"Telugu Word Sense Disambiguation: Building a Polysemy Dataset and Evaluating NLP Models"

Syam Immanuel Paul Bondada
230853737
Haim Dubbosarsky
MSc Big data Science

***Abstract*: One of the linguistically rich Dravidian languages is Telugu. Like other languages, it also has ambiguous words, depending on the situation, have unique meanings. These words are known as polysemous words, or words with several experiences. Telugu is not well-represented in linguistic resources. This study focusses on Word Sense Disambiguation (WSD), a core job in Natural Language Processing (NLP) that involves figuring out the meanings of words with multiple senses. The study highlights the challenges associated with WSD, particularly in contexts including multiple languages, and presents a new Telugu WSD dataset facilitating the development and assessment of contextualized models. One of the main contributions of this work is assessing the degree to which multilingual models handle polysemy in low-resource contexts by effectively transferring knowledge between languages. The result emphasizes how crucial language-specific modifications are to the advancement of NLP in Telugu, as well as how important resources like the recently released Telugu WiC dataset are. The knowledge gathered from this study is helpful in tackling the NLP issues that widely spoken but under resourced languages confront, and it clarifies the complexities associated with transfer learning in these situations.**

***Keywords:* Embeddings, Transformer Models, Word Sense Disambiguation (WSD), Clustering, Dimensionality Reduction.**

I. INTRODUCTION

Artificial Intelligence (AI) has made great progress in recent years, changing how industries work, boosting economies, and altering how we interact with technology. However, the benefits of these advancements are not shared equally. AI development has mostly focused on languages and regions that have plenty of resources, like English and other widely spoken languages. This creates a major challenge: making AI accessible and useful to everyone, no matter their language or available resources.

The field of Natural Language Processing (NLP) is a prime example of this challenge. NLP has achieved impressive results in areas like machine translation, sentiment analysis, and chatbots, but these successes depend heavily on large amounts of high-quality data, especially in English. On the other hand, languages with fewer resources, even those spoken by many people, lack the necessary linguistic data, such as annotated text collections. This lack of resources makes it difficult to develop NLP models for these languages, widening the gap between well-resourced and low-resource languages. Closing this gap is essential to ensuring that AI benefits everyone, promoting fairness and inclusivity in technological advancements. Though these developments are largely beneficial to languages with abundant resources, Natural Language Processing (NLP) has made great progress in recognizing and creating human language. Many low-resource languages, including indigenous and minority languages like Telugu, my native language, are left behind. Telugu faces challenges in developing strong NLP tools because there aren't enough annotated datasets, linguistic resources, or computational tools. This lack of resources leads to poorer performance of NLP models, making it difficult to create technologies like machine translation, speech recognition, and language understanding for Telugu and similar languages. The unique characteristics and diversity of these low-resource languages make the problem even harder, requiring new and creative solutions to overcome these challenges.

Low-resource languages in NLP have particular problems that need for a focused and methodical approach. The creation of high-quality datasets is essential to enable the training of models that can perform effectively in these languages. Unlike high-resource languages, where large-scale datasets are readily available, low-resource languages often lack such resources, leading to a significant performance gap. By developing and curating datasets specifically tailored to the linguistic characteristics of low-resource languages, it is possible to train models that approach the performance levels of those in rich-resource languages. This solution not only addresses the immediate need for better NLP tools in low-resource languages but also contributes to the broader goal of democratizing AI.

Polysemy, the phenomenon where a single word has multiple meanings, is a particularly challenging aspect of language understanding (Navigli, 2009). This is especially true in low-resource languages, where the nuances of polysemy are often underrepresented in the available datasets. Understanding and disambiguating polysemous words is crucial for tasks like translation, sentiment analysis, and information retrieval (Navigli, 2009). The focus of this work on polysemy stems from its significance in ensuring accurate language processing. In low-resource languages, where the richness of meaning is often lost due to limited data, addressing polysemy is vital for developing more sophisticated and context-aware NLP models.

This dissertation adds a great deal to the field of Natural Language Processing (NLP) by addressing the complex issue of polysemy in the Telugu language, a low-resource Dravidian language. The primary contribution is the creation of a novel Telugu Word Sense Disambiguation (WSD) dataset specifically designed to capture the nuanced meanings of polysemous words in Telugu. Additionally, this work evaluates

the effectiveness of various NLP models, including both language-specific and multilingual models, in handling polysemy within this dataset. By rigorously testing these models, the research highlights the limitations of general-purpose multilingual models and demonstrates the superior performance of language-specific models in accurately disambiguating word senses. This study not only provides valuable resources for advancing NLP in Telugu but also sets a foundation for improving NLP capabilities in other low-resource languages through similar approaches.

The thesis begins by reviewing the related work in the field of NLP for low-resource languages, particularly focusing on Telugu, to establish the context and foundation for this research. It then describes the various resources utilized in the study, which are essential for understanding the scope and limitations of the data and tools available for Telugu NLP. Following this, the thesis provides a detailed overview of the Telugu language, highlighting its unique linguistic features and the challenges these pose for NLP. The methodology section then outlines the process of creating the Telugu polysemy dataset and the approaches taken to evaluate different NLP models. The results and analysis section presents the findings from the experiments, offering insights into the effectiveness of the models tested. Finally, the thesis concludes with a discussion of the key findings, the limitations of the study, and potential directions for future research in this area.

II. Related Work

In exploring advancements in Natural Language Processing (NLP) for low-resource languages, particularly Telugu, several foundational studies have been instrumental in shaping my approach to handling polysemy and dataset creation. A pivotal contribution comes from Eluri and Siddu (2020), who proposed a novel algorithm for word sense disambiguation. Their method utilizes a combination of intersection, hierarchy, and distance metrics to compute semantic similarity between words. This approach was crucial for my research, as it highlighted the importance of integrating hierarchical metrics and intersection-based similarity to enhance the accuracy of sense annotation. Their work provided a theoretical foundation for developing a robust methodology for sense representation in my datasets, ensuring that polysemous words were effectively disambiguated based on contextual cues.

Further advancing the understanding of disambiguation techniques, Palanati and Kolikipogu (2021) explored the decision list algorithm for word sense disambiguation. Their study emphasized how corpus size and the clustering of documents with ambiguous words influence the accuracy of sense identification. Their findings were instrumental in guiding the creation of a more diverse and comprehensive dataset for my research. By incorporating their insights on corpus size and document clustering, I was able to enhance the quality and breadth of the datasets used in my study. Their work also informed my approach to evaluating clustering methods, highlighting the need for effective clustering techniques to improve the precision of word sense disambiguation.

Eluri and Pilli (2020) employed the ShotgunWSD approach, which combines unsupervised machine learning with knowledge-based methods to disambiguate Telugu words. This methodology, incorporating word embeddings and semantic relations, was particularly influential in shaping the dimensionality reduction and clustering methods used in my research. The ShotgunWSD approach demonstrated the efficacy of integrating machine learning with semantic analysis to improve disambiguation results. Additionally, Marreddy (2022) made significant contributions by developing extensive Telugu NLP resources, including annotated datasets and various embeddings for different tasks. Their work on sentiment, emotion, and hate-speech lexicons provided a valuable resource base that informed the creation of high-quality Telugu datasets and supported the evaluation of different NLP models.

A notable study by Dairkee and Dubossarsky (2024) further expands on the concept of polysemy in low-resource languages by introducing a new polysemy dataset in Hindi. Their research highlighted the challenges associated with cross-lingual transfer in word sense disambiguation tasks, particularly in the context of multilingual models. This study underscored the limitations of current models in effectively transferring knowledge across languages, thereby emphasizing the need for language-specific datasets and approaches. The insights from their work were critical in shaping my understanding of the cross-lingual challenges involved in word sense disambiguation and reinforced the importance of developing tailored datasets for each language. This was particularly relevant in the context of Telugu, where similar challenges are observed.

The methodologies and insights from these studies were integral to developing high-quality Telugu datasets and refining clustering techniques in my research. By adapting and building upon the approaches proposed by these researchers, I was able to create robust datasets and employ effective clustering methods tailored to Telugu polysemy. This integration of related work facilitated a comprehensive analysis of Telugu polysemy, enabling a more nuanced understanding of NLP techniques for low-resource languages. The advancements achieved through this integration contribute to the broader goal of enhancing NLP capabilities for languages with limited resources, paving the way for future research and development in this field.

III. Resources

Telugu NLP research benefits from a range of resources, despite the challenges posed by their limited availability. A foundational asset is the **AI4Bharat-IndicNLP Dataset** (AI4Bharat, 2020), which provides an extensive corpus of 2.7 billion words spanning 10 Indian languages, with notable representation for Telugu. This dataset is particularly valuable as it includes pre-trained word embeddings that support a wide array of NLP tasks, offering a robust basis for model training and evaluation. Another critical resource is the Telugu Wikipedia dump (Wikipedia, 2024), which provides a broad and diverse collection of textual data across various domains. However, it is important to note that this resource lacks spe-

cialized annotations, which can limit its utility for tasks requiring specific contextual information. For research focused on polysemy, the Telugu Sense Dataset is of great importance. It contains annotated instances crucial for evaluating word embeddings and understanding the nuances of word senses in context (Eluri, 2020). Additionally, the **Telugu-English Dictionary** by C. P. Brown (1903) remains an essential cross-lingual resource, facilitating the translation and interpretation of Telugu terms (Brown, 1903).

The development and evaluation of NLP models have been significantly supported by several advanced models. The "l3cube-pune/telugu-bert"(Huggingface.co.,2022) model, available through the Hugging Face Model Hub, was a key resource for generating contextual embeddings. This model, specifically fine-tuned for Telugu, is instrumental in capturing deep contextual nuances and improving the accuracy of tasks such as semantic analysis and clustering. Additionally, the ''Google/Muril-Base-Cased'' model has been utilized for its strong performance in multilingual settings, including Telugu, offering valuable capabilities for handling diverse linguistic contexts. For further enhancing contextual understanding, models such as ''Google/mt5-large'', "google/muril-large-cased" and "Google-BERT/bert-base-multilingual-cased" are noteworthy for their advanced multilingual features, although they were not directly used in this research.

Supporting resources that have played a role in this research include the **Telugu WordNet**, a comprehensive lexical database that provides semantic relationships among words. This resource is particularly useful for understanding word senses and their interconnections, even though it was not directly utilized in this study. Additionally, the **Telugu-English Dictionary**, accessible at Telugu Dictionary, was employed for cross-referencing and validating word senses and translations. Together, these resources and models have contributed significantly to the comprehensive analysis and robust development of Telugu language datasets and methodologies, enhancing the overall research outcomes in Telugu NLP.

IV. Telugu

Telugu is a Dravidian language mostly spoken in the Indian state of Andhra Pradesh and Telangana. With approximately 85 million native speakers, it ranks as the third most-spoken language in India and the 14th most-spoken language globally. Telugu has a rich cultural heritage, with its roots tracing back over 2,000 years. It has a classical status, which reflects its long literary tradition and historical significance. The language's development has been influenced by various dynasties, including the Chalukyas, the Kakatiyas, and the Vijayanagara Empire, each contributing to its literary and cultural evolution. Today, Telugu serves as a vital medium of communication, education, and media in its regions, with a robust presence in literature, cinema, and the arts.

One of the distinctive features of Telugu is its extensive polysemy-where a single word can have multiple meanings depending on context. This characteristic is a result of Telugu's rich morphological and syntactic structure. For instance, consider the word **"కోసం"** (kosam). The primary meaning of **"కోసం"** is "for" or "in order to", used to indicate purpose or benefit. However, it can also carry nuanced meanings depending on the context in which it is used:

1. **Purpose**: " మీరు యాపిల్స్ ఇష్టపడతారు, కాబట్టి మీ ఆరోగ్యం కోసం యాపిల్స్ తినాలి "
   - "You like apples, so you should eat more apples for health."
2. **Beneficiary**: "ఈ బతుకులు కేవలం నీ కోసం."
   - "These lives are only for you."
3. **Intention**: " నేను నీ కోసం ఎదురు చూస్తున్నాను."
   - "I am waiting for you."

Telugu, with its rich morphological and syntactic flexibility, exemplifies the depth of its polysemy, where a single word can convey multiple meanings based on contextual clues. This capacity for nuanced expression highlights the intricate nature of the language and makes it a compelling subject for linguistic studies, particularly in the realm of polysemy.

Adding to the language's complexity is its unique writing system. Telugu uses a syllabic script derived from the Brahmi script, characterized by its circular and curvilinear shapes. This phonemic script closely corresponds to the sounds of the language, which facilitates accurate pronunciation and reading. Featuring a set of consonants and vowels that combine to form syllables, along with distinctive diacritical marks to modify sounds, the Telugu script enhances the language's intricate phonological and grammatical structure. Together, these elements underscore the complexity and beauty of Telugu, reflecting its rich linguistic heritage.

V. Method

The approach taken to create the Telugu polysemy dataset, includes the collection of polysemous words, the sampling of their usages, and the subsequent annotation and validation processes. The methodology is divided into four key phases.

A. Collection of Candidate Polysemous Words

The initial step in constructing the dataset involved identifying and collecting polysemous words in Telugu. This process began with a thorough review of existing Telugu linguistic resources, including dictionaries, thesauri, and lexical databases such as the Telugu WordNet (Vishala., 2020). Leveraging these resources, a preliminary list of words known to have multiple meanings based on their usage in different contexts was compiled.To enhance the comprehensiveness of the candidate list, corpus-based methods were also employed. Large-scale Telugu corpora taken from the AI4Bharat-IndicNLP Dataset (AI4Bharat, 2020), Blogspot.com. (2021) -telugumalika and the Telugu Wikipedia dump (Wikipedia, 2024) were utilized to identify polysemous words that may not be extensively documented in traditional linguistic resources. Through natural language processing techniques, such as word frequency analysis and context clustering,

words that frequently appeared in varied contexts, indicating potential polysemy, were detected.

B. Sampling and Validation of Word Senses:

Once the candidate list of polysemous words was established, the next phase involved sampling sentences from the corpora that exemplified the different senses of these words. Initially, I created a dataset manually by reading the sentences for each word sense, collecting 20 sentences per sense. This manual process allowed for an in-depth understanding of the context and nuances of each word. After the manual dataset was established, I identified the best model, clustering method, and dimensionality reduction technique for automating and scaling this process.

For dimensionality reduction, I experimented with several techniques, including Principal Component Analysis (PCA)**,** t-distributed Stochastic Neighbor Embedding (t-SNE)**,** Uniform Manifold Approximation and Projection (UMAP)**,** and Multidimensional Scaling (MDS)**.** Each of these techniques was evaluated for its ability to effectively separate different word senses in a reduced-dimensional space. Among these, **UMAP** and **PCA** performed particularly well in preserving the local structure of the data, with UMAP providing the best balance between computational efficiency and the preservation of global structure, making it the preferred method for this study.

For clustering, I tested K-Means Clustering and Agglomerative Clustering**.** K-Means was evaluated for its simplicity and efficiency, while Agglomerative Clustering was considered for its ability to capture hierarchical relationships between word senses. After thorough testing, **K-Means Clustering** demonstrated superior performance, particularly in maintaining the integrity of closely related word senses, making it the chosen method for clustering in this study.

Context-sensitive word embedding models, specifically the "google/muril-base-cased" and "l3cube-pune/telugu-bert" models (Hugging Face, 2024), were used to generate contextual representations of the polysemous words across different sentences. By applying UMAP for dimensionality reduction and Agglomerative Clustering to these contextual embeddings, I identified distinct sense groups corresponding to different meanings of each word. Representative sentences that captured the diversity of meanings were sampled from these clusters. To ensure the accuracy of sense annotation, each sampled sentence was manually reviewed by linguists proficient in Telugu. (Eluri, 2020) These experts verified the sense of the polysemous word in each context, ensuring that the clustering process had correctly identified and separated the different meanings. In cases where the clustering was ambiguous or overlapping, the sentences were re-evaluated and re-clustered based on expert feedback.

C. Summary Statistics

After completing the collection, validation, and pairing processes, summary statistics were compiled to quantify the scope and scale of the dataset. The final Telugu polysemy dataset comprises:

A total of 50 polysemous words were identified and validated for this study. Each word, on average, exhibited 2 to 3 distinct senses, leading to a dataset encompassing approximately 100 unique word senses. The dataset comprises 4581 sentences, with each word sense being represented by around 50 sentences.

Total Words	50
Total Senses	102
Total Number of sentences	4581
Average number of senses/word	2.04
Average number of sentences/word	44.9

Table 1:Overview of Dataset Statistics

This methodical approach to dataset creation ensures that the Telugu polysemy dataset is both comprehensive and reliable, providing a valuable resource for advancing NLP research in low-resource languages. The balance between automatic processes and manual validation has resulted in a high-quality dataset that effectively captures the complexity of polysemy in Telugu.

VI. Results & Analysis

In this study, several linguistic resources were leveraged to identify and validate polysemous words in Telugu, which served as the foundation for the dataset creation and subsequent analyses. These resources included traditional Telugu dictionaries, annotated linguistic corpora, and digital lexical databases such as the Telugu WordNet. Each resource played a critical role in distinguishing between the different senses of polysemous words, which is essential for accurate word sense disambiguation (WSD) in NLP tasks.

Sample Entry from Dictionary:

Word: "**వేరు"** (veru)

- Sense 1: Different – Used to describe something that is distinct or not the same as something else.
- Sense 2: Plant Root – Refers to the part of a plant that typically lies below the surface of the soil and anchors the plant, absorbing water and nutrients.

This entry illustrates how a single Telugu word can have distinct meanings based on the context in which it is used. The ability to correctly identify these senses is crucial for tasks like machine translation, sentiment analysis, and information retrieval.

The full list of 50 polysemous words, along with their identified senses, is provided in the appendix of this thesis. This list was systematically compiled by cross-referencing entries from multiple sources to ensure that each word included was genuinely polysemous and represented distinct, context-dependent meanings. The candidate list was then used as a basis for extracting and validating sentences that illustrate the various senses, ensuring a comprehensive and reliable dataset for further analysis.

For a more in-depth exploration, five polysemous words were selected for detailed manual annotation. This process involved determining the correct sense of each word in a set

of sampled sentences, a critical step for training and evaluating the NLP models used in this study. The manual annotation was particularly important for selecting the most effective model and dimensionality reduction technique for accurately projecting word senses in semantic space.

Word: నవ(nava)

- Sense 1: Nine
  "అమ్మ నవ మాసాలు బిడ్డను మోసి, ఎన్నో కష్టాలు, నొప్పులను కూడా లెక్క చేయకుండా తన బిడ్డకు ప్రాణం పోస్తుంది."
  "A mother carries a child for nine months and gives life to her child without even counting the many hardships and pains."

- Sense 2: New
  "నవ వధువు అస్వస్థకు గురికావడంతో స్థానికంగా ఆస్పత్రికి తరలించారు."
  "The new bride fell ill and was taken to a local hospital."

Word: కూలి (kooli)

- Sense 1: Labour
  "కూలి పని చేసుకొని కుటుంబాన్ని పోషిస్తున్నారు."
  "He is working as a labourer and supporting his family."

- Sense 2: Destroyed/Collapsed
  "స్కూల్ గోడ కూలి ఇద్దరు విద్యార్థినులు మృతిచెందారు."
  "Two female students died after the wall of the school collapsed."

Word: గడ్డ (gadda)

- Sense 1:Land/Country
  "పవన్ కళ్యాణ్ పవర్‌కి అమెరికా గడ్డ దద్దరిల్లింది."
  "Pawan Kalyan's power shakes the American soil."

- Sense 2: Clot/Clump
  "దీంతోపాటు, మెదడులో రక్తం గడ్డ కట్టింది."
  "In addition, there is a blood clot in the brain."

Word: గాజు (gaaju)

- Sense 1: Glass
  "గాజు గ్లాసులు, కప్పుల్లో మాత్రమే తాగాలి."
  "Drink only in glass glasses and cups."

- Sense 2: Bangle
  "చెల్లికి కొత్త గాజు తెచ్చాను."
  "I brought a new bangle for my sister."

Word: తాళం (talam)

- Sense 1: Lock
  "వాళ్ళ ఇల్లు కుడా తాళం వేసి ఉంది."
  "Their house is also locked."

- Sense 2: Rhythm
  "ఆ గాయకుడు తన పాటల్లో తాళం ని అద్భుతంగా వినియోగించడం ద్వారా ప్రసిద్ధి పొందాడు."
  "The singer is known for his brilliant use of rhythm in his songs."

The consistently high accuracy and F1 scores observed in the evaluation could suggest that Telugu, in the context of polysemy, presents relatively clear distinctions in word meanings, making it easier for models to disambiguate. Additionally, the specific examples chosen for this study may have been particularly straightforward, further contributing to these results. The fine-tuning of models on Telugu-specific data and the effective use of clustering techniques likely enhanced the models' ability to accurately capture and differentiate between the nuanced meanings in the dataset.

To evaluate the effectiveness of different NLP models, scatter plots were generated for each word. These plots visually represent how sentences with different senses are separated in a lower-dimensional space, which is crucial for understanding the model's ability to disambiguate polysemous words.

A. Model 1: BERT (Telugu-specific):

The first model evaluated was a BERT model specifically fine-tuned for the Telugu language. This model demonstrated an impressive performance with an average accuracy of 90%

Word	Model	Clustering Method	Accuracy	F1 Score
నవ	google/muril-base-cased	K-Means	0.99	0.99
	l3cube-pune/telugu-bert		1.00	1.00
కూలి	google/muril-base-cased	K-Means	0.95	0.95
	l3cube-pune/telugu-bert		0.94	0.94
గడ్డ	google/muril-base-cased	K-Means	0.99	0.99
	l3cube-pune/telugu-bert		0.99	0.99
గాజు	google/muril-base-cased	K-Means	1.00	1.00
	l3cube-pune/telugu-bert		1.00	1.00
తాళం	google/muril-base-cased	-Means	1.00	1.00
	l3cube-pune/telugu-bert		1.00	1.00

Table 2: Comparison of Accuracy and F1 Scores for "google/muril-base-cased" and "l3cube-pune/telugu-bert" on Selected Polysemous Words in Telugu

and an F1-Score of 0.89. BERT's strength lies in its deep understanding of context, which was particularly evident in the scatter plot analysis for the word "కూలి" (kooli). In Telugu, "కూలి"(kooli) can mean both "laborer" and "collapsed", depending on the context. The BERT model was able to clearly distinguish between these different senses, as indicated by a distinct separation of data points in the scatter plot. This clear differentiation underscores BERT's capacity to effectively manage polysemy in Telugu, making it a powerful tool for tasks that require nuanced understanding and precise contextualization in the language. The model's performance highlights the value of language-specific fine-tuning, especially in handling the complexities of Telugu, where words often carry multiple meanings based on subtle shifts in context.

B. Model 2: MURIL (Multilingual):

The second model evaluated was MURIL, a multilingual model specifically designed with a focus on Indian languages, including Telugu. MURIL achieved a commendable average accuracy of 91% and an F1-Score of 0.91, closely aligning with the performance of the Telugu-specific BERT model. In many cases, MURIL demonstrated strong capabilities, particularly in handling multilingual contexts and diverse linguistic environments. For instance, in scenarios where the word "కూలి" (kooli) had overlapping meanings such as "laborer" and "collapsed" MURIL was able to maintain a level of precision comparable to that of the Telugu BERT model. However, while it generally performed well, there were instances where MURIL slightly outperformed BERT, particularly in handling nuanced distinctions in less colloquial contexts. This suggests that while MURIL's multilingual design makes it versatile across various languages, it can still approach, and in some cases exceed, the performance of language-specific models like BERT in certain Telugu-specific tasks.

C. Model 3: XL-Lexeme (Multilingual):

The third model assessed was 'pierluigic/xl-lexeme', another model designed for multilingual tasks. This model recorded an average accuracy of 77% and an F1-Score of 0.76. However, when it came to distinguishing between different meanings of words in Telugu, its performance was somewhat less sharp compared to the BERT model.

Model	*Average Accuracy*	*Average F1-Score*
pierluigic/xl-lexeme	0.77	0.76
google/muril-base-cased	0.91	0.91
l3cube-pune/telugu-bert	0.90	0.89
google/muril-large-cased	0.70	0.70

Table 3: Comparison of Average Accuracy and F1-Score for Different NLP Models on Telugu Word Sense Disambiguation Tasks.

Evaluation of Other Models

In addition to BERT, MURIL, and XL-Lexeme, several other models were evaluated for their performance in handling Telugu language tasks. Several multilingual models were evaluated for their performance in handling Telugu language tasks, but many struggled with the nuances and complexities of the language. The 'google/mt5-large' model, though designed for a variety of tasks across different languages, faced challenges in accurately capturing the polysemy in Telugu. Similarly, 'google-bert/bert-base-multilingual-uncased' and 'google-bert/bert-base-multilingual-cased' versions of BERT, despite their multilingual design, did not perform as well in Telugu-specific contexts, highlighting the limitations of a one-size-fits-all approach for languages with distinct linguistic features. Even **'google/muril-large-cased'**, part of the MURIL family known for focusing on Indian languages, did not deliver the desired results, illustrating the difficulty in achieving high performance across multiple languages with a single model. Lastly, 'distilbert/distilbert-base-multilingual-cased', a distilled version of BERT designed to be lighter and faster, underperformed in Telugu-specific tasks, emphasizing the trade-off between model size and effectiveness in dealing with nuanced language tasks.

The evaluation of these models highlights the importance of language-specific fine-tuning and the limitations of applying general-purpose multilingual models to tasks in languages like Telugu. While multilingual models offer the advantage of broad applicability, they often fall short in handling the specific challenges posed by individual languages, particularly when it comes to polysemy and deep contextual understanding. For languages such as Telugu, where words can carry multiple meanings depending on subtle contextual cues, models like BERT, which are fine-tuned for the language, offer superior performance. This suggests that for tasks requiring precise language understanding, particularly in languages with rich polysemy, language-specific models may be the best approach.

The sentence sampling process was a critical step in constructing a robust and diverse dataset, ensuring that it captures the full range of meanings for words with multiple senses. This process involved clustering the sentences based on their contextual usage and calculating the centroid for each sense. The sentences closest to these centroids were selected, representing the most typical instances of each sense. These selected sentences were then manually reviewed to ensure accuracy and appropriateness.

The screenshot provided illustrates the process of smart sentence sampling used in this study for Word Sense Disambiguation (WSD) in Telugu. In this example, the word "కూలి" (kooli) is analysed, which has multiple meanings, such as "labour" and "destroyed/collapsed." To effectively distinguish between these senses, we first embedded the sentences containing "కూలి" using the "telugu-bert" model. These high-dimensional embeddings were then reduced to a 2D space using Uniform Manifold Approximation and Projection (UMAP), a technique that preserves the local structure of the data. The reduced embeddings were subsequently clustered using K-Means, which grouped the sentences into distinct clusters

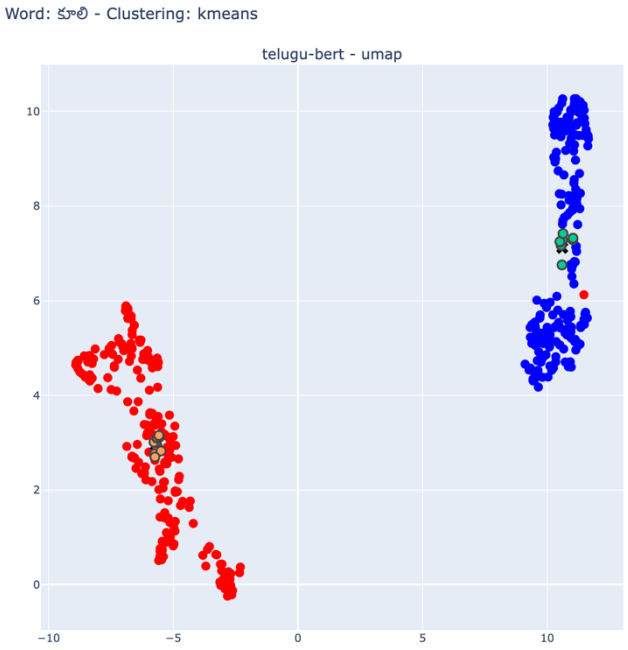

Figure 1: *Visualization of the word "కూలి" using UMAP and K-Means clustering with the* telugu-bert *model.*

based on their contextual similarities. Each cluster, represented by different colours in the screenshot, corresponds to a different sense of the word "కూలి."

Following the clustering, the sentences closest to each cluster's centroid were identified as representative examples of the respective word senses. These centroid sentences are crucial as they most accurately capture the context and meaning of the word within that cluster. To ensure the reliability of this automatic process, the selected sentences were manually reviewed by linguists proficient in Telugu. This manual validation step confirmed that the sentences accurately reflected the intended word sense and that the clustering had correctly differentiated between the different meanings. The combination of advanced clustering techniques and meticulous manual validation ensured that the resulting dataset is both comprehensive and reliable, effectively capturing the nuances of polysemy in Telugu.

VII. Discussion

The research presented in this thesis highlights significant advancements in the field of Natural Language Processing (NLP) for low-resource languages, specifically focusing on Telugu. Through the development of a robust Telugu polysemy dataset, this study addressed the critical challenge of Word Sense Disambiguation (WSD), which is a fundamental task in NLP(Koppula.N, 2019). The discussion draws upon the key findings from the dataset creation, the evaluation of various NLP models, and the methodological innovations introduced in this research.

One of the primary challenges in NLP for low-resource languages like Telugu is the scarcity of annotated linguistic resources. The lack of comprehensive datasets and advanced models tailored to these languages has led to a significant gap in NLP performance when compared to well-resourced languages. This research aimed to bridge this gap by focusing on polysemy, a particularly challenging aspect of language understanding. Polysemous words, which have multiple meanings depending on the context, are prevalent in Telugu. (Eluri & Siddu, 2020) Accurately disambiguating these words is crucial for tasks such as machine translation, sentiment analysis, and information retrieval.

The creation of the Telugu polysemy dataset involved a meticulous process of identifying, sampling, and annotating polysemous words. The methodology incorporated both automated and manual approaches, ensuring that the dataset captured a wide range of contextual usages for each word. By leveraging clustering techniques and centroid calculations, the most representative sentences for each word sense were identified. (Palanati & Kolikipogu, 2021) This approach not only enhanced the dataset's comprehensiveness but also ensured that it was well-suited for training sophisticated NLP models capable of handling Telugu polysemy.

A significant portion of the research involved evaluating various multilingual models to assess their ability to disambiguate polysemous words in Telugu. The findings revealed that while multilingual models offer broad applicability, they often struggle with the nuances and complexities of individual languages. For example, models such as 'google/mt5-large' and 'FacebookAI/xlm-roberta-large' faced challenges in accurately capturing the polysemy in Telugu. These models, although effective in multilingual contexts, did not perform as well when dealing with the specific linguistic features of Telugu. This suggests that a one-size-fits-all approach in multilingual models may not be sufficient for low-resource languages with rich morphological and syntactic structures (Palanati & Kolikipogu, 2021; Eluri & Pilli, 2020).

In contrast, language-specific models like **BERT**, fine-tuned for Telugu, demonstrated superior performance in WSD tasks. The BERT model's ability to deeply understand and contextualize polysemous words was evident in its scatter plot analysis, where it successfully separated different senses of words like "కూలి" (kooli). This underscores the importance of fine-tuning models for specific languages to achieve high accuracy in NLP tasks. The superior performance of the BERT model highlights the need for continued development of language-specific models, particularly for low-resource languages like Telugu (Eluri & Pilli, 2020).

IX. Future Work & Limitations

This research has made significant progress in addressing Word Sense Disambiguation (WSD) in Telugu, but there are several areas where future work could enhance these findings. One important direction is the expansion of the Telugu polysemy dataset to include a broader range of polysemous words and more diverse sentence contexts. This expansion could improve the dataset's utility and help NLP models generalize better across various linguistic scenarios. Additionally, exploring cross-lingual transfer learning, particularly between Telugu and other Dravidian languages, could leverage shared linguistic features to improve WSD performance across multiple low-resource languages. Automated approaches to manual annotation, such as integrating active

learning techniques, could also make the dataset creation process more scalable and reduce the dependency on extensive manual labour.

However, this research has some limitations that should be acknowledged. The current dataset, while comprehensive, represents only a fraction of the complexity found in the Telugu language, which may limit the generalizability of the models trained on it. The manual annotation process, though crucial for accuracy, is resource-intensive and potentially introduces human bias, which could affect the consistency of the dataset. Additionally, the performance of multilingual models in this study highlighted their struggles with the nuances of Telugu polysemy, suggesting that the trade-offs between language-specific fine-tuning and the broad applicability of multilingual models remain unresolved. The clustering techniques used for sense separation, while effective, may not fully capture the intricate contextual factors that influence word meaning, indicating a need for more sophisticated methods.

Looking ahead, there is considerable potential for improving the effectiveness of NLP models in low-resource languages like Telugu. Future research should focus on enhancing the dataset with more diverse linguistic data, refining clustering techniques, and exploring the potential of cross-lingual transfer learning. Additionally, practical applications of these models, such as in machine translation and sentiment analysis, should be explored to evaluate their real-world effectiveness. These efforts will be essential for advancing NLP capabilities and ensuring that AI technologies are inclusive and accessible to speakers of all languages.

X. Conclusion

This thesis has contributed significantly to the field of Natural Language Processing (NLP) by addressing the challenge of polysemy in the Telugu language. Through the development of a robust Telugu polysemy dataset and the evaluation of various NLP models, the research has demonstrated the limitations of general-purpose multilingual models and the superiority of language-specific models like BERT for tasks involving Word Sense Disambiguation (WSD). The methodological innovations introduced, including smart sentence sampling and clustering techniques, have proven effective in creating a comprehensive and balanced dataset, which is crucial for advancing WSD in Telugu.

The findings underscore the importance of continued development in language-specific models and resources for low-resource languages. By enhancing NLP capabilities for Telugu, this research lays a strong foundation for future efforts aimed at closing the gap between high-resource and low-resource languages. Ultimately, this work contributes to the broader goal of democratizing AI, ensuring that technological advancements are accessible and beneficial to speakers of all languages, regardless of their resource availability.

Acknowledgements

Dr. Haim Dubossarsky's encouragement and constructive feedback greatly contributed to the development of the Telugu polysemy dataset and the evaluation of the NLP models, ultimately helping to shape the overall direction and impact of this research. His mentorship has been essential in navigating the challenges of this project, and his contributions are deeply appreciated.

References

AI4Bharat (2020). *GitHub - AI4Bharat/indicnlp_corpus: Description Describes the IndicNLP corpus and associated datasets*. [online] GitHub. Available at: https://github.com/AI4Bharat/indicnlp_corpus?tab=readme-ov-file#ai4bharat-indicnlp-dataset [Accessed 20 Aug. 2024].

Blogspot.com. (2021). నానార్థాలు. [online] Available at: https://telugumalika.blogspot.com/2012/05/blog-post_13.Html [Accessed 20 Aug. 2024].

Brown, C.P. and Library, A. (2023). A Telugu-English Dictionary. *Uchicago.edu*. [online] doi:http://www.purl.oclc.org/dsal/dictionaries/PL4776.B680.

Dairkee, F. and Dubossarsky, H. (1534). *Strengthening the WiC: New polysemy dataset in Hindi and lack of cross lingual transfer*. [online] p.15341. Available at: https://aclanthology.org/2024.lrec-main.1332.pdf.

Eluri, S. and Pilli, V.K. (2020). Global Word Sense Disambiguation of Polysemous Words in Telugu Language. *International Journal of Engineering and Advanced Technology*, 10(1), pp.420–425. doi:https://doi.org/10.35940/ijeat.a1915.1010120.

Huggingface.co. (2022). *l3cube-pune/telugu-bert · Hugging Face*. [online] Available at: https://huggingface.co/l3cube-pune/telugu-bert.

Huggingface.co. (2024). *Models - Hugging Face*. [online] Available at: https://huggingface.co/models/ [Accessed 20 Aug.2024].

Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi:https://doi.org/10.18653/v1/2020.acl-main.560.

Marreddy, M. (2023). *Text Classification for Telugu: Datasets, Embeddings and Models for Downstream NLP Tasks*. [online] Google.com. Available at: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=Ikqyo5sAAAAJ&citation_for_view=Ikqyo5sAAAAJ:fQNAKQ3IYiAC [Accessed 21 Aug. 2024].

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni and Radhika Mamidi (2022). Multi-Task Text Classification using Graph Convolutional Networks for Large-Scale Low Resource Language.

2022 International Joint Conference on Neural Networks (IJCNN). doi:https://doi.org/10.1109/ijcnn55064.2022.9892105.

Navigli, R. (2009). Word sense disambiguation. *ACM Computing Surveys*, 41(2), pp.1–69. doi:https://doi.org/10.1145/1459352.1459355.

Palanati, D. and Kolikipogu, R. (n.d.). Decision List Algorithm for Word Sense Disambiguation for TELUGU Natural Language Processing. *International Journal of Electronics Communication and Computer Engineering*, [online] 4(6). Available at: https://www.ijecce.com/Download/conference/NCRTCST-2/38NCRTCST-13084.pdf [Accessed 22 Aug. 2024].

Suneetha Eluri and Vishala Siddu (2020). A Knowledge Based Word Sense Disambiguation in Telugu Language. *International Journal of Engineering and Advanced Technology*, 10(1), pp.440–445. doi:https://doi.org/10.35940/ijeat.a1911.1010120.

Koppula, N., Rani, B.P. and Srinivas Rao, K. (2019). Graph-based word sense disambiguation in Telugu language. International Journal of Knowledge-based and Intelligent Engineering Systems, 23(1), pp.55–60. doi:https://doi.org/10.3233/kes-190399.

APPENDEX

Table 1: Comprehensive List of Polysemous Words with Translations, Definitions, Part of Speech, and Example Sentences

Word	Word Trans-literation	Sentence	Sense	Meaning Word	Part of Speech	Sentence Translation
అంటు	aMTu	ఆకాశాన్ని చూస్తూ ఈ విశ్వం అనంతం అనే సత్యాన్ని తలచుకుంటూ అసలు అదెలా సాధ్యం అంటు తెగ ఆశ్చర్య పడేవాడిని.	0	Saying	Noun	Seeing the sky, They are surprised by saying that this universe is the truth of infinite
అంటు	aMTu	పుష్కరాల వల్ల పాప నాశనం కాదు చర్మ వ్యాధులు, అంటు రోగాలు ప్రభలుతాయి.	1	Touch/ Infectious	Noun	Skin diseases and infectious diseases are not destroyed by flowers.
అడుగు	aDugu	మీ అభిప్రాయం గురించి అడుగు.	0	Ask	Verb	Ask about your opinion.
అడుగు	aDugu	తెలంగాణ మార్పులో ఒక అడుగు ముందే ఉంటుంది.	1	Step	Noun	Telangana change is one step in advance.
అంతస్తు	aMtastu	5వ అంతస్తు నుంచి కిందకు దూకి ఆత్మహత్యాయత్నం చేశాడు.	0	Storey/ floor	Noun	He jumped from the 5th floor and attempted suicide.
అంతస్తు	aMtastu	రాజకీయ నాయకులంతా తమ అంతస్తు కోసం పోటీ పడతారు.	1	Status	Noun	All politicians compete for their status.
అసలు	asalu	మనోడు మాత్రం అసలు ఒక క్యూట్ లవ్ స్టోరీతో కామెడీగా చంపేశాడంతే.	0	Really/Actually	Adverb	Actually, our guy killed a comedy with a cute love story.
అసలు	asalu	అతను తన అసలు సొమ్మును కంపెనీలో పెట్టుబడి చేసాడు.	1	Principal Amount/Original Amount	Noun	He invested his original money in the company.
ఆట	ATa	నిన్న ఆ ఆట చూడడం నాకు సంతోషం కలిగించింది.	0	Dance/show	Noun	It made me happy to watch that show yesterday.
ఆట	ATa	2007 ఆగ్నేయ ఆసియా క్రీడలలో పోలో ఆట చేర్చబడింది.	1	Sport	Noun	The Polo game was included in the 2007 Southeast Asian Games.
ఆత్రుత	Atruta	విమానం ఎక్కేముందు మితిమీరిన ఆత్రుత పనికిరాదు .	0	Anxious	Adjective	Excessive anxiety is useless before boarding the plane.
ఆత్రుత	Atruta	ఈ కేవలో సస్పెన్స్, ముందేం జరుగుతుందన్న ఆత్రుత కరువై పోయాయి.	1	Eagerness	Noun	In this category, the suspense and the eagerness of pre -happening.
ఆసు	Asu	ఆసు యొక్క కర్రలు బలంగా ఉండాలి.	0	Looming Eqipment	Noun	The looming equipment of sticks should be strong.
ఆసు	Asu	ఆసు అతనికి విజయాన్ని ఇచ్చింది.	1	Ace (in cards)	Noun	Ace gave him success.
ఉత్తరం	uttaraM	తెరిచి చూస్తే అందులో కొన్ని ఎండు ఖర్జూరాలు, ఒక ఉత్తరం వున్నాయి.	0	Letter	Noun	Opening it has some dried kharjuras and a letter.
ఉత్తరం	uttaraM	ఇది ఉత్తరం వైపు తెలంగాణా రాష్ట్రానికి చివరి జిల్లా.	1	North	Noun	This is the last district of Telangana state on the north.
ఎద	èda	తన ఎద అందాలపై ఎలాంటి ఆచ్ఛాదన లేకుండా కేవలం చేతులని అడ్డుపెట్టుని వీడియో చేసింది.	0	Chest	Noun	The video was made to interfere with his hands without any impact on her chest beauty.
ఎద	èda	ఫలితంగా పవన్ కల్యాణ్ పచ్చబొట్టును తన ఎద పక్కన పూనమ్ కౌర్ పొడిపించుకుంటున్నారు.	1	Heart	Noun	As a result, Poonam Kaur is getting Pawan Kalyan's tattoo next to her heart.
కట్ట	kaTTa	అయితే, మనుగడకోసం ఎంతో కొంత రాజీ పడే ధోరణి కాస్తా అడ్డూ అదుపూ	0	Bundle	Noun	However, the tendency to survive has increased in the

		లేకుండా పెరిగిపోయి ఇతరేతర రంగాల్లో వుండే అవలక్షణాలన్నీ కట్ట గట్టుకుని ఈ రంగంలో ప్రవేశించాయి.				field, which has become a bit of a compromise.
కట్ట	kaTTa	కృష్ణా నది కట్ట పైన పర్యాటకుల సంఖ్య పెరుగుతోంది, దీంతో స్థానిక వ్యాపారాలు పుంజుకున్నాయి.	1	bank/shore/dam/em-bankment	Noun	The number of tourists on the Krishna River bank is increasing, leading to local businesses.
కమ్మ	kamma	రాష్ట్రంలో అత్యధికంగా కమ్మ వారు ఉండే విజయవాడలో కచ్చితంగా జగన్ కథ అంతే అనుకున్నారంతా.	1	Regional Community	Noun	In Vijayawada, where there is a majority of Kamma community people in the state, Jagan's story is exactly what everyone thought.
కమ్మ	kamma	ఆ వంటకారుని ప్రసిద్ధి కమ్మ రుచికి సంబంధించినది.	2	Subtle Taste	Adjective	The chef is famous for his subtle/good taste.
కమ్మ	kamma	ప్రముఖ డిజైనర్ రూపొందించిన కమ్మ లతో మోడల్స్ రాంప్ పై మెరిసిపోయారు.	0	Earring	Noun	Models dazzled on the ramp in earrings created by the famous designer.
కళ	kaLa	కళ రంగం నేటి సమాజానికి మార్గదర్శకంగా ఉన్నది.	0	Art/Skill	Noun	The art sector is a guide to today's society.
కళ	kaLa	వజ్రాలు తమ కళ తో ప్రదర్శనలో మెరిసిపోయాయి.	1	Ray of light/bright-ness/Radiance	Noun	The diamonds shone in the show with their brightness.
కాంతి	kAMti	65 లక్షల కాంతి సంవత్సరాల దూరంలో ఉంది.	0	Light	Noun	65 million light years away.
కాంతి	kAMti	పచ్చని చీరలో ఆమె కాంతి ఎంతో విశాలంగా ఉంది.	1	brilliancy/ lustre/ gleam	Noun	Her lustre in the green saree is very broad.
కాపు	kApu	కాపు రిజర్వేషన్లపై ఓసారి కావాలని, ఇప్పుడు కుదరదని జగన్ మాట మారుస్తున్నారన్నారు.	1	Regional Commuity	Noun	Jagan's word is changing not to be once again on the Kapu community reservation.
కాపు	kApu	పిల్లలకు కాపు గా ఉన్నప్పుడు తల్లిదండ్రులు సంతోషిస్తారు, ఎందుకంటే వాళ్లకు భద్రత కల్పించారు.	0	Protection	Noun	Parents are happy when they are protectors for children, as they are secured.
కారు	kAru	అతడిని నమ్మి కారు ఎక్కాను.	0	Car	Noun	I boarded a car believing him.
కారు	kAru	కారు మేఘాలు కమ్ముకోవడం వలన వాతావరణం చల్లబడింది.	1	Black Colour	Noun	The weather cooled due to black clouds.
కూలి	kUli	ఆ తర్వాత రైతు కూలి సంఘం అమరులైన గిరిజనుల స్మారక స్థూపాన్ని నిర్మించింది.	0	Labour Charges or Labour	Noun	The farmer labour union then built a commemorative stupa of the tribal community.
కూలి	kUli	మంగళవారం భండారా జిల్లాలో భారీ వర్షానికి ఇల్లు కూలి ఒకే కుటుంబానికి చెందిన ముగ్గురు మృతి చెందారు.	1	Destroyed	Verb	Three members of the same family died after their house collapsed due to heavy rain in Bhandara district on Tuesday.
కొట్టు	kòTTu	నిన్న రాత్రి కొట్టు యజమానిని అప్రమత్తం చేయడం ద్వారా దొంగతనాన్ని నివారించారు.	0	Shop	Noun	Last night the theft was prevented by alerting the shop owner.
కొట్టు	kòTTu	నిందితుడి కొట్టు లో భద్రతా సిబ్బంది కఠిన నిఘా ఏర్పాటు చేశారు.	1	Prison Cell	Noun	Security personnel have been rigorous surveillance in the accused cell.
కొమ్ము	kòmmu	ఈ కొమ్ము ద్వారానే ప్రవర్థనం చెందుతుంది.	0	Horn	Noun	This horn can be translated itself.

కొమ్ము	kòmmu	ఈ కథలు రాసేవాళ్ళూ, వాళ్ళ కొమ్ము కాసేవాళ్ళూ నింగినీ నేలనీ నిలవకండ అదిరిపడుతున్నారు.	1	Support	Noun	The people who write these stories and those who support them are pushing themselves to the ground.
క్షేత్రము	kShetramu	ఈ క్షేత్రము కొవ్వూరుకు 25 కి.మీ.	0	A sacred spot	Adjective	This sacred spot is 25 km away.
క్షేత్రము	kShetramu	ఆత్మకు తన ఒక్క శరీర క్షేత్రము గురించే తెలుసు.	1	Field	Noun	The soul knows about his one body field.
గజం	gajaM	గజం తన గొప్ప శరీరంతో చెట్లు కూల్చివేస్తుంది.	0	Elephant	Noun	The elephant tear down the trees with its great body.
గజం	gajaM	ఇక్కడ గజం 45 వేల నుండి 50 వేల వరకు ఉంది.	1	Yard/Sq.Yard	Noun	The yard here ranges from 45 thousand to 50 thousand.
గడ్డ	gaDDa	గడ్డ కట్టుకుని పోయిన రక్తం బైటికి దుమకడానికి సిద్ధంగా ఉంది.	0	Clot/Clump	Noun	The clotted blood is ready to flow out.
గడ్డ	gaDDa	తెలంగాణ గడ్డ పోరాటాలకు నెలవుగా చెప్పుకోవచ్చు.	1	Land/Country	Noun	Telangana can be said to be the month of land struggles.
గాజు	gAju	గాజు గ్లాసులు మీద పడ్డాయి.	0	Glass	Noun	The glass glasses fell on me.
గాజు	gAju	చెల్లికి ఆ గాజు ఇచ్చింది, అది చూసి ఆమె చాలా హర్షించింది.	1	Bangle	Noun	She gave the bangle to the sister, and she looked very excited.
గుండు	guMDu	ఆ టెన్నిస్ గుండు తో ఆటగాళ్లు కఠినంగా పోరాడి విజయం సాధించారు.	0	cylindrical/spherical object/ball	Noun	The players fought hard with that tennis ball and won.
గుండు	guMDu	యాత్రలో పాల్గొన్న భక్తులు గుండు తీయించుకుని, తమ పునీతత్వాన్ని వ్యక్తం చేశారు.	1	Shaven head	Noun	Devotees who participated in the trip were shaved and expressed their saints.
చిత్రము	chitramu	సుమారు రెండు గంటల పరిమితి గల చిత్రము స్వర్గనీమ.	0	Movie/picture	Noun	Heavenly, a movie/picture of about two hours of limit.
చిత్రము	chitramu	ప్రతి ఒక్క పాత్రకు ఒక చిత్రము ఉంది మరియు ఒక వివరణాత్మక పరిచ్ఛేదము కూడా.	1	Painting	Noun	For every single character there is a painting and also a detailed accomplishment.
చుక్క	chukka	ఒక్క చుక్క నీటిని వృథా పోనీయనన్నారు.	0	Dropplet/drop	Noun	A single drop of water should not be wasted.
చుక్క	chukka	చుక్క వెలుగు నా దారిని వెలుగులతో నింపుతుంది.	1	Star	Noun	A star light illuminates my path.
చుట్ట	chuTTa	చుట్ట తాగే సమయంలో, ఆయన సముద్రపు అలలను చూస్తూ ఉండేవాడు.	0	Cigar	Noun	While taking cigar, he used to watch the ocean waves.
చుట్ట	chuTTa	భూమిని చుట్ట చుట్టి తన జేబులో పెట్టుకున్న భయంకర రాక్షసుడు.	1	Roll/wrapped	Noun	He is a terrible monster who wrapped the earth in his pocket.
ఛాయ	ChAya	వీరి చర్మ ఛాయ నల్లగా ఉన్నప్పటికీ వీరి మనస్సు చాలా స్వచ్ఛంగా ఉంటుంది.	0	Colour	Noun	Although their skin shade is black, their mind is very pure.
ఛాయ	ChAya	ఇద్దరి లోనూ విప్లవ ఛాయ గోచరిస్తుంది.	1	Shade/shadow	Noun	The revolutionary shade is visible in both of them.

తార	tAra	బాలీవుడ్ అందాల తార ఐశ్వర్యరాయ్ కూడా కంటతడి పెట్టుకొన్నారు.	0	beautiful woman	Noun	Bollywood beauty star Aishwarya Rai is also in trouble.
తార	tAra	తార ల కాంతి భూమిపై పడుతూ ప్రకృతి రంగులను వెలుగులోకి తెస్తుంది.	1	Star	Noun	The light of the stars lies on the earth and lights the colors of nature.
తాళం	tALaM	తలుపుకు తాళం వేసి బయటకు వెళ్లిపోయాడు.	0	Lock	Noun	He locked the door and went out.
తాళం	tALaM	నిన్న జరిగిన సంగీత సదస్సులో తాళం పై ప్రముఖ సంగీతకారులు మాట్లాడారు.	1	Musical Rhythm	Noun	At yesterday's music conference, prominent musicians spoke on the musical rhythm.
తీర్థం	tIrthaM	తిరుపతికి ఉత్తరంగా, తిరుపతి కొండలకు ఆనుకుని అలిపిరి దిగువకు వెళ్తే మనోహరమైన ఈ తీర్థం కనిపిస్తుంది.	0	Pilgrimage	Noun	If you go north of Tirupati, along the Tirupati Hills and below Alipiri, you will find this lovely pilgrimage.
తీర్థం	tIrthaM	అయితే ఈ ఎన్నికల సీజన్లో ఆమె తన భర్త పనిచేస్తోన్న కాంగ్రెస్ పార్టీని కాదని కాషాయ తీర్థం పుచ్చుకున్నారు	1	Being a member of a group/Taking Membership	Noun	But in this election season, She said that she is not the Congress party that her husband is working for and took membership of kasha.
దక్షిణ	dakShiNa	గృహప్రవేశం పూర్తయిన తరువాత దక్షిణ కోరుకొమ్మంటుంది పార్వతీదేవి.	0	Gift	Noun	After entering the house, Goddess Parvati seeks the gift.
దక్షిణ	dakShiNa	ఈ మొక్క దక్షిణ భారతదేశంలో కన్నా, ఉత్తర భాగాన ఎక్కువగా పెరుగుతుంది.	1	South	Noun	This plant grows more in northern India than in southern India.
దర్శనం	darshanaM	ఆలయంలో అమ్మ దర్శనం చేసుకుని అందరం వసుంధరక్కయ్య దగ్గరకు చేరుకున్నాం.	0	Visiting	Noun	After seeing Amma in the temple, we all reached Vasundharakkaya.
దర్శనం	darshanaM	ఈ ఆలోచనలను సమకాలీన సామాజిక పరిస్థితులకు అన్వయం చేసి ఈ దేశానికి ఒక దిశ దర్శనం చూపించింది రాష్ట్రీయ స్వయంసేవక సంఘము.	1	Show	Noun	These ideas were applied to contemporary social conditions and showed a direction to this country.
దళం	daLaM	యుపీఏ హయాంలో భారత వైమానిక దళం ఆమోదం తెలిపిన తరువాతే ఫ్రాన్స్	0	Force/Army	Noun	France after the Indian Air Force approved the UPA regime
దళం	daLaM	రైతు సంఘం నాయకులు మద్ధులు మాట్లాడుతూ కర్నూలు జిల్లాలో లింగాల బుర్రకథ దళం ఆదర్శమని చెప్పారు.	1	Group	Noun	Farmers' Union leaders have said that the Lingala Burkatha group in Kurnool district is ideal.
దారి	dAri	కొండల మధ్య ఎత్తైన దారి పై నిర్మాణ పనులు ప్రారంభమయ్యాయి.	0	Path/Route	Noun	Construction work began on the high path between the hills.
దారి	dAri	విద్యార్థులు విజయం సాధించేందుకు సరైన దారి అవలంబించాలని ఉపాధ్యాయులు సూచించారు.	1	Method/way off	Noun	Sports have been recognized as a way to boost enthusiasm among the youth.
దిక్కు	dikku	ఒక మార్గానికి నాలుగు మూలలు, నాలుగు దిక్కులు ఉంటే కేవలం ఒక్క మూల కాని, ఒక్క దిక్కు గాని చూపిస్తారు.	0	Direction	Preposition	If there are four corners and four directions per path, you can show only one corner or one direction.
దిక్కు	dikku	ఒక్క పెద్ద దిక్కు కోల్పోయిందని, దాసరి మృతి చెందడం బాధిస్తోందని యాంకర్ సుమ, నటుడు రాజీవ్ కనకాల దంపతులు పేర్కొన్నారు.	1	Support	Noun	The couple of anchor Suma and actor Rajeev Kanakala said that they have lost a big support and Dasari's death is painful.

ధర్మం	dharmaM	సనాతన ధర్మం నేడు అధర్మం అవుతున్నది.	1	Ancient axiom	Noun	The Ancient axiom is worse today.
ధర్మం	dharmaM	నీకు తెలియని విషయం కాదు, స్నేహం చేసేటటువంటివాడికి ఒక ధర్మం ఉంది.	0	duty	Noun	It is not a thing you don't know, but a friendship has a virtue.
నవ	nava	ఆ శబ్దం పంకజమ్మ నవ నాడుల్లో ప్రవేశించింది.	0	Nine	Noun	That sound entered Pankajamma's nine nerves.
నవ	nava	టెట్ రాసిన నవ వధువు, నవ మాత	1	New	Noun	New bride, New mother wrote TET.
పక్షం	pakShaM	అదేమంటే ఆ పక్షం ప్రభుత్వం ఏర్పాటుకు అనుచిత విధానాలు అనుసరిస్తోందన్నారు.	0	Party	Noun	Similarly, the party is following inappropriate policies for the formation of the government.
పక్షం	pakShaM	ప్రతి నెల హిందూ కాలెండర్ ప్రకారం శుక్ల పక్షం మరియు కృష్ణ పక్షముల నందు రెండు చవితులు వస్తుంటాయి.	1	forthnight	Noun	Every month there are two Chavits Shukla and Krishna fortnights according to Hindu calendar.
పక్షం	pakShaM	పాలకులు ప్రజల పక్షం ఉండాలని కోరుకుంటడు.	2	Side	Noun	The rulers want to be on the side of people.
పటం	paTaM	పండుగ సందర్భంలో దేవాలయంలో పుష్పాలంకృత పటం సుగంధాన్ని వెదజల్లుతోంది.	0	Image/Picture	Noun	In the case of the festival, the floral map/image is dispersed in the temple.
పటం	paTaM	పటంగుల పోటీల్లో పసుపు రంగు పటం లు ప్రథమ స్థానంలో నిలిచాయి.	1	Kite	Noun	The yellow kite tops the ranks in the Kites competitions.
పట్టు	paTTu	పట్టణాభివృద్ధిపై ప్రభుత్వం పట్టు బిగించింది.	0	Grip/Hold	Noun	The government has been holding up on urban development.
పట్టు	paTTu	పట్టు దుస్తులు పై కొత్త సాంకేతికతను ఉపయోగించి మెరుగైన ఉత్పత్తులు అందిస్తున్నారు.	1	Silk Cloth	Noun	Improved products are being offered using new technology on silk dresses.
పత్రం	patraM	శ్రీదేవి మరణ ధ్రువీకరణ పత్రం మధ్యాహ్నం సమయంలో జారీ అయింది.	0	Certificate	Noun	Sridevi's death certificate was issued in the afternoon.
పత్రం	patraM	మళ్ళీ ఈ దేశాలన్నీ మానవ హక్కుల పత్రం పాటిస్తామని సంతకాలు చేసిన వారే.	1	Letter	Noun	Again, all these countries have signed a human rights document/letter.
పాదం	pAdaM	తన పాదం మీద పోసాను.	0	Foot	Noun	I poured on his foot.
పాదం	pAdaM	పై పద్యం ఏ పాదానికి ఆ పాదం విడిపోతుంది.	1	A line in a stanza	Noun	The above poem breaks into stanzas.
ప్రగతి	pragati	తెలంగాణ ప్రభుత్వం ప్రతిష్టాత్మకంగా నిర్వహించబోతున్న ప్రగతి నివేదన సభపై హైకోర్టులో పిటిషన్ దాఖలయ్యింది.	0	Progress report	Noun	A petition has been filed in the High Court against the Telangana government's ambitious progress reporting meeting.
ప్రగతి	pragati	అలాంటి ఓ నూతన ప్రక్రియ ప్రగతి అను లఘుచిత్రం ద్వారా తెలుగు సాహిత్య రంగంలోకి అడుగుపెట్టింది.	1	Refers to name/news paper	Noun	Such a new process entered the field of Telugu literature through the short film Pragati.
ప్రభువు	prabhuvu	ఆ ప్రభువు ఈ ప్రభువు మధ్య తేడా తెలియకున్నది.	0	Master/Ruler	Noun	The difference between that ruler and this ruler is unknown.
ప్రభువు	prabhuvu	వారెన్నడూ (తమ ప్రభువు సన్నిధిలో) గర్వపడరు.	1	Lord/God	Noun	None of them are proud (in the presence of their Lord).
ప్రళయం	praLayaM	సర్వభూత ప్రపంచానికీ మూడు విధాల ప్రళయం ఉంటుంది.	0	Apocalypse/deluge	Noun	There will be three types of deluge for the omnipresent world.
ప్రళయం	praLayaM	ఇంకా ఎంతకాలం ఈ ప్రళయం కొనసాగుతుందో, ఎంత నష్టం సంభవిస్తుందో అర్థం కాని పరిస్థితి.	1	Flood	Noun	How long this flood will continue and how much damage is not understood.
ఫలము	phalamu	ఈ సీజన్‌లో ఆమ్రపాలి మామిడి ఫలము కు భారీ డిమాండ్ నెలకొంది.	0	Fruit	Noun	This season, there is a huge demand for the fruit of Amrapali mango.
ఫలము	phalamu	ఈ సంవత్సరం కృషి చేసిన విద్యార్థులకు పరీక్షలలో మంచి ఫలము లు దక్కాయి.	1	Result/Produce	Noun	Students who worked this year have got good result in the exams.

రాశి	rAshi	నీ గోధుమ పంట రాశి అగ్నికి ఆహుతి అవుతూ ఉంది" అని కొండ్యాతో అన్నారు.	0	Heap/Pile	Noun	he said to Kondya that his wheat crop heap is being consumed by fire.
రాశి	rAshi	ఈ వారం మీ రాశి ఫలాలు విజయవంతంగా ఉండే అవకాశాలు ఉన్నాయి.	1	Zodiac Sign/horo-scope	Noun	Chances say that your hor-oscope will be successful this week.
వేరు	veru	పాటలు అన్నీ చిత్రీకరణ వేరు వేరుగా వుంటుంది.	0	Different	Adjective	Filming of all the songs is different.
వేరు	veru	ఇక లిల్లి మొక్కల వేరు భాగంలో ఉండే దుంపలు అత్యంత పోషక విలువలున్నవిగా జపాన్	1	Plant root	Noun	And tubers in the root of lily plants are the most nutri-tious in Japan
శూన్యం	shUnyaM	కానీ ఆచరణలో మాత్రం ఆ నిబద్ధత శూన్యం అని పదే పదే రుజువు అవుతోంది.	0	Null/Void	Adjective	But in practice, that com-mitment has been proven time and again to be null and void.
శూన్యం	shUnyaM	ఆ శూన్యం అనేది లేనప్పుడు జీవితం రంగులమయమే కదా.	1	universe	Noun	Isn't life colorful in the ab-sence of that universe?
సూత్రం	sUtraM	ఈ సూత్రం ఉపయోగించాలంటే ఫెర్మి-డిరాక్ గణాంకాలు వాడాలి.	0	Principle	Noun	In order to use this princi-ple, Fermi-Dirac statistics must be used.
సూత్రం	sUtraM	వాళ్ళు చెప్పినట్టు గానే ఆవిడ మెడలో మంగళ సూత్రం కూడా లేదు.	1	Nuptial Chain	Noun	As they say, she does not even have the Nuptial Chain on her neck.

Scatterplots of the 5 words which taken in the VI. Results & analysis:

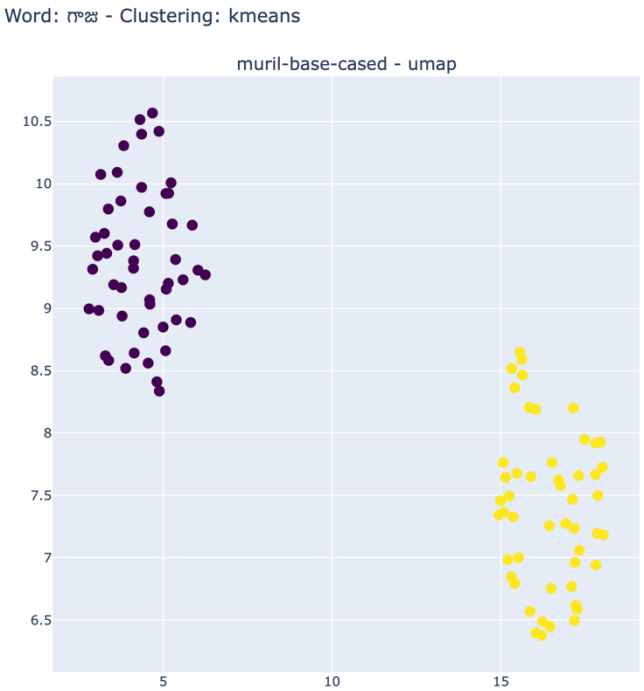

Figure A: Scatterplot of the word " గాజు "(gaaju) using muril base cased

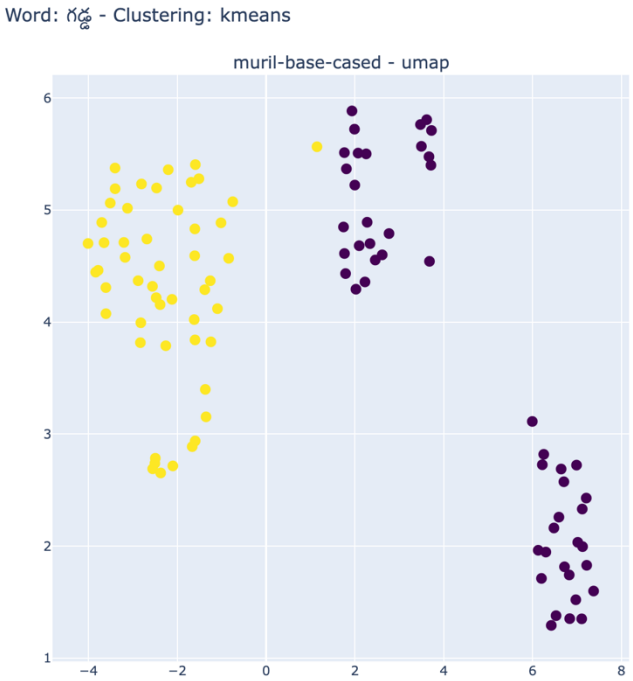

Figure B: Scatterplot of the word " గడ్డ "(gadda) using muril base cased

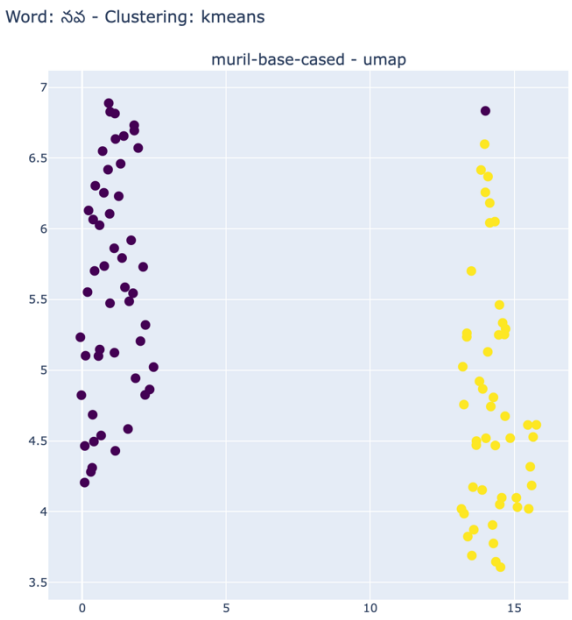

Figure C: Scatterplot of the word " నవ "(nava) using muril base cased

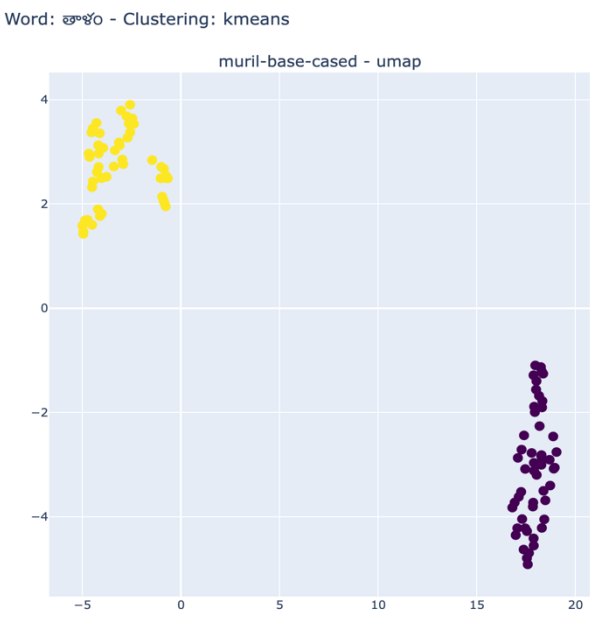

Figure D: Scatterplot of the word " తాళం "(taalam) using muril base cased

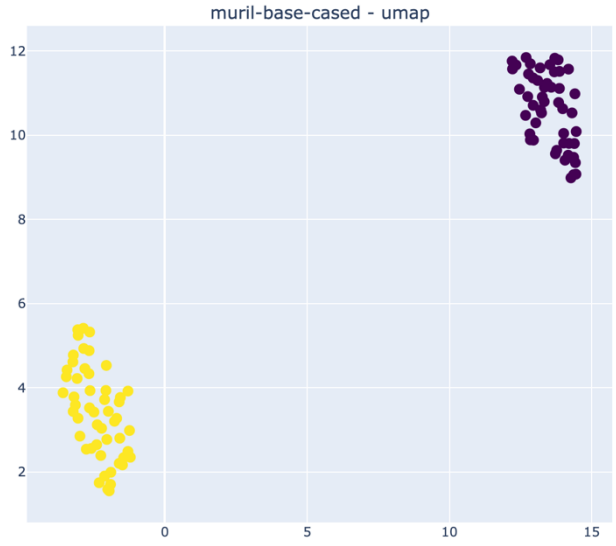

Figure E: Scatterplot of the word " కూలి "(kooli) using muril base cased