Scope
Volume 13 Number 4 December 2023

A Systematic Review on Various Approaches for News Headlines Categorization in Malayalam Language

Rameesa K.[1]**& K. T. Veeramanju**[2]

[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore – 575001, Karnataka India, ORCID ID: 0000-0003-0971-1661

[2]Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore – 575001, Karnataka India, ORCID ID: 0000-0002-7869-3914

Abstract

Purpose: It is said that newspaper is something that we cannot get rid of. It is very necessary for us to know about our surroundings and to know about the things in our world. In today's situation, people like to read everything online, unlike newspaper printed on paper which is commonly found in today's technological world. For the reader, it is very good to save time to get each news item separately. But readers have made everything online and the biggest problem is in the area of reading newspapers once they get all the news together without separating them. Not only that, but in our newspaper industry, sorting the news into each category at the time of printing requires a lot of labor. Therefore, many research works has been done in various languages regarding the categorization of news. This paper discusses a detailed literature survey of various approaches used to categorize news headlines in Malyalam language. **Design/Methodology/Approach:** The details collected for this review paper were obtained by analysing and comparing different research articles from recognized resources. **Objective:** To find a research gap and appropriate solutions for news headlines categorizations in Malayalam. **Results/ Findings:** Review of this paper gives a proper understanding of news headlines categorization in Malayalam and other languages in connection with machine-learning and deep learning approaches. **Originality/Value:** The review of this paper exhibits an analysis of machine learning algorithms for news headlines categorization in Malayalam and other languages and suggests the importance of news headlines categorization.

Type of Paper: Literature Review.

Keywords: languages, news headlines, machine learning, categorization, algorithms, Malayalam, deep learning, rule based, statistical approaches, summarization, automatic, offensive, methods, hybrid, accuracy, performance

1. Introduction :

The categorization of news headlines has become a crucial problem in the fields of information retrieval and natural language processing (NLP) in this era of abundant and broadened information. The task of efficiently classifying news is very difficult because there are many different languages, each with peculiar syntactic and semantic peculiarities. With the growing digital landscape, the vast quantity of news content demands advanced computational methods to filter, arrange, and rank news for end users. This systematic review aims to bring together the variety of approaches used in the categorization of news headlines in various language contexts, emphasising the unique characteristics and innovations that define this ever-evolving field of research.

News headlines are the point of contact between readers and information; they are a brief summary intended to enlighten, draw in, or even influence public opinion (Nguyen, D. et al.(2022). [1]). Because of this, their classification is more than just a theoretical exercise; it has important implications for journalism, media consumption, and public discourse (Watson, H. et al. (2021). [2]). Classification techniques have developed over time, moving from straightforward rule-based algorithms to advanced machine learning and deep learning models, all of which aim to negotiate the challenging field of linguistic diversity (Zhou, N. et al. (2022). [3]).

This review is highly significant since it discusses the cultural complexities and technology developments that influence headline classification. The task is far from uniform in a multilingual society, where each language poses distinct challenges: the contextual richness in Chinese, Korean, Arabic, Hindi, the agglutinative structures in Malayalam, or the subtle metaphorical language in English headlines (Jha, A. et al. (2023).[4]). This work integrates studies from a diverse array of sources, analysing the performance and flexibility of alternative classification models in the face of linguistic nuances.

We set down the foundation for identifying best practices, areas requiring more research, and the possibility for cross-linguistic application of categorization models by offering a thorough evaluation about the existing literature. This paper implements a comparative analysis of the efficacy of modern algorithms, including transformers, recurrent neural networks (RNNs), and convolutional neural networks (CNNs), in the field of news headline classification.

2. Objectives of Review Paper :

(1) To review the present state of research on the categorization of news headlines in Malayalam and other languages.

(2) To evaluate the relative merits of rule-based, machine learning, and deep learning models for news headline classification

(3) To determine the merits and demerits of current approaches for precision, effectiveness, and expandability in the classification of news headlines in several languages with differing grammatical and syntactic characteristics.

(4) Tosummarise the results of earlier research to enhance comprehension of the strategies that have worked best and why, as well as to point out any research gaps that may need to be solved.

(5) Tooffer suggestions for future lines of inquiry, covering possible enhancements to existing models, investigating hybrid strategies, and utilising innovative technologies like transformer-based models like BERT and GPT.

(6) Toinvestigate the useful applications of news headline classification, including how it affects news aggregation, content creation, and news feed personalization in multilingual environments.

(7) Understanding how algorithms trained in one language may need to be modified or fail when applied to another is necessary to evaluate the transferability of classification models across languages and cultures.

(8) To gather and examine datasets in various languages that are used for training and validating headline classification models, taking note of the data sets' representativeness, size, quality, and accessibility.

3. Methodology :

Data collected from a range of sources, such as scholarly journals, conference papers, websites, and publications, forms the basis of the analysis.

4. Review of Literature/ Related Works:

4.1 Approaches to Text Classification

One of the most crucial tasks of NLP is text classification. Sometimes it is mentioned as text categorization, is the act of automatically categorising text documents into specified labels or groups based on their content. Training a model using a labeled dataset, in which each document is categorized, constitutes a task in supervised machine learning. Once trained, the model may allocate the proper categories to fresh, unclassified documents. As part of text classification, the text is first preprocessed, then features are extracted, and finally the model is created. All the techniques and models for this are available from NLP. By producing innovative results in a various natural language retention jobs, well-known NLP models like BERT, GPT-3, and others have considerably advanced the field of text classification.

Applications of Text Classification

Following are the important applications of text classification.

1. An important function of text classification is to distinguish between what we are receiving and whether it is spam or not. This process is called spam email detection [24].
2. After reading a text message, decide if it expresses a neutral, negative, or positive emotion. This procedure known as sentiment analysis.
3. Topic categorization means that we can categorize the different types of news articles or block posts that we get by looking at the type of topic under which they fall that is, it can be distinguished whether it is politics, sports, or technology etc.
4. Language identification means that once an article or a document is received, it can be distinguished in which language it is written.
5. After seeing a text, it is possible to identify what the customer meant within it. It is called intent recognition. It is mainly applied in chatbots and virtual assistants.

The review focuses on the categorization of news headlines in Malayalam. Malayalam language has been declared as a classical language by Indian Govt. in 2013. In India, the regional languages are more popular in educational institutions, local people communicate with it and are more comfortable. Hence as the less work has been done in this area, we focus on Malayalam language. Opportunities also exist in this area as to convert the categorized headlines to voice. There is always a choice or selection possibilities based on the listener or viewer's interest. This type of choice is not available in regular news channels. There are already a number of surveys on text categorization that extensively incorporate methods from various text representation schemes, but leave out a number of additional approaches that have been studied in addition to the fundamental methods.

4.1.1 Rule-Based Classification Methods

In the context of classifying news headlines, rule-based classification methods refer to systems that classify headlines into various groups or classes using a set of manually created rules. This method is for constructing classifiers using logical combinations of elementary rules[56]. These guidelines are predicated on particular elements that have been taken from the text, like the existence of particular words, phrases, or syntactic constructions. A rule-based classifier consists of a set of "IF-THEN" rules obtained by statistically apprehending the training data[57]. C4.5 and Partial Decision Tree (PART) are very popular algorithms among them and both have many empirical features such as continuous number categorization, missing value handling, etc [58].Driven by the business advantage, people continuously lucubrate decision-tree algorithm to improve the precision of classification and for this purpose, a decision-tree rule classification algorithm is proposed[59]. Classification rules can be built using one of the two methods -direct method and indirect method[60].Acombination and pruning algorithm are applied to combine the two rule sets to generate a final classification rule set [61].In the context of news headlines, rule-based classification generally functions as follows:

1. Feature Identification: The first step for analysts is to determine which features could be important for classification. When it comes to news headlines, these could be particular nouns that denote persons or locations, verbs that speak to important occasions, or adjectives that express the article's tone.
2.Rule Creation: Experts create a set of guidelines that may direct the classification based on these features. For instance, a headline containing the word "election" may fall under the category "Politics," or it may fall under the category "Sports" if it contains the name of a sports team.
3.Pattern Matching: After that, the system examines incoming headlines and categorises them using these rules. To identify the relevant category, it looks for the predefined patterns or features in the text.
4. Making a Decision: A headline is placed in the appropriate category as specified by a rule if it meets the requirements outlined in that rule.
Among the benefits of rule-based systems are:

- **Transparency:** Since the classification criteria are derived directly from the guidelines established by the analysts, they are comprehensible and transparent.
- **Control:** Since they can modify the rules as necessary, the developers have complete control over the classification procedure.
- **Simplicity:** Since they don't require machine learning algorithms or training data, they can be implemented in a fairly simple manner.

Rule-based classification systems do, however, have certain drawbacks, particularly when working with natural language:

- **Maintenance:** As the language changes over time, the system may become difficult to maintain due to the growing complexity and number of rules.
- **Scalability:** Because every new scenario may call for a new rule, they don't scale well with the volume of data or the variety of languages.
- **Rigidity:** Although headlines that don't match the predefined patterns may still be semantically similar to them, rule-based systems are generally less flexible and may misclassify them.
- **Context Sensitivity:** They might struggle with language quirks like idioms, sarcasm, and context-dependent meanings, which can be especially difficult in languages like Malayalam that are rich in these elements.

Models that can learn from data have replaced rule-based systems utilising the advent ofartificial learning, in particular, deep learning. These demonstrations are often more adept at managing the intricacy and variability of natural language, and they are also capable of automatically recognising patterns. Nonetheless, rule-based techniques are still in use, particularly in hybrid systems that combine machine learning and rule-based techniques to best utilise their respective advantages.

4.1.2 Statistical Learning Approaches

Statistical learning techniques for classifying news headlines involve utilising statistical models to forecast a headline's category based on textual features that are extracted. These strategies are different from rule-based techniques in that they rely on patterns found in data rather than a predetermined set of rules. Statistical learning models use statistical properties to predict outcomes and deduce the underlying structure from the training dataset.The process of using statistical learning to categorise news headlines is as follows:

1. **Feature extraction:** Formatting headlines so that statistical models can process them is the first step. Typically, this means showing the text as a feature vector. Common strategies include Bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), and n-grams.
2. **Model Training:** Using a labelled dataset in which each headline's correct category is known, a statistical model is trained.By altering its settings to lessen the discrepancy between the categories it predicts and the

real ones, the model "learns." The algorithms that are employed could be support vector machines, logistic regression, or Naive Bayes.

3. **Pattern Recognition:** Using the statistical characteristics it has acquired, the trained model recognises patterns in fresh headlines that belong to a specific category. It might discover, for instance, that headlines containing particular word frequencies are probably related to sports news.
4. **Prediction:** The model utilises the learnt patterns to determine the most likely category when a new headline is given to it.
5. **Evaluation:** Performance metrics for the model include accuracy, precision, recall, and F1 score.. Usually, a different test set that wasn't used for training is used to validate the model.

Several benefits are provided by statistical learning approaches:

1. **Flexibility:** They don't require human assistance to automatically adjust to new patterns and trends in the data.
2. **Scalability:** Because they can manage massive data volumes, they are more scalable than rule-based systems.
3. **Generalisation:** Their ability to draw inferences from examples is superior, enabling them to generate reasonable predictions for headlines that deviate slightly from the training set.

But there are also restrictions:

1. **Data Dependency:** To achieve good results, these methods need a sizable and representative labelled dataset.
2. **Feature Selection:** Performance can be significantly affected by the features and representation used, and it's not always easy to identify the best features.
3. **Interpretability of the Model:** Certain statistical models, particularly the more sophisticated ones, can be challenging to interpret, which makes it challenging to comprehend the reasoning behind a given classification choice.

Machine learning in text classification was made possible by statistical learning techniques, and many of the ideas and methods that were developed in statistical learning are still essential to more sophisticated approaches, such as deep learning.

4.1.3 Machine Learning Techniques

In order to classify news headlines using machine learning techniques, algorithms are usually trained to find patterns in datasets and then utilise those patterns to predict the category of incoming headlines. Machine learning is capable of handling high-dimensional datasets and identifying sophisticated nonlinear relationships within the data, in contrast to rule-based or purely statistical approaches. There are typically multiple steps involved in classifying news headlines using machine learning:

1. **Data Preprocessing:** Cleaning the data, handling missing values, normalising the text, tokenizing (dividing the text into discrete words or phrases), and stemming or lemmatizing (reducing words to their base or root form) are all examples of data preprocessing.
2. **Feature Engineering**: To represent the headlines in a format that machine learning models can understand, features are then taken out of the preprocessed text. Word embeddings (such as Word2Vec or GloVe), TF-IDF vectors, bag-of-words models, or more complex representations like sentence embeddings could all be examples of this.
3. **Model Selection:** Machine learning models that can be used for classification include Support Vector Machines (SVM), Decision Trees, Random Forests, Gradient Boosting Machines, and Neural Networks. Every model has advantages, and the qualities of the dataset and problem can be taken into consideration when selecting a model.

4. **Model Training:** The chosen model is trained using a labelled dataset. The model gains ability to link headlines' characteristics to the appropriate categories. This is giving the model the target (the right classification) and the features (input data) in supervised learning.
5. **Cross-validation:** During training, the model's ability to generalise to a different dataset is evaluated using cross-validation techniques like k-fold cross-validation.
6. Model Tuning: To maximise the performance of the model, hyperparameter tuning is carried out. To determine the ideal set of parameters, methods like random search, grid search, or Bayesian optimisation can be employed.
7. **Evaluation:** The model is evaluated using metrics such as accuracy, precision, recall, and F1-score on a test set.. Confusion matrices can also be used to assess how well each category is being classified by the model.
8. **Prediction:** After being trained and adjusted, the model can use its newfound knowledge to categorise previously unseen headlines.

Machine learning techniques offer a number of advantages:

1. Adaptability: If they are retrained using updated data, they can adjust to new data without the need for human assistance.

2. Complex Pattern Recognition: Deep learning models in particular, which are more sophisticated than simpler models, are able to recognise complex patterns and interactions between features that humans or other models may miss.

3. Automation: A trained model can swiftly and consistently classify a large number of headlines.

But there are also difficulties:

1. Labelled Data Requirement: A significant amount of labelled training data is needed for supervised learning.

2. Overfitting: When a model gets too complicated and performs well on training data but poorly on unknown data, it is said to be overfitted.

3. Interpretability: Deep learning models in particular can operate as "black boxes," making it challenging to comprehend how they make decisions.

4. Computational Resources: The training and operation of certain machine learning techniques, particularly deep learning, call for a substantial amount of computational power and resources.

4.1.3 Deep Learning Models

To automatically extract features and representations from the textual data, deep learning approaches to news headline classification use layered neural network architectures. This allows the model to make predictions or categorise text without the need for explicit rule-based instructions. The complicated structure and minute details of natural language found in news headlines are well-suited for these techniques.

An outline of deep learning's applications in news headline classification is provided below:

1. Word and Sentence Embeddings: Using strategies like contextual embeddings (BERT, GPT) or word embeddings (Word2Vec, GloVe), the first step typically entails converting words and sentences into numerical representations. The semantic meaning and context that these embeddings capture are essential for comprehending headlines.

2. Neural Network Architectures: Text classification tasks use a variety of neural network architectures. While Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks are examples of Recurrent Neural Networks (RNNs), are particularly good at capturing information across longer sequences, which can be helpful for understanding the structure and context of headlines, Convolutional Neural Networks (CNNs) are used to identify patterns in text sequences.

3. Sequence Models: Complex models such as the Transformer, on which architectures such as BERT, RoBERTa, and T5 are based, use mechanisms such as self-esteem formeasure the proportional value for each word in a sentence. With headline data, these models can be refined to yield cutting-edge classification outcomes.

4. Training and Fine-Tuning: In deep learning, a model is usually pre-trained on a large corpus of text.

5. Transfer Learning: Transfer learning is The practise of altering a model that has been trained in a variety of tasks that is similar. For instance, a model that has already been trained to understand general language can be adjusted specifically for the job of dividing news headlines into groups such as entertainment, politics, or sports.

6. Evaluation: Recall, accuracy, precision, and F1 score are among the metrics used to evaluate deep learning models, just like they are with other machine learning techniques. To make sure a model works well with data that hasn't been seen yet, it's usually tested on a held-out dataset.

7. Inference: The deep learning model can rapidly classify new headlines using the learned representations after it has been trained.

The following are some benefits of deep learning methods:

- They are capable of handling delicate feature interactions and very large datasets.
- They have the ability to extract and apply sequential and contextual information from text, which is necessary to comprehend the nuanced language used in headlines.
- When it comes to natural language tasks, they frequently perform better than conventional machine learning models.

  But they also encounter challenges:
- A significant amount of labelled instruction setsis needed for deep learning models.
- They could be "black boxes," offering little information about the decision-making process.
- Deep learning model training can be a laborious and computationally demanding process.
- Deep learning models must be skilled at comprehending contextual nuances and linguistic quirks in order to classify news headlines in a variety of languages, including those with complex structures like Malayalam. This may require creating or modifying models especially for those languages.

4.1.4 Text Classification In Malayalam

Table 1: Scholarly literature review on text classification in Malayalam

S.NO.	Area and Focus of the Research	Outcome of the Research	Reference
1.	The study focuses on sentiment analysis, specifically using a rule-based approach to classify Malayalam words into positive, negative, and ambiguous categories.	It demonstrated the effectiveness of the rule-based approach in identifying the polarity of words within Malayalam documents. It displayed the quantity of both positive and negative terms found in the documents that were examined.	Jose, L.(2023).[5]
2.	The study aims to address the challenges of text classification for Dravidian languages, focusing on a multilingual framework due to the scarcity of publicly available resources and the difficulty of combining multiple Dravidian languages in NLP	The study developed a multilingual text classification framework that significantly improved the performance of multilingual text classification tasks for Dravidian languages, addressing language-specific nuances and the correlation among different languages	Lin, X. et al. (2021).[6]

3.	The study focuses on the analysis of text classification in Indian languages, addressing the challenges posed by NLP in this context	According to the study, supervised learning techniques like Naive Bayes, SVM, and ANN are effective for text classification tasks in various Indian languages, including morphologically rich ones like Tamil and Kannada. It also noted the success of specific methods like the Vector Space Model and Artificial Neural Networks in Tamil language classification	Kaur, J. et al. (2015).[7]
4.	The study focuses on natural language inference in Malayalam, using Siamese networks to classify text hypothesis pairs into entailment and contradiction classes	The results indicate that Siamese networks with newer embedding models like XLM-R provide promising performance for entailment recognition in Malayalam. The study demonstrates adequate classification performance with these models	Renjit, S. et al. (2021).[8]
5.	This paper addresses the gap in multilingual Aspect-based sentiment analysis(ABSA) by developing a tool for text classification based on multilingual aspects that targets delicate and varied Indian languages such as Bengali, Hindi, Tamil, Malayalam, Urdu, Telugu, and Sinhalese.	Every feature in the embedding document are categorised to their appropriate classes (flowers, plants, animals, sports, politics, etc.) by the LSFECO customised reLU-BiLSTM architecture. The proposed technique performs better than traditional methods in terms of entropy, coverage, purity, processing time, accuracy, F1-score, recall, and precision.	Suresh Kumar, K. et al.(2022).[9]
6.	The paper describes the offensive language identification and troll meme classification tasks for Dravidian languages (Tamil, Malayalam, Kannada).	The team achieved significant results in offensive language identification having F1 scores of 0.75 (Tamil), 0.95 (Malayalam), and 0.71 (Kannada).	Ghanghor, N. et al. (2021).[10]
7.	The study focuses on identifying insulting phrases for under-resourced Dravidian languages—Tamil, Malayalam, and Kannada in social media.	It highlights the performance of various models, with F1-scores indicating significant success in the discovery of offensive language for these languages.	Bharathi Raja Chakravarthi. et al.(2021).[11]
8.	Using multilingual BERT and Text CNN for semantic extraction and text classification, the study aims to classify Dravidian languages from social media into categories like Not-Offensive, Off-Untargeted, Off-Target-Individual, etc.	The study highlights the effectiveness of combining multilingual BERT with Text CNN in knowing offensive language in Dravidian languages and focusing on the significance of learning social media data that is mixed with codes.	Chen, S. et al. (2021). [12]
9.	Focusing on the classification of code-mixed and code-switched (CMCS) text, particularly addressing	Results imply that adapter-based PMLM fine-tuning approaches perform as well as or better than	Rathnayake, H. et al. (2022).[13]

	the challenges in low-resourced languages such as Sinhala, with an exploration of adapter-based fine-tuning of pre-trained multilingual language models (PMLMs) for CMCS text classification	standard PMLM model fine-tuning. A unique Sinhala–English CMCS dataset that has been annotated for a variety of classification tasks is also introduced in this study.	
10.	This study focuses on the challenges of analyzing social media data, particularly in code-mixed contexts for languages using non-English scripts, with a focus on Malayalam and other low-resource languages	The research highlights the classification of code-mixed data and acknowledges the limited availability of datasets for such languages, noting specific datasets for Malayalam-English and other language pairs	Chathuranga, S. et al. (2021). [14]
11.	The paper focuses on the application of Support Vector Machines (SVM) for effective text classification, particularly for news filtering in the Malayalam language	The study demonstrates the efficiency of SVM in classifying Malayalam texts, highlighting its effectiveness in high-dimensional spaces and its versatility with different kernel functions. The system shows high accuracy in classifying unknown Malayalam text into predefined classes	Nibeesh, K. et al. (year does not found).[15]
12.	The study explores the effectiveness of multilingual joint training for a variety of NLP tasks, such as the detection of hate speech, the sentiment analysis of Dravidian languages in code-mixed text, and the event detection of news articles.	The study concludes that multilingual joint training offers significant advantages in processing complex, noisy datasets like multilingual social media conversations. It notes that this approach is more effective for multilingual datasets than for monolingual datasets from more controlled sources like newspapers.	Kumar, R. et al.(2020).[16]
13.	Investigating the classification of YouTube responses, which are mixed between English and Malayalam, with a focus on viewer feedback and sentiment.	The study found that Random Forest with Term Frequency Vectorizer was the best among traditional models with 63.59% accuracy. The paper also experimented with multilingual transformer-based models like BERT, showing XLM as the top-performing model with 67.31% accuracy	Kazhuparambil, S. et al. (2020). [17]
14.	The study, which focuses on the relationship information between entities in social media news text, particularly in Malayalam and other Indian languages, presents a statistically based classifier for text classification.	The study achieved high accuracy rates in categorizing relationships within text, particularly in the Telugu language dataset. Logistic regression was found to be the most effective classifier among the ones tested	Krishna, N. S. et al. (2018). [18]
15.	In order to create a personalised travel recommender system for Malayalam, the study will employ	The research demonstrated that the recommender system's accuracy and efficiency were greatly increased by the	Muneer, V. K. et al (2023). [19]

	content-based clustering and collaborative filtering techniques. It emphasises using social media travelogues and reviews as the main data source.	clustering technique. While the agglomerative hierarchical clustering approach achieved an accuracy of 85% and an F1 score of 84.25%, K-means clustering achieved an accuracy of 91% and an F1 score of 85%.	
20.	The study aims to adopt on the pervasive challenge of fraudulent news in the Malayalam language, particularly targeting misinformation spread on various online platforms	Specific findings or experimental results were not detailed in the sections provided, but the approach aims to address the growing need for specialized solutions in regional languages like Malayalam.	Sujan, A. S. et al. (2023). [20]
21.	The study uses a code-mixed dataset of posts on YouTube in Tamil, Malayalam, and Kannada to identify vulgar phrases in Dravidian languages.	With weighted F1 scores of 0.64 (Kannada), 0.95 (Malayalam), and 0.71 (Tamil) on the test dataset, the team's best method produced notable outcomes. This suggests that these languages are capable of identifying offensive language effectively.	Chakravarthi, B. R. et al. (2023). [21]
22.	In order to detect offensive content from multilingual code-mixed data on social media platforms, including languages like Malayalam, this paper focuses on developing an automated system.	While the paper does not explicitly detail its findings in the quoted sections, it emphasizes the significance of applying robust computational systems to recognise and restrict objectionable content on social media	Sharif, O., Hossain, E. et al. (2021). [22]
23.	In code-mixed or non-native script multilingual social media posts in Dravidian languages, such as Malayalam, Tamil, and Kannada, the paper seeks to identify words that are offensive.	Specific findings or key results were not detailed in the quoted sections. The paper emphasizes the growing importance of analyzing social media user-generated material for objectionable language, particularly in situations with multiple languages and mixed codes.	Yasaswini, K. et al. (2021). [23]

4.2 News Headlines Categorization

4.2.1 News Headlines Categorization in Various Languages

Lis Jose(2023) experimented the polarity classification of Malayalam Document- a Rule Based Approach[5]. It explains the reasons why the researchers follow Rule Based approach for the categorization in Malayalam over corpus based approach. The main reasons are the unavailability of large corpora, its cost.

The different algorithms has been used for the categorization of news headlines in various languages. Aysha Gazi Mouri. et al.(2022) submitted a paper about the empirical investigation using various machine learning techniques to categorise Bengali news headlines[25].Using NLP, the Bengali news headlines are divided into six different categories. This paper presented an analysis of data using deep learning models (LSTM, Bi-LSTM, GRU, Bi-GRU, CNN), transformer learning models (Bangla-BERT, XLM –RoBERTa), and machine learning algorithms (Logistic Regression, Random Forest Classifier, Multinomial Naïve Bayes, and RBF Support Vector Machine). After the evaluation XLM-RoBERTa gained a better accuracy of 86.5 %.

Amran Hossain. et al. (2021) introduced LSTM and GRU neural networks models to categorize Bangla News Headlines[26]. Among this models LSTM acquires less accuracy than GRU. For this research GRU used 64 units and LSTM used 128 bits. Therefore the GRU training performance is faster than that of LSTM.

Benjamin Chanakot. et al.(2022) conducted a research on Thai news headline classification with an artificial neural network[27]. They had collected around 1200 headlines and then classified into political news, sports news, economic news and crime news. Chi-square, information gain, and term frequency inverse class frequency (TFICF) are used to measure the distribution of headlines. Then the news headlines are classified using artificial neural networks.Ankita dhar. et al. portrays the text categorization in the past as well as present [28]. This literature review paper confirms that text categorization tasks has been done in different languages like English, Arabic, and Chinese etc by different authors. It is also comments that very few work has been done for Indian Languages. According to this survey, The hierarchy of text categorization algorithms includes deep learning, conventional, and fuzzy logic based algorithms. The conventional algorithms includes Handcrafted, Nature inspired and Graph based. It donates the information about the various libraries like NLTK, Fuzzy Logic toolbox, Numpy, Scipy, Scikit-learn, Theano, TensorFlow, Keras, PyTorch, Pandas, TextBlob supports the various algorithms. Ettilla Mohiuddin. et al. presented a paper on Bengali News Headlines: Multilevel Categorization Using Bidirectional Gated Recurrent Unit[29]. It is categorized into six classes of news Headlines.With training data, the suggested model achieved an accuracy of 97%, and with validation data, it achieved 84%. Madhus Smitha. et al.(2023)first classified Indian news headlines using the LSTM model and word members technique[30].Headlines are usually a very short sentence so that the reader can understand exactly what it is about after reading this sentence.The models utilised to categorise news headlines into distinct groups include word embedding, cosine similarity index, Bidirectional Encoder Representations from Transformers (BERT), and Long Short Term Memory (LSTM) networks. In this work it was possible to label unlabelled data as wellmaking use of the BERT sentence encoder. Rizwana Kallooravi Thandil. et al.(2021) K published a paper titled as NLP of Malayalam text for predicting its authenticity[31].It talks about how to create a customised data scraping tool that can retrieve Malayalam-language text from facebook.com. This paper conveys a novel method for determining the authenticity of news articles and text written in English. News prediction is accomplished through the use of methods such as TF-IDF, Bag of Words, and NLP. This paper mentions about the preprocessing tools for Malayalam language are NLTK, rootpack and Tokenization classes. Some of the research works be regarded as sentiment analysis is also applicable for news headlines classification. It has been experienced for Telugu news headlines by Jalaja Kumari Dygani. et al. [32]. Ijaz. et al.(2009) designed a comparison of statistical methods using naïve Bayes and support vector machines (SVM) for Urdu text categorization (NB)[33]. Recently the morophological analyzers are added for the Korean language news articles' topic classification [34]. Korean is an agglutinative language[35] like Malayalam, tamil etc. For the morphologically rich languages morphological analysis is an important step as part of NLP. Fatima Jahara. et al.(2022) introduced multilayer perceptron for the automatic categorization of news articles and headlines[36]. Vukyam Sri Sravya. et al. represents the the telugu news headlines classification with the help of constructing personalised Deep Learning and Machine Learning models using count and prediction based word embeddings[37]. Rameesa, K. et al.(2021) came up with a new idea that Malayalam news headlines categorization is possible through sentiment analysis[38].Jisha, P. et al.text classification of Malayalam documents which is web based has been done[39].Parvathavarthini, S. et al.announced that Bi-directional Encoding Representational Transformers (BERT), also known as NLP transformers, can be apply as news headlines categorization tool[40]. It is clear from this study that there has been no process of categorization of news headlines in Malayalam so far.

4.2.2 News Headlines Categorization in Malayalam

Table 2: Scholarly literature review on news headlines categorization in Malayalam

S.NO.	Area and Focus of the Research	Outcome of the Research	Reference
1.	Explores the news programming practices in Malayalam newspapers and how these align with readers' content preferences	Important discoveries would include perceptions into the readership preferences and news programming ways of Malayalam newspapers. These findings would shed light on the alignment or gaps between what is presented in the newspapers and what readers prefer.	Harikumar, M. S. (2008). [41]
2.	The study reviews various machine translation systems developed for Indian languages, including systems for Malayalam-English and English-Malayalam translation	successful use of monolingual and bilingual corpora for translation, application of order conversion rules to address structural differences, and the employment of AI techniques for context disambiguation and compound word splitting in Malayalam-English translation	Godase, A. et al. (2015). [42]
3.	The paper aims to review and summarize 60 key published papers related to automaticallyidentifying falsified information on social media, focusing on the developments from 2011 to 2022. This includes examining key models, techniques, and challenges in the field of fake news detection on social media	a deep dive into the complex and evolving field of fake news detection on social media highlights the progress made in the last decade, outlines the current state of research, and identifies key areas for future exploration.	Manish, M. K. S. et al. (2022). [43]
4.	The study aims to automate the evaluation of English essays, traditionally a manual and challenging task, by leveraging natural language processing techniques	The study showcases high accuracy and efficiency in grading essays, with specific measures like word embedding and sequence analysis contributing to the grading system's effectiveness.	Rajest, S. S. et al. (2023). [44]
5.	Developing an abstractive summarization system for Malayalam documents. Addressing the challenges posed by the agglutinative and morphologically rich nature of Malayalam.	successful development of an abstractive summarization system specifically for Malayalam language documents particularly in the context of cricket.	Sunitha, C.et al. (2019). [45]
6.	It aims to create a system that can efficiently summarize information from multiple documents, particularly for Malayalam news documents. The	The performance was evaluated using intrinsic evaluation methods by comparing the system-generated summaries with human-	Manju, K.(2017).[46]

	study utilizes a sentence scoring technique and incorporates an online Malayalam Wordnet for semantic similarity checking.	generated reference summaries. Metrics like precision and recall were used to assess the effectiveness of the summarization system, demonstrating its capability in multi-document summarization for the Malayalam language	
7.	The research investigates the syntactic-semantic role of touch in ritual structures, drawing an analogy between language and ritual. The study is grounded in the cognitive approach to ritual competence and communication.	The research demonstrates the viability of the proposed analytical framework by categorizing touching events in shared festivals.	Gamliel, O. (2019). [47]
8.	The research is centered on developing a shallow parser for Malayalam to facilitate machine translation, particularly between Malayalam and Tamil.	The work provides valuable resources for students and professionals in NLP.	Sankaravelayuthan, R. et al.2019). [48]
9.	This research focuses on applying natural language processing techniques to categorize news stories into broad topic categories.	The results of the system were very promising, achieving an average recall rate of 93% and an average precision rate of 93%. This means that the system correctly made 93% of the topic assignments and had only 7% false negatives and false positives in its assignments. The categorization was achieved in an average time of around 15 seconds per story	Hayes, P. J. et al. (1988). [49]
10.	The study is focused on real-time classification of news headlines into predefined categories using various machine learning algorithms	Several machine learning models, such as Multinomial Naïve Bayes, Logistic Regression, Support Vector Machine, and Neural Networks, were compared and assessed in the study. Three metrics were used to analyse the performance: recall, precision, and F1 score.	Chhajerh, M.S. et al.(2021).[50]
11.	The research focuses on generating emojis from Malayalam language news headlines using machine learning techniques. Specifically, Naive Bayes (NB) and Support Vector Machine (SVM) classifiers are used.	Experimental results showed that the SVM model, especially when combined with TF-IDF features, outperformed the NB model, achieving an accuracy of 74.3%	Soumya S. et al. (2019). [51]
12.	This study compares the efficacy of	machine learning models,	Chavan, S.S. (2018).

	Semantic Oriented Approach (SOA) and various machine learning techniques for sentiment classification of news headlines.	especially SVM, are more effective in sentiment classification of news headlines than the SOA based model.	[52]
13.	The study aims to explore and classify multiword expressions in Malayalam based on three idiosyncrasies: semantic, syntactic, and statistic idiosyncrasies. It addresses the less-studied area of MWEs in Malayalam, focusing on their classification and features.	The study concludes that understanding and classifying MWEs in Malayalam is crucial for effective language processing and machine translation. It emphasizes the need for more linguistic analysis in this area, considering the unique properties of Malayalam and the prevalence of code-mixed expressions	Cyriac, T. et al. (2022). [53]
14.	This study explores the use of SVM for effective text classification in the Malayalam language, particularly for news filtering applications.	The findings emphasize the effectiveness of SVM in handling high-dimensional spaces and its versatility in various classification tasks.	Nibeesh, K. et al.[54]
15.	The study focuses on the automatic categorization of news articles into specific groups using the k-mean clustering algorithm and TF-IDF technique.	It concludes that the combination of k-mean clustering and TF-IDF is effective for automatic text categorization of news articles, suggesting its potential for broader applications in news filtering and digital content management.	Sandhya, M. et al. (2016). [55].

5. Current Status & New Related Issues :

Based on the review conducted, news headlines classification is most important for every newspaper in Malayalam. Different machine learning algorithms and approaches are studied. By using a hybrid model of these algorithms, the accuracy of classification can be improved.

- Decision-making by considering various data sources has to be explored using various approaches.
- Various contextual factors related to news headlines and categorization have to be studied.
- Advanced machine learning and deep learning approaches can be studied to get more accurate classification results.
- The lack of rich annotated datasets in Malayalam

6. Ideal Solution, Desired Status & Improvements Required: (Based on Current Status)

(1) How the accuracy can be improves with larger data incorporated
(2) Hybrid categorization by decision based on accuracy
(3) Find out what are the modifications can be done in already existing methods
(4) Compare and select the model based on the type of news headlines
(5) The creation and training of more precise models is complicated by the known lack of rich, annotated datasets in Malayalam.

7. Research Gap :

(1) The lack of extensive, annotated datasets dedicated to Malayalam news headlines is one of the major gaps. This constraint interferes with the advancement of increasingly complex models that are capable of absorbing linguistic complications and accurately classifying information.

(2) Code-mixed content, which combines English and other regional languages, is frequently seen in Malayalam. The difficulties presented by such mixed-language content, which is becoming more and more common in digital news platforms, have not been sufficiently addressed by current research.

(3) Although real-time categorization has advanced, it is still difficult for these systems to scale up to meet the growing amount and diversity of news content. There is currently little research on scalable and effective real-time processing models for Malayalam news.

(4) The primary focus of current models is text syntactic processing. Research on more in-depth semantic learning and contextual analysis is lacking, which is important for precise classification—especially in a language like Malayalam that is full of idioms and cultural references.

(5) Although sentiment analysis of Malayalam news headlines has been investigated, more work needs to be done to determine how accurately sarcasm and other subtle emotional aspects can be detected.

8. Research Agendas Based On Research Gap :

(1) Tocreate and maintain broad, annotated datasets with a variety of categories and sources for Malayalam news headlines.

(2) Tocarry out comprehensive research on the creation of algorithms that can effectively handle code-mixed news content in Malayalam.

(3) Todevelop unique, scalable models that can effectively handle massive amounts of data in real-time Malayalam news headline processing.

(4) Toconcentrate on creating sophisticated NLP models with a concentration on Malayalam semantic comprehension and contextual analysis.

(5) Toimprove Malayalam news headline sentiment analysis methods, with a particular focus on sarcasm and subtle emotion detection.

9. Analysis of Research Agendas:

(1) The rich annotated datasets will greatly increase machine learning model training and accuracy. It enables a more sophisticated comprehension of the language, encompassing its colloquial expressions and cultural minute details.

(2) Since code-mixing is common in Malayalam digital content, handling code-mixed content successfully will increase the models' applicability in real-world situations.

(3) For the purpose of promptly classifying and distributing news, real-time processing capabilities are essential for improving user experience and information accessibility.

(4) Improving contextual and semantic analysis will result in a more thorough and precise comprehension of news content, which is necessary for sentiment analysis and efficient classification.

(5) For a variety of stakeholders, improved sentiment analysis will provide an improved grasp of public perception and emotions expressed in news headlines.

All of these agendas are meant to make a major contribution to the field of Malayalam news classification. Every agenda, however, also offers distinct difficulties that call for diverse methods, substantial computer resources, and cutting-edge research techniques. If these agendas are successfully implemented, the Malayalam-speaking community and the field of natural language processing (NLP) as a whole will benefit greatly from more effective, efficient, and contextually rich information processing systems.

10. Research Proposal

Following a study and evaluation of the available research literature review, the article suggests conducting mega-research on proposing news headlines classification in Malayalam with high accuracy models towards newspaper companies.

10.1 Proposed title: Build a large dataset with a variety of Malayalam news items that have been annotated for different categories and emotions and introduce hybrid categorization to improve the performance.

10.2 Purpose: Effective machine learning model training requires a large, annotated dataset. The dataset's diversity and comprehensiveness determine how well the model learns and comprehends the complex elements of the Malayalam language. As a result, the classification and analysis of news content is more accurate. Hybrid categorization helps to get high-accuracy classification models for better decision-making based on accuracy.

11. Suggestions to Implement Research Activities :

(1) Dataset development and annotation.
(2) Hybrid model development
(3) Technical infrastructure and tools
(4) Collaborations and partnerships
(5) Testing and validations
(6) Ethical legal considerations
(7) Documentation and publication
(8) Limitations and future work

12. Conclusion:

The literature review concludes by presenting a picture of a dynamic field of study with lots of room for impact and innovation. If the proposed research activities are implemented successfully, Malayalam news categorization could undergo a significant change, becoming more precise, effective, and culturally sensitive. In addition to helping the Malayalam-speaking community, this will provide insightful new knowledge in the international fields of information management and language processing.

References:

1. Nguyen, D., & Hekman, E. (2022). The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation. *AI & SOCIETY*, 1-15.
2. Watson, H., & Thompson, L. F. (2021). News in the age of algorithmic personalization: Analyzing the impact of news content classification on media practices and public discourse. *Journal of Communication Technology*, 4(2), 55-75.
3. Zhou, N., Yao, N., Zhao, J., & Zhang, Y. (2022). Rule-based adversarial sample generation for text classification. *Neural Computing and Applications*, *34*(13), 10575-10586.
4. Jha, A., & Patil, H. Y. (2023). A review of machine transliteration, translation, evaluation metrics and datasets in Indian Languages. *Multimedia Tools and Applications*, *82*(15), 23509-23540.
5. Lin, X., Lin, N., Wattanachote, K., Jiang, S., & Wang, L. (2021). Multilingual text classification for dravidian languages. *arXiv preprint arXiv:2112.01705*.
6. Kaur, J., & Saini, J. R. (2015). A study of text classification natural language processing algorithms for Indian languages. *VNSGU J Sci Technol*, *4*(1), 162-167.

7. Renjit, S., & Idicula, S. M. (2021, September). Siamese networks for inference in Malayalam language texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1167-1173).
8. Suresh Kumar, K., & Helen Sulochana, C. (2022). Local search five-element cycle optimized reLU-BiLSTM for multilingual aspect-based text classification. *Concurrency and Computation: Practice and Experience*, *34*(28), e7374.
9. Ghanghor, N., Krishnamurthy, P., Thavareesan, S., Priyadharshini, R., & Chakravarthi, B. R. (2021, April). IIITK@ DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages* (pp. 222-229).
10. Chakravarthi, B. R., Priyadharshini, R., Jose, N., Mandl, T., Kumaresan, P. K., Ponnusamy, R., ... & Sherly, E. (2021, April). Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages* (pp. 133-145).
11. Chen, S., & Kong, B. (2021, April). cs@ DravidianLangTech-EACL2021: Offensive language identification based on multilingual BERT model. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 230-235).
12. Rathnayake, H., Sumanapala, J., Rukshani, R., & Ranathunga, S. (2022). Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. *Knowledge and Information Systems*, *64*(7), 1937-1966.
13. Chathuranga, S., & Ranathunga, S. (2021, September). Classification of code-mixed text using capsule networks. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2021)* (pp. 256-263).
14. Nibeesh, K., Sreejith, C., & Raj, P. R. Malayalam Text Classification for Efficient News Filtering using Support Vector Machine.
15. Kumar, R., Lahiri, B., Ojha, A. K., & Bansal, A. (2020, December). ComMA@ FIRE 2020: Exploring Multilingual Joint Training across different Classification Tasks. In *FIRE (Working Notes)* (pp. 823-828).
16. Kazhuparambil, S., & Kaushik, A. (2020). Cooking is all about people: Comment classification on cookery channels using bert and classification models (malayalam-english mix-code). *arXiv preprint arXiv:2007.04249*.
17. Krishna, N. S., Bhattu, S. N., & Somayajulu, D. V. (2018). idrbt-team-a@ IECSIL-FIRE-2018: Relation Categorization for Social Media News Text. In *FIRE (Working Notes)* (pp. 166-173).
18. Muneer, V. K. (2023). A Comparative Study of Collaborative Filtering and Content-Based Approaches for Improving the Accuracy of Travel Recommender Systems for Malayalam Language. *International Journal of Advanced Networking and Applications*, *14*(6), 5717-5721.
19. Sujan, A. S., Benny, A., & Anoop, V. S. (2023). MalFake: A Multimodal Fake News Identification for Malayalam using Recurrent Neural Networks and VGG-16. *arXiv preprint arXiv:2310.18263*.
20. Chakravarthi, B. R., Priyadharshini, R., Jose, N., Mandl, T., Kumaresan, P. K., Ponnusamy, R., ... & Sherly, E. (2021, April). Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages* (pp. 133-145).
21. Sharif, O., Hossain, E., & Hoque, M. M. (2021). Nlp-cuet@ dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. *arXiv preprint arXiv:2103.00455.*
22. Yasaswini, K., Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., & Chakravarthi, B. R. (2021, April). IIITT@ DravidianLangTech-EACL2021: Transfer learning for offensive language

detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 187-194).

23. Zareapoor, M., & Seeja, K. R. (2015). Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business*, *7*(2), 60.
24. Mouri, A. G., Talukder, P., Anik, T. R., Rahman, I. S. I., Joy, S. K. S., Shawon, M. T. R., ... & Mandal, N. C. (2022, December). An Empirical Study on Bengali News Headline Categorization Leveraging Different Machine Learning Techniques. In *2022 25th International Conference on Computer and Information Technology (ICCIT)* (pp. 312-317). IEEE.
25. Hossain, E., Chaudhary, N., Rifad, Z. H., & Hossain, B. M. (2020). Bangla-news-headlines-categorization. *GitHub*.
26. Chanakot, B., & Sanrach, C. (2023). Classifying thai news headlines using an artificial neural network. *Bulletin of Electrical Engineering and Informatics*, *12*(1), 395-402.
27. Dhar, A., Mukherjee, H., Dash, N. S., & Roy, K. (2021). Text categorization: past and present. *Artificial Intelligence Review*, *54*, 3007-3054.
28. Mohiuddin, E., & Matin, A. (2021, July). Multilevel Categorization of Bengali News Headlines using Bidirectional Gated Recurrent Unit. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)* (pp. 1-6). IEEE.
29. Khuntia, M., & Gupta, D. (2023). Indian News Headlines Classification using Word Embedding Techniques and LSTM Model. *Procedia Computer Science*, *218*, 899-907.
30. Thandil, R. K., Basheer KP, M., & VK, M. (2021). Natural Language Processing of Malayalam Text for predicting its Authenticity. *Proceedings of the Yukthi*.
31. Kumari Bygani, J., Venkateshwaralu, Y., & Ramana, K. V. (2023). A Sentence Level Classification of Telugu News Document using Sentiment Analysis. In *E3S Web of Conferences* (Vol. 391, p. 01037). EDP Sciences.
32. Ali, A. R., & Ijaz, M. (2009, December). Urdu text classification. In *Proceedings of the 7th international conference on frontiers of information technology* (pp. 1-7).
33. Ahn, S. (2023). Experimental Study of Morphological Analyzers for Topic Categorization in News Articles. *Applied Sciences*, *13*(19), 10572.
34. Kumar, A., Padró, L., & Oliver, A. (2015, September). Learning agglutinative morphology of Indian languages with linguistically motivated adaptor grammars. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 307-312).
35. Jahara, F., Sharif, O., & Hoque, M. M. (2022). Automatic Categorization of News Articles and Headlines Using Multi-layer Perceptron. In *Intelligent Computing & Optimization: Proceedings of the 4th International Conference on Intelligent Computing and Optimization 2021 (ICO2021) 3* (pp. 155-166). Springer International Publishing.
36. Sravya, V. S., Kumar, S., & Soman, K. P. (2022, June). Text Categorization of Telugu News Headlines. In *2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE.
37. Rameesa, K., & Veeramanju, K. T. (2022, November). Sentiment Analysis for Headlines Categorization in Newspaper Industry Malayala Manorama Company Limited. In *2022 1st International Conference on Computational Science and Technology (ICCST)* (pp. 32-41). IEEE.
38. Jayan, J. P., & Govindaru, V. (2022, August). Automatic Text Classification for Web-Based Malayalam Documents. In *International Symposium on Intelligent Informatics* (pp. 187-200). Singapore: Springer Nature Singapore.
39. Parvathavarthini, S., Shreekanth, M., & Santhosh, N. S. (2023, August). News Category Classification using Natural Language Processing Transformer. In *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)* (pp. 1185-1189). IEEE.

40. Harikumar, M. S. (2008). *News programming practices and readers' content preferences: A study of Malayalam newspapers* (Doctoral dissertation, University of Calicut).
41. Godase, A., & Govilkar, S. (2015). Machine translation development for Indian languages and its approaches. *International Journal on Natural Language Computing*, *4*(2), 55-74.
42. Manish, M. K. S., Jawed, J. A., Alam, M. A. A., Kamlesh, K. K. R., & Sachin, S. K. (2022). A Comparative Study of Computational Fake News Detection on Social Media.
43. Rajest, S. S., Regin, R., & Shynu, T. (2023). A New Natural Language Processing-Based Essay Grading Algorithm.
44. Sunitha, C., Jaya, A., & Ganesh, A. (2019). Automatic summarization of Malayalam documents using clause identification method. *International Journal of Electrical and Computer Engineering*, *9*(6), 4929.
45. Manju, K. (2017, March). An extractive multi-document summarization system for Malayalam news documents. In *First EAI International Conference on Computer Science and Engineering* (pp. 218-227).
46. Gamliel, O. (2019). The syntactic roles of touch in shared festivals in Kerala: towards an analysis of ritual categories. *Entangled Religions*, *10*.
47. Sankaravelayuthan, R., & Krishnakumar, K. (2019). A Comprehensive Study of Shallow Parsing and Machine Translation in Malaylam. *Coimbatore: Amrita Vishwa Vidyapeetham, Coimbatore*.
48. Hayes, P. J., Knecht, L. E., & Cellio, M. J. (1988, February). A news story categorization system. In *Second Conference on Applied Natural Language Processing* (pp. 9-17).
49. Chhajerh, M.S., Kvs, A., Meleet, P.M., & Murthy, D.R. (2021). Real Time News Headlines Classification Using Machine Learning.
50. Soumya, S.,Pramod, K. V. (2019). Prediction of Emoji from News Headlines using Machine Learning Techniques. *International Journal of Recent Technology and Engineering*.
51. Chavan, S.S. (2018). Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning.
52. Cyriac, T., & Lalitha Devi, S. (2022). Classification of Multiword Expressions in Malayalam. *WILDRE*.
53. Nibeesh, K., Sreejith, C., & Raj, P. R. Malayalam Text Classification for Efficient News Filtering using Support Vector Machine.
54. Sandhya, M., Sarika, S., Anukriti, S., & Sushila, A. (2016). Automatic Text Categorization on News Articles. *International Journal of Engineering and Techniques*, *2*(3), 33-38.
55. Michailidis, G., & Shedden, K. (2003). The Application of Rule-Based Methods to Class Prediction Problems in Genomics. *Journal of computational biology : a journal of computational molecular cell biology, 10 5*, 689-98 .
56. Hossain, S.K., Ema, S.A., & Sohn, H. (2022). Rule-Based Classification Based on Ant Colony Optimization: A Comprehensive Review. *Appl. Comput. Intell. Soft Comput., 2022*, 2232000:1-2232000:17.
57. Mazid, M.M., Ali, A.B., & Tickle, K. (2010). Input space reduction for rule based classification. *WSEAS Transactions on Information Science and Applications archive, 7*, 749-759.
58. Jian-hua, W. (2010). Classification algorithm of rule based on decision-tree. *Computer Engineering and Design*.
59. Datta, R.P., & Saha, S. (2016). Applying rule-based classification techniques to medical databases: an empirical study.
60. Yuan, Y., Shen, J., & Song, Q. (2003). A new rule-based video classification approach. *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693), 1*, 225-230 Vol.1