MalUpama - Figurative Language Identification in Malayalam -An Experimental Study

Reenu Paul
Research Scholar
Department of Computer Applications,
Hindustan Institute of Technology & Science,
Chennai, India
rp.23703012@student.hindustanuniv.ac.in

Wincy Abraham
Assistant Professor
Department of Computer Science,
Assumption College (Autonomous),
M.G University,
Changanasserry, Kerala,India
wincya@gmail.com

Anitha S Pillai
Professor
Department of computer Applications,
Hindustan Institute of Technology &Science,
Chennai,India
anithasp@ hindustanuniv.ac.in

Abstract

Figurative language, particularly in under represented languages within the Dravidian family, serves as a critical medium for conveying emotions and cultural meaning. Despite the rich literary traditions of languages such as Malayalam, Tamil, Telugu, and Kannada, there has been minimal progress in developing computational techniques to analyze figurative expressions. Historically, Malayalam was known by various names, such as Malayanma and Malabari. Similarly Kerala was known as Malanadu before adopting its current name, which metaphorically refers to the land between the Indian Ocean and the Western Ghats. In this study, we introduce the UPAMA Model(MalUpama), designed to identify Similes in Malayalam, an under-resourced Dravidian language mostly spoken in the state of southern India, Kerala. The current research focuses on detection of presence of Simile in Malayalam prose using the 'Upama model'. This paper outlines the detection of Simile in Malayalam sentences and a detection accuracy of 94.5% is achieved by the proposed method. To the best of our knowledge this is the first work in the Malayalam language, explores computational techniques with a particular focus on applying machine learning to analyze figurative expressions which can be adopted for other Dravidian Languages too. The dataset developed for this study is made publicly available, allowing scholars to contribute and explore more on the category 'Upama' of Figurative Languages ('Alankarangal') of Malayalam language.

1 Introduction

Although Dravidian languages are spoken by many, they still remain underrepresented and have not received appropriate attention in computational linguistic research. Figurative language like Simile can make strong emotional response and the lack of representation create a significant gap in research, especially in the Dravidian languages. This limitation is particularly noticeable in Malayalam, where Similes, or 'Upama Alankara', are important rhetorical persuasion and plays a major role in expressing complex emotions in poetic expressions. The current paper presents a machine learning model, to analyze 'Upama Alankara'(Simile) in the Dravidian language Malayalam. A.R. Rajarajavarma(a prominent Malayalam poet and scholar) in his Book "*BhashaBhushanam"* in the chapter "Alankara Pramanagal", states that a figurative language pro vides readers immense enjoyment, by converting a simple sentence into extra ordinary art form. Additionally, it serves as cultural markers that protect language tradition in addition to enhancing poetic expressions.

The phrase "My life burns and melts like a candle" conveys an emotional depth, denoting intense distress and suffering beyond a simple expression of love. Thus, the study of Figurative languages, especially Similes, is vital for sentiment analysis because it may capture complex emotional states.

The lyrics "പുഷ്പപാതകംപുറത്തുവച്ച്"
"നീ നഗ്നപാതയായിഅകത്തുവരൂ"

(Pushpa pathakam purathu vechu, nee nagnapathayayi akathu varu) carry a deeper expression of thought. The lines convey the traditional practice of removing one's footwear before entering a space, representing a special importance of respect and purity when stepping into a sacred or meaningful environment. The computational analysis of figurative languages often opens vast avenues for researchers to delve deep in to the cultural and linguistic heritage of a land and thus contribute to the study in various research areas. We propose Machine Learning model MalUpama to detect Simile in Malayalam.

2 Motivation

Figurative language, particularly Simile in underrepresented languages like Malayalam in the Dravidian family, plays a crucial role in expressing deep emotions and cultural significance. Taraka Rama et.al (2024) highlighted that Dravidian languages are spoken by millions of people worldwide. NLP in Dravidian languages face many challenges and at the same time give exciting research opportunities in the field of sentiment analysis, stress detection, machine translation etc.

The main objective of this research is to create an 'UPAMA model-MalUpama' to identify the figurative language Simile in sentences and thus contribute to the language technology in resource-poor Dravidian languages like Malayalam. In Malayalam poetry and prose, *Upama Alankaram* (Simile) enriches the sentence's aesthetic beauty by crafting vivid imagery. This helps readers or listeners to clearly perceive the comparisons between different objects or ideas. The creation of the 'UPAMA Model-MalUpama' offers a method for detecting Similes in Malayalam. It sets the groundwork for computational studies across various Dravidian languages.

3 Literature review

Simile detection in Natural Language Processing (NLP) is a special and developing area of research. We conducted an extensive search for literature focusing on Dravidian languages. However, the available studies in this domain are found to be quite limited. So, we expanded our research area to encompass English, cross-lingual languages and Indian Languages thus enabling a deeper understanding of the current research initiatives in Simile identification within natural language processing. This expansion provided deeper insights into the methodologies and progress made in this field. Advances in deep learning and machine learning techniques have improved Simile detection systems' efficiency and precision in a variety of languages. The literature review is organized into three main categories.

3.1 Simile Detection in English

A self-verifying approach was introduced by Longxuan Ma et al. (2024) which aimed to improve the pre-trained language models to detect and understand Similes. They used a multi-layered Simile recognition framework combined with the integration of diverse forms of Simile. To enhance the variety of Simile structures in training datasets, Chang et al. (2023) developed the I-WAS (Iterative Word Replacement and Sentence Completion) technique. They used GPT-2 for metaphor detection. The metaphor identification models accuracy can be influenced by biases in dataset design, as shown by Joanne Boisson et al. (2023). They suggest that natural corpora is essential for accurate results.

Neural multitask learning is another area based on the study by Lizhen Liu et al. (2018) for Simile recognition research. It highlights the advantages of using a multitask learning framework. Three primary objectives are introduced by this framework: language modeling , Simile extraction ,and Simile sentence classification. This study suggests that even though there are difficulties, multitasking gradually improves the performance in both classification and component extraction tasks.Tuhin et.al(2022) used a contextual model ParaCOMET (Paragraph-level Commonsense Transformers with Recurrent Memory) to interpret the Figurative language .

3.2 Simile Detection in Cross Lingual

Yulia Tsvetkov et al.(2024) introduced a model trained on English data that can be effectively transferred to other languages like Spanish, Farsi, and Russian using a bilingual dictionary. This supports the hypothesis that metaphors are conceptual rather than purely lexical, meaning the underlying conceptual mappings can be detected across different languages. Xiaoyue et.al(2023) proposed a graph-based methodology which extracts Simile components such as tenor and vehicle using advanced.

3.3 Simile Detection in Indian Languages

Abhishek et al. (2024) have analysed figurative language especially hyperboles, metaphor and their creation and focus on blending to figurative language by analysing the text. They give importance more to the practical application of text. J. Auhiainen et al. (2021) highlight the differences between conventional and contemporary techniques. The study employed a dataset of 16,674 YouTube comments in Roman script. Bharathi Raja Chakravarthi and colleagues (2020) provide a manually annotated dataset particularly for sentiment analysis and for identifying offensive language in three low-resource Dravidian languages—Tamil, Kannada, and Malayalam—which includes over 60,000 comments sourced from YouTube, and it is useful for multilingual research. Xiaotian Lin et al. (2021) developed the Masked Language Model (MLM) technique for language-specific terms and adversarial training to solve text information biases in multi-task learning. Experimental results show significant performance improvements in multilingual text classification tasks, leveraging existing language representations. A Hande et al. (2021) demonstrate the effectiveness of multi-task learning approach in sentiment analysis and offensive language detection in various Dravidian languages like Tamil, Malayalam, and Kannada. The study discusses that multi-task learning models outperform the single-task learning models, with higher F1 scores for sentiment analysis. The authors compare various language models like mBERT and Distil BERT, in order to identify the best performance for each language under different parameters.

Joseph Marvin et al. (2024) use STANDARDIZE framework to align large language models for content generation . They use a contextual learning approach demonstrating a notable improvement in the accuracy of content generation, especially within the educational domain using CEFR and CCS guidelines. Goyal et al.(2022) focused to increase translation accuracy of Dravidian languages, particularly in low-resource environments. They use models like OpenNMT, IndicNLP used for machine translation projects. Language Transformer models, such as XLM-RoBERTa have contributed promising capabilities in handling multilingual data. Sangeetha et al. (2024) generated code-mixed corpora in Tamil and English specifically for sentiment analysis.

Abu Bakkar Siddique Raihan et al. (2024) introduced advanced transformer-based models to deal with stress detection in low-resource languages like Tamil and Telugu. The models achieve macro F1-scores of 0.71 for Tamil and 0.72 for Telugu. Debjoy Saha et al. (2022) demonstrated the effectiveness of multiple feature extraction methods and transformer topologies in stress detection. Chattu et.al (2024) emphasize the performance of deep learning models in sentiment prediction accuracy. BERT tokenizer and BERT Embeddings help to capture contextual information of each word based on the surrounding context according to Anuja K et.al (2024).

While performing literature analysis we came across papers related to sentiment analysis but could not find any in the area of Figurative Language Identification in Dravidian Languages. This indicates that there is a significant disparity in studies among languages, primarily due to the fact that some languages lack sufficient resources. Working with such under-resourced languages is challenging, and researchers encounter numerous obstacles since no foundational work has been done previously.To address this significant gap, we have developed the ‘Upama model’. The research question then becomes: In what ways can the development of the ‘Upama model’ help mitigate the challenges associated with resource-scarce languages and aid in bridging this gap? We have focused on objectives such as the development, optimization, and implementation of the ‘UPAMA model’ for Malayalam language to identify Similes in Malayalam script.

4 Dataset Preparation

In order to effectively train any machine learning model, a large amount of data is essential. The dataset preparation for ’UPAMA Model-MalUpama’ required several key steps to ensure that the data is appropriate for both analysis and model training. Below is a summary of the dataset preparation process as carried out in this research work.

Malayalam boasts a wealth of literary and general texts, offering a broad selection of material for linguistic study. We have tried to extract lines from Malayalam news portals and various social media platforms. However, Similes are scarce in sentences, appearing in roughly two out of every hundred sentences. Despite the existence of many books, newspapers, magazines, and online resources, there is still a significant lack of a comprehensive text corpus, particularly in the area of figurative language. The development of NLP tools for Malayalam and computational linguistics research are hampered by this limitation.

To overcome this limitation, we developed a specialized Malayalam text corpus centered around Similes. This corpus is designed to aid computational linguists, NLP researchers, and scholars by providing vital resources for linguistic feature analysis, research, and information retrieval tasks. Understanding the significance of such a corpus, we curated a collection of Malayalam Simile sentences, which plays an essential role in Simile model creation.

The process of creating this dataset was carefully planned and executed, using large language model tools such as ChatGPT, Google Gemini, and Claude AI. ChatGPT, a dialogue generation tool based on transformer architecture, was used for this task, having been trained on extensive conversational data to generate human-like text. Similes are heavily used in various forms in normal conversations in Malayalam, the machine generated sentences can be easily anlayzed for correctness. With assistance from selected volunteers, we compiled a dataset of 1,000 Simile sentences and 1,000 non-simile sentences in Malayalam. Figure 1 shows the block diagram of the method we used for the data set creation.

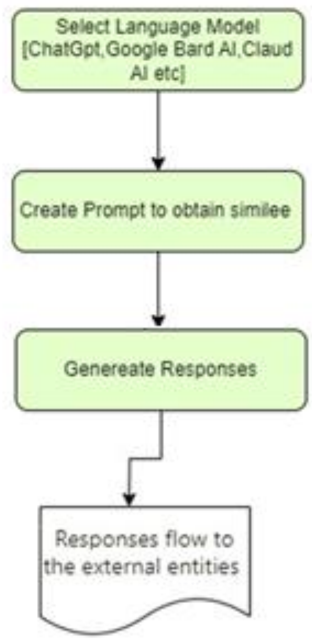

Figure 1: Data Preparation

4.1. Prompt Design

We used a prompt to ChatGPT to generate the sentences in the required form.

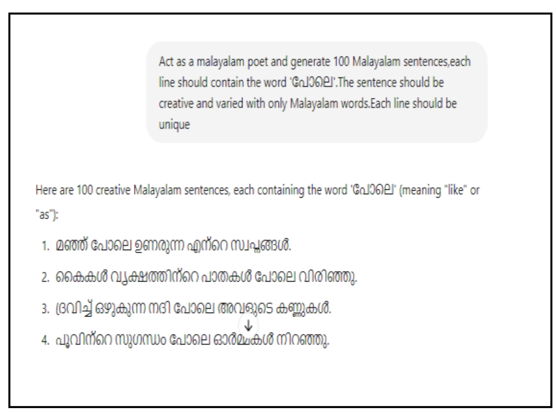

Figure 2: Prompt and corresponding simile sentences

Figure 2 and Figure 3 show the prompt given to ChatGPT for generating Simile and non-simile sentences respectively and the corresponding sequences obtained for both Simile and non-simile sentences.

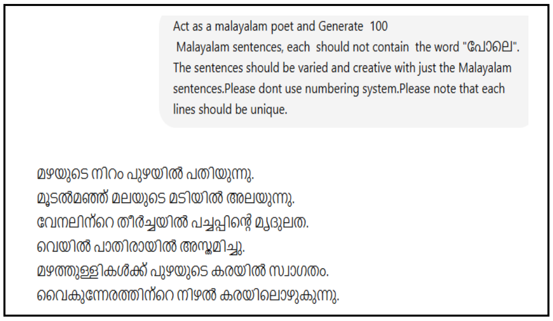

Figure 3: Prompt and corresponding non-simile sentences

The comparative analysis of Simile generation using the various approaches is displayed in Figure 4.

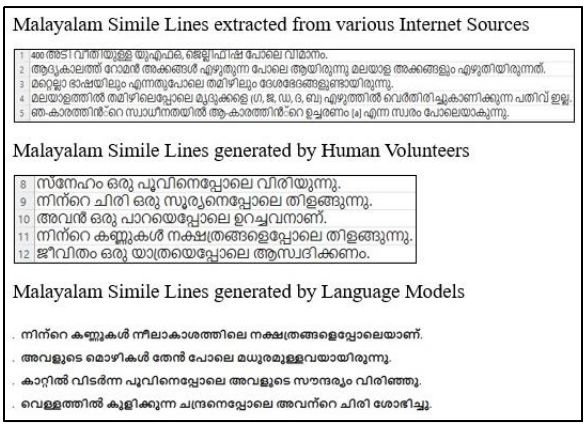

Figure 4: Comparative Analysis of Various Approaches

We can draw the conclusion that the language model's output is on par with that of humans. It is quite challenging to determine if it is created by a machine or a human- a true indication of the advancements in computational language technology.Prompt is supplied to ChatGPT in such a way that Simile sentences and non-simile sentences are produced with the least chance of mistakes. With careful evaluation of the results from ChatGPT and data preprocessing we could achieve our target of 2000 sentences comprising of both Simile and non-simile sentences.

4.2 Data Preprocessing

In order to develop Malayalam 'Upama model' to identify Similes ('upama') in Malayalam, preprocessing the data is crucial to ensure the quality and accuracy of the model. The preprocessing of the automatic generated sentences are performed to ensure that the text data is suitable for training and prediction. Short lines or sentences with less than three words are not used for Simile detection and thus removed as a preprocessing step. We also ensured that only sentences with proper context are included and others are eliminated. This helps in avoiding ambiguous or meaningless sentences in the dataset. A final validation was made to make sure that each line met the length requirements and contained valid words before passing them through the model. So these steps form the foundation for the successful implementation of the Simile model, enabling accurate identification of upama in Malayalam lines.

4.3 Data Labelling

As the model is expected to perform binary classification, we have used binary labelling, numerical value 1 is used to label Simile sentences and 0 to label non simile sentences.

5 Methodology

The development of Simile detection Model- 'MalUpama' in Malayalam focuses on detecting and analyzing the figurative language *Upama* (Simile) using machine learning techniques. The Flow Diagram of MalUpama is displayed in Figure 5.

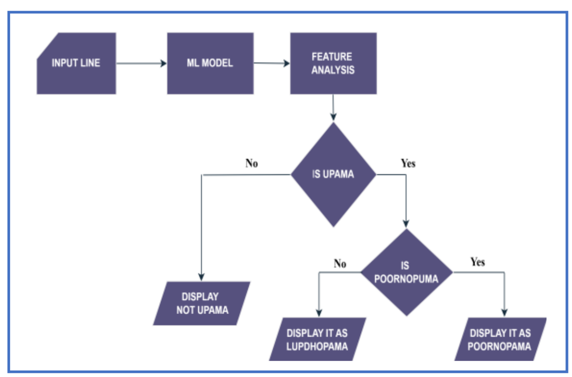

Figure 5: Flow Diagram of the Upama Model

5.1 Loading the data set

The dataset is loaded from an external file using the DataFrame of Pandas library in Python. Sentences and the corresponding labels are extracted for further processing. The dataset, stored in an Excel file, contains both text data with the corresponding labels. The text data serves as the input features, while the labels act as the target variables that the model is trained to predict or classify. This step is crucial for preparing the data in a structured manner.

5.2Text Vectorization Using CountVectorizer

Since machine learning models cannot directly process text data, it needs to be converted into a numerical format**.** In this work, the CountVectorizer of the scikit-learn library is utilized to perform the conversion. Specifically, Count Vectorizer(binary=True) is applied to represent the text data in a binary format. This converts the text into a numerical vector by using a binary representation. The binary parameter indicates that the vectorized data will only contain 0s and 1s, representing the presence or absence of a specific token (in this case, the Malayalam word 'പോലെ'). The pattern token pattern=r'പോലെ' defines 'പോലെ' as all Simile sentences in Malayalam are expected to include the word 'പോലെ'. The vectorizer will search for this word in each sentence, assigning a value of 1 if it appears and 0 if it does not. By fitting the vectorizer to the text data (sentences), it transforms the data into a sparse matrix X, where each row corresponds to a sentence, and each column indicates the presence or absence of the word 'പോലെ'. This step is a form of text vectorization, an essential part of text data preprocessing where we convert raw text into a structured numerical form that machine learning models can work with. The presence of the pattern 'പോലെ' is used for the Vectorizer setup and its presence is important for the classification task. Removing this will degrade the model performance.

5.3 Splitting the Data into Training and Test Sets

The dataset is divided into training and testing subsets using the train_test_split function available in the scikit-learn library in Python. In this work, 80% of the data is used for training while the remaining 20% is kept aside for testing.

5.4 Model Selection

Once the dataset is ready and features are extracted, the next concern is regarding the selection of the appropriate model for classification. A Rule-based classifier is simple and can perform well for cases where a simple if -then rule can do the classification. Another choice is the Linear Regression Model which can fit the data to a straight line that works well when the data carries linear relationships. Support Vector Machine classifier is good for classification especially when the data has some non-linear relationships. SVM particularly excels at managing high-dimensional data. An advantage of SVM classifier is the presence of hyper-parameters which makes it more versatile by making it aligned with the data by setting appropriate values to these parameters. We experimented with the supervised learning models in machine learning and the Support Vector Machine(SVM) classifier is found to perform better than the other models.

Thus, in this work we have chosen a Support Vector Machine (SVM) classifier to perform the binary classification, using the Radial Basis Function (RBF), kernel (kernel='rbf'), which is effective for handling non-linear data. The model is first trained on the training dataset.

6 Experimental analysis

We ensured that all lines that passed the validation steps during preprocessing are ready for training and prediction. Two thousand lines of Malayalam text were created for this purpose and used for training and testing. The training phase used the presence/absence of ”പോലെ” as the feature (X) and the corresponding label (y) as the target variable. Hyper-parameter **'rbf'** random state in SVM is set empirically as 0.42.

The model is trained using 80% of the data and its performance evaluated using the test set, the remaining data in the dataset. The model shows the accuracy of 94.5%, and a precision of 98% as shown in Table 1. During the error analysis we found that some sentences got identified wrongly as Simile due to the absence of contextual evaluation. This error has to be fixed in the future work.

Method	Accuracy	Precision	Recall
MalUpama-Model	94.5 %	98%	94.4%

Table 1:Performance Metrics

7. Conclusion & Future Scope

'Upama', Simile in Malayalam, can be found in Poetry and Prose : Our current research focuses on detection of presence of Simile in prose using the 'Upama model' an SVM binary classifier. Additional contextual features (e.g., the length of the line or surrounding words) can be considered to improve the model's performance and reduce misclassifications as a future work. Fine tuning this model to predict Upamanam(Tenor), and Upameyam(Vehicle) based on the context also is set as a future work. For that use of part of speech tags is required to identify 'Upamanam' and 'Upameyam' which is a cumbersome task especially in low resource language like Malayalam. In addition to this the developed model should benefit from a more complex and diverse dataset. A significant contribution of this study is the successful creation of a dataset for 'Upama', in Malayalam and creation of a machine learning model to detect an essential element of figurative language- Simile . Future work will emphasis on deep learning techniques which will have a contribution towards sentiment analalysis of Malayalam.

References

Rama, T., & Vajjala, S. (2024). Dravidian language family through Universal Dependencies lens. arXiv preprint arXiv:2406.14680.

Ma, L., Ke, C., Zhou, S., Sun, C., Zhang, W., & Liu, T. (2024, May). A Self-verified Method for Exploring Simile Knowledge from Pre-trained Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1563-1576).

Chang, Y., Zhang, R., & Pu, J. (2023, August). I-WAS: A Data Augmentation Method with GPT-2 for Simile Detection. In *International Conference on Document Analysis and Recognition* (pp. 265-279). Cham: Springer Nature Switzerland.

Boisson, J., Espinosa-Anke, L., & Camacho-Collados, J. (2023). Construction artifacts in metaphor identification datasets. *arXiv preprint arXiv:2311.00790*.

Ma, L., Ke, C., Zhou, S., Sun, C., Zhang, W., & Liu, T. (2024, May). A Self-verified Method for Exploring Simile Knowledge from Pre-trained Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1563-1576).

Liu, L., Hu, X., Song, W., Fu, R., Liu, T., & Hu, G. (2018). Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1543-1553).

Chakrabarty, T., Choi, Y., & Shwartz, V. (2022). It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, *10*, 589–606. https://doi.org/10.1162/tacl_a_00478

Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., & Anastasopoulos, A. (2024). DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *arXiv preprint arXiv:2403.11009*.

Wang, X., Song, L., Liu, X., Zhou, C., & Su, J. (20222). Getting the most out of simile recognition. *arXiv*. https://doi.org/10.48550/arXiv.2211.05984

Chauhan, S., Kumar, R., Saxena, S., Kaur, A., & Daniel, P. (2024). SemSyn: Semantic-Syntactic Similarity Based Automatic Machine Translation Evaluation Metric. *IETE journal of Research*, *70*(4), 3823-3834.

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, *11*, 36120-36146.

Jauhiainen, T., Ranasinghe, T., & Zampieri, M. (2021). Comparing approaches to Dravidian language identification. *arXiv preprint arXiv:2103.05552*.

Chakravarthi, A., & Raja, B. (2020). *Leveraging orthographic information to improve machine translation of under-resourced languages* (Doctoral dissertation, NUI Galway).

Lin, X., Lin, N., Wattanachote, K., Jiang, S., & Wang, L. (2021). Multilingual text classification for dravidian languages. *arXiv preprint arXiv:2112.01705*.

Hande, A., Hegde, S. U., Priyadharshini, R., Ponnusamy, R., Kumaresan, P. K., Thavareesan, S., & Chakravarthi, B. R. (2021). Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *arXiv preprint arXiv:2108.03867*.

Imperial, J. M., Forey, G., & Madabushi, H. T. (2024). Standardize: Aligning Language Models with Expert-Defined Standards for Content Generation. *arXiv preprint arXiv:2402.12593*.

Aggarwal, S. (2022). *Exploiting Indian Languages' Similarity for Different NLP Applications* (Doctoral dissertation, International Institute of Information Technology Hyderabad).

Sangeetha, M., & Nimala, K. (2024). DL-TBAM: Deep Learning Transformer based Architecture Model for Sentiment Analysis on Tamil-English Dataset. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-15.

Raihan, A., Rahman, T., Rahman, M., Hossain, J., Ahsan, S., Das, A., & Hoque, M. M. (2024, March). CUET_DUO@ StressIdent_LT-EDI@ EACL2024: Stress Identification Using Tamil-Telugu BERT. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (pp. 265-270).

Nayak, S., Schuler, C., Saha, D., & Baumann, T. (2022, November). A Deep Dive Into Neural Synchrony Evaluation for Audio-visual Translation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (pp. 642-647).

Chattu, Kannaiah & Sumathi, D.. (2024). Sentiment Analysis Using Deep Learning Approaches on Multi-Domain Dataset in Telugu Language. Journal of Information & Knowledge Management. 23. 10.1142/S0219649224500187.

Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *arXiv preprint arXiv:1904.09077*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Anuja K., Reghu Raj, P. C., & Remesh Babu, K. R. (2024). Emotion Detection System for Malayalam Text using Deep Learning and Transformers https://dl.acm.org/toc/tallip/justaccepted

MalUpama - Figurative Language Identification in Malayalam -An Experimental Study(Supporting Document)

Figure 1: Data Preparation

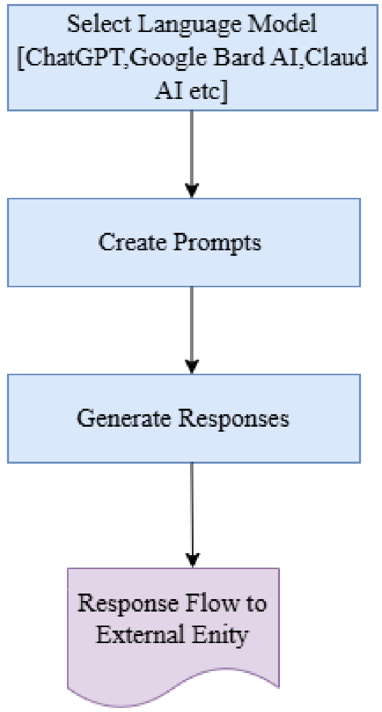

Figure 2: Prompt and corresponding simile sentences

MalUpama - Figurative Language Identification in Malayalam -An Experimental Study(Supporting Document)

Figure 3: Prompt and corresponding non-simile sentences

Act as a malayalam poet and Generate 100 Malayalam sentences, each should not contain the word "പോലെ". The sentences should be varied and creative with just the Malayalam sentences.Please dont use numbering system.Please note that each lines should be unique.

മഴയുടെ നിറം പുഴയിൽ പതിയുന്നു.
മൂടൽമഞ്ഞ് മലയുടെ മടിയിൽ അലയുന്നു.
വേനലിന്റെ തീർച്ചയിൽ പച്ചപ്പിന്റെ മൃദുലത.
വെയിൽ പാതിരായിൽ അസ്തമിച്ചു.
മഴത്തുള്ളികൾക്ക് പുഴയുടെ കരയിൽ സ്വാഗതം.
വൈകുന്നേരത്തിന്റെ നിഴൽ കരയിലൊഴുകുന്നു.

Figure 4: Comparative Analysis of Various Approaches

Malayalam Simile Lines extracted from various Internet Sources

1	400 അടി വീതിയുള്ള യുഎഫ്ഒ, ജെല്ലിഫിഷ് പോലെ വിമാനം.
2	ആദ്യകാലത്ത് റോമൻ അക്കങ്ങൾ എഴുതുന്ന പോലെ ആയിരുന്നു മലയാള അക്കങ്ങളും എഴുതിയിരുന്നത്.
3	മറ്റെല്ലാ ഭാഷയിലും എന്നതുപോലെ തമിഴിലും ദേശഭേദങ്ങളുണ്ടായിരുന്നു.
4	മലയാളത്തിൽ തമിഴിലെപ്പോലെ മൃദുക്കളെ (ഗ, ജ, ഡ, ദ, ബ) എഴുത്തിൽ വേർതിരിച്ചുകാണിക്കുന്ന പതിവ് ഇല്ല.
5	ഞ-കാരത്തിൻ്റെ സ്വാധീനതയിൽ ആ-കാരത്തിൻ്റെ ഉച്ചരണം [a] എന്ന സ്വരം പോലെയാകുന്നു.

Malayalam Simile Lines generated by Human Volunteers

8	സ്നേഹം ഒരു പൂവിനെപ്പോലെ വിരിയുന്നു.
9	നിന്റെ ചിരി ഒരു സൂര്യനെപ്പോലെ തിളങ്ങുന്നു.
10	അവൻ ഒരു പാറയെപ്പോലെ ഉറച്ചവനാണ്.
11	നിന്റെ കണ്ണുകൾ നക്ഷത്രങ്ങളെപ്പോലെ തിളങ്ങുന്നു.
12	ജീവിതം ഒരു യാത്രയെപ്പോലെ ആസ്വദിക്കണം.

Malayalam Simile Lines generated by Language Models

. നിന്റെ കണ്ണുകൾ നീലാകാശത്തിലെ നക്ഷത്രങ്ങളെപ്പോലെയാണ്.

. അവളുടെ മൊഴികൾ തേൻ പോലെ മധുരമുള്ളവയായിരുന്നു.

. കാറ്റിൽ വിടർന്ന പൂവിനെപ്പോലെ അവളുടെ സൗന്ദര്യം വിരിഞ്ഞു.

. വെള്ളത്തിൽ കുളിക്കുന്ന ചന്ദ്രനെപ്പോലെ അവന്റെ ചിരി ശോഭിച്ചു.

MalUpama - Figurative Language Identification in Malayalam -An Experimental Study(Supporting Document)

Figure 5: Flow Diagram of the Upama Model

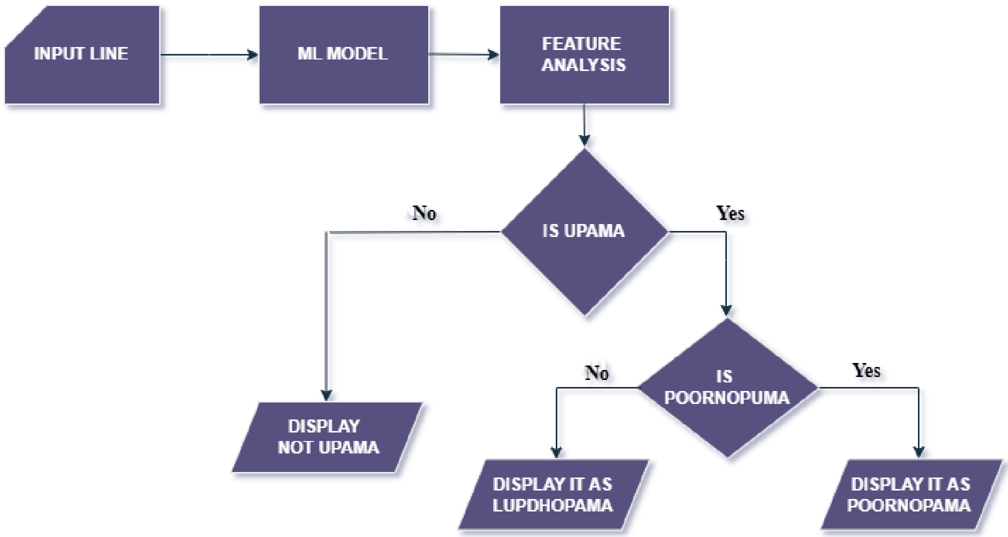

Table 1:Performance Metrics

Method	Accuracy	Precision	Recall
MalUpama-Model	94.5 %	98%	94.4%