENCES

# REFERENCES


- Bonaccorso.G.(2017,October,6).Passive Aggressive Algorithms. https://www.bonaccorso.eu/2017/10/06/ml-algorithms-addendum-passive-aggressive-algorithms/
- Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with the perceptron algorithm. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '02).
- Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science, 17, 26–32.
- Jabeen Sultana, M. (2020). Deep Learning Based Recommender System using Sentiment Analysis to reform Indian Education. Learning and Analytics in Intelligent Systems, 15, 143–150.
- List of languages by total number of speakers. (2020). Encyclopedia. https://en.wikipedia.org/wiki/ List_of_languages_by_total_number_of_speakers
- Liu & Bing. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies.
- Mukku, S. S., Choudhary, N., & Mamidi, R. (2016). Enhanced Sentiment Classification of Telugu Text using ML Techniques. 25th International Joint Conference on Artificial Intelligence.
- Mukku, S. S., & Mamidi, R. (2017). Act as: Annotated corpus for telugu sentiment analysis. In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems (pp. 54–58). Association for Computational Linguistics.
- Naga Sudha, D., & Madhavee Latha, Y. (2021). Semi Supervised Multi Text Classifications for Telugu Documents. Turkish Journal of Computer and Mathematics Education, 12(12), 644–648.