Extraction of Textual Data from Unstructured Malayalam Web Resources

Jisha P.Jayan[1], **Anju Vinod[2]** , **Suresh K.S.[3]** and **Jayaraj N.[4]**
Research & Development Division
Centre for Development of Imaging Technology, Thiruvananthapuram
[1]jishapjayan@cdit.org, [2]anjuvinod@cdit.org, [3]sureshks@cdit.org, [4]jayaraj@cdit.org

Abstract:
The exponential growth of digital content in Malayalam language has led to an urgent need for advanced methods to extract valuable insights from this vast corpus. This paper presents a comprehensive study of text mining techniques tailored to Malayalam, aiming to bridge the gap between linguistic intricacies and computational approaches. This work describes a pioneering investigation into the creation of an integrated Malayalam Text Mining Tool combined with a comprehensive Part-of-Speech (POS) tagger. To handle the Malayalam language's distinctive linguistic characteristics, the research combines Natural Language Processing and Machine Learning concepts. The text mining tool is meant to rapidly extract relevant insights from Malayalam text, meeting the growing demand for language-centric technologies in a variety of applications. Concurrently, the POS tagger improves the tool's capabilities by precisely recognizing and labelling parts of speech in Malayalam phrases, enhancing the analysis.This tool serves as a facilitator for gathering corpora from diverse newspapers, employing data processing capabilities in TXT and CSV file formats, which are indispensable for a multitude of Natural Language Processing applications. Techniques such as tokenization, stemming, and lemmatization are employed to standardize word representations. Feature extraction methods like TF-IDF and word embeddings capture semantic relationships and local patterns, which enhance text comprehension for further analysis and machine learning. The analyzed data then undergoes classification to extract valuable insights. Model performance is assessed using evaluation metrics, while visualization techniques are employed to present results comprehensively for interpretation and communication. In future, further exploration could involve integrating additional machine learning algorithms for comparative analysis, thus paving the way for a deeper understanding and more advanced applications of Malayalam text mining across various domains.

Keywords:
Text Mining, Information Extraction, Unstructured Text, BeautifulSoup, Conditional Random Field (CRF)

1 INTRODUCTION

Text mining (TM) is the process of transforming unstructured text into meaningful and actionable information. TM, also known as text data mining or text analytics, addresses the challenge of extracting valuable patterns and trends from vast collections of text documents. There are numerous methods and resources available for mining text documents to extract important data [1]. Selecting the appropriate text mining technique is crucial for enhancing the speed and efficiency of retrieving valuable information. The availability of digital data continues to increase, with a substantial portion existing in unstructured textual form. This has led to the emergence of information extraction and text mining as popular research areas dedicated to uncovering valuable information from textual data. This can be applied in various areas, including web mining and merging with traditional data mining processes. Feature selection is an important aspect of text mining, involving the process of selecting a subset of important features for model creation. By identifying topics, patterns, and relevant keywords, text mining allows one to obtain valuable insights without needing to go through all the data manually. TM combines notions of statistics, computational linguistics, information retrieval, data mining, and machine learning to create models that learn from training data and can predict the results of new information based on their previous experience. Hence, it is nothing short of a multidisciplinary field. It deals with natural language texts either stored in semi-structured or unstructured formats or identifies facts, relationships, and assertions that would otherwise remain buried in the mass of textual big data. These facts are extracted and turned into structured data, for analysis, visualization, integration with structured data in databases or warehouses, and further refinement using machine learning systems. Document categorization or classification, information retrieval, document clustering, information extraction, and performance assessment are a few applications of text mining.

TM is a knowledge discovery process used to extract interesting and non-trivial patterns from natural language [2]. With the advent of big data platforms and deep learning algorithms that can evaluate large amounts of unstructured data, TM has become increasingly useful for users, including data scientists. Text mining and analysis can be used by enterprises to extract potentially useful business insights from a variety of text-based data sources, including corporate papers, customer emails, call center logs, verbatim survey answers, social media posts, and medical information.

TM skills are also being added to Artificial Intelligence (AI) chatbots and virtual agents, which businesses use to automatically respond to consumers as part of their marketing, sales, and customer support processes. TM classifies each document according to its primary topic, intent, and sentiment to make sense of a large amount of unstructured text. It analyses unstructured material using Natural Language Processing (NLP) techniques and then classifies the documents using AI techniques like Machine Learning (ML). Patterns and relationships that would otherwise remain hidden in the text are revealed through this method. It is also possible for machine learning algorithms to generate models that forecast novel behaviours and trends. Up to 80% of commercial data is thought to be composed of unstructured data, like text. Analysing all of this data might be very time-consuming or even impossible without an automated method. Moreover, information produced by automatically processing text documents can be more precise and reliable. Text mining can assist companies in anticipating competitive threats, responding more swiftly to manufacturing or customer service issues, and offering more individualized customer care.

Many different approaches are used in text mining, and they are important. The methods are not all the same. Although the information extraction technique extracts information from organized databases, the retrieval technique uses unstructured text to obtain valuable information. The document is summarized using the summarization approach, which shortens its length while maintaining its significance. The method of categorization is controlled and employs a predetermined set of documents based on their content. In contrast, clustering is used to identify underlying structures in data and group related data into subgroups for additional research and analysis.

With its many advantages, text mining is transforming how businesses gather and use data from enormous volumes of unstructured text. Discovering insightful patterns and insights in data to support well-informed decision-making is one of the main benefits. By examining consumer attitudes, market developments, and competition, businesses can obtain a competitive advantage. Understanding consumer input and making necessary adjustments to products and services leads to better customer experiences. By collecting pertinent information from medical literature, text mining helps in drug discovery and speeds up research in the healthcare industry. Effective information retrieval, improved search features, and anomaly detection help avoid fraud across a range of businesses. TM helps language translation programs by providing more precise and context-aware translations.

Despite its strength, TM has several drawbacks due to the complexity of natural language and variety of textual data. Ambiguity is a significant barrier since words can have several meanings and interpretations, which makes it difficult to understand context clearly. Text mining algorithms also face challenges due to the dynamic nature of language, which includes slang, acronyms, and changing terminology. Handling unstructured data can provide some difficulties, including uneven language usage and different document formats. Text mining algorithms may have trouble processing negations and comprehending context, which could result in misunderstandings. Moreover, the enormous amount of textual data necessitates complex computer power, and the requirement for annotated training data can be a bottleneck, particularly in specialized fields. Complications arise from addressing privacy issues and ensuring proper use of textual data.

This paper is organized into several sections. The introduction portion is covered in the first section. Major works in this field of research are reviewed in the second section. The third section deals with the need for TM for Indian languages with importance to Malayalam. The fourth section covers the proposed method and implementation. The next section deals with the results and discussions. The sixth section deals with the limitation and future scope of the present work. The seventh section concludes the paper with future works that can be done as an outcome of this work.

2 LITERATURE SURVEY

Text mining is an interdisciplinary field that involves the application of NLP, ML, and data mining techniques to extract meaningful information from large volumes of text data. The authors of [3] described text mining as a development of data mining methodology. Numerous fields have used text mining, including business intelligence, social media analysis, healthcare, and scientific literature analysis. Text mining has its origins in the fields of computational linguistics and information retrieval. Early research works, such as the Vector Space model for information retrieval [4] and Latent Semantic Analysis (LSA) [5], paved the path for further improvements. Specifically, Word2Vec's introduction [6] transformed word representation, and Latent Dirichlet Allocation (LDA) [7] offered strong topic modelling capabilities.

A deep learning approach using a parallel architecture of Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) for sentiment intensity scoring of English tweets is presented in [8]. The authors extract two sets of features to aid the network in judging emotion intensity and describe experiments on different models and various feature sets, along with the analysis of the results. The data for this task consists of tweets across various domains, classified into four emotions: joy, sadness, anger, and fear. Each tweet is represented as a 35 x d matrix, where d is the output dimension of embedding of a single word. GloVe Word Embeddings, trained on 2 billion tweets from Twitter, are used for the datasets corresponding to anger, fear, and joy emotions. The processed text is converted to word embeddings, representing each word of the text into a d-dimensional vector.

The authors of the paper [9] focus on the use of text mining algorithms to extract knowledge from unstructured political information found in newspapers, specifically the Nigerian Guardian newspaper. The algorithm involves natural language processing techniques such as tokenization, text filtering, and refinement, followed by Association Rule mining for knowledge extraction. The main contribution

of the technique is the integration of an information retrieval scheme (Term Frequency Inverse Document Frequency) with a data mining technique for association rules discovery. The program is applied to pre-election information from the Nigerian Guardian newspaper, and the extracted Association Rules provide important features and informative news related to the concluded 2007 presidential election. The system presented in the paper aims to provide useful information that can help sanitize the political environment and protect the nascent democracy.

The paper [10] demonstrates how to analyze Turkish written text using Big Data technologies such as Nutch, Spark, and MongoDB, which can be applied to text mining in any language. The study focuses on the linguistic aspects of the Turkish language, which presents challenges for text mining due to its word order-free and agglutinative nature and the complexity of word stemming. The study highlights the need for well-defined natural language characteristics for Turkish, including different styles, formal and informal language, ambiguities, and context. The study aims to enhance text mining efficiency in Turkish by demonstrating the issues, methods, and applications of qualitative approaches, including preprocessing, feature selection, and topic clustering.

In [11] highlights the importance of idea mining in improving strategic decision making and discusses the efficient computational methods for idea characterization based on concept extraction from unstructured texts. It explores various successful text mining tools and text classification techniques that can be used to extract ideas from different sources such as patents, publications, reports, documents, and the internet. They emphasize the need for a combination of idea-mining measures to effectively extract and characterize ideas from the text. The paper provides insights into the methods and tools available for mining ideas from the text, showcasing their potential to enhance decision-making processes.

The paper [12] provides a review of text mining, discussing its research status, general models, and applications such as text categorization, text clustering, association rule extraction, and trend analysis. The work presented in [13] provides a review of text mining, which is the process of deriving high quality information from text. It discusses the applications of text mining in various areas such as customer care service, fraud detection, contextual advertisement, and healthcare.

Paper [14] discusses machine learning methods for text mining, specifically focusing on unstructured natural-language documents rather than structured databases. It reviews standard classification and clustering methods for text, such as Naive Bayes (NB), Rocchio, Nearest Neighbor, and Support Vector Machines (SVMs) for classifying texts, and hierarchical agglomerative, spherical k-means, and Expectation Maximization (EM) methods for clustering texts. The paper also covers Information Extraction (IE) methods that use sequence information to identify entities and relations in documents, including Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) for sequence labeling and IE. The discussed methods are motivated by applications in spam filtering, information retrieval, recommendation systems, and bioinformatics.

Text and content mining [15] are subcategories of data mining that extract information from web content, such as web pages and search logs. The extracted information can be used in various applications, including extracting opinions from online sources and analyzing web hierarchy for better insights and knowledge. The systematic review included 18 research papers that focused on the applications, techniques, and issues of text and web mining. The research papers provided a good foundation for the topic, explaining different techniques used in text and web mining and highlighting potential future research areas.

The paper [16] provides an overview of recent advances in text mining of Indian languages, including NER systems, feature evaluation methods, Dimensionality Reduction Techniques, association rules, clustering algorithms, and sentiment analysis. The authors of the paper discuss a comparative study of Feature Dimensionality Reduction Techniques applied to Hindi and Bengali named entity recognition (NER). They studied various feature evaluation methods like TF, tf-idf, Mutual Information (MI), Information Gain (IG), chi-square test, and relief. They also selected features using three methods: wrapper, filter, and embedded. Finally, they extracted features using Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA). Clustering methods were used for dimensionality reduction. The paper by [17] discusses the challenges of text mining in Indian languages, such as tokenization, sentence boundary disambiguation, lack of annotated corpora, and non-availability of full-scale gazetteer lists.

The authors [18] present a bibliometric analysis of articles on the influence of Covid-19 on consumer behaviour, using Biblioshiny and VOSviewer applications. It identifies the most influential documents, authors, affiliations, countries, and journals in this research domain. The paper conducts citation, co-citation, and keyword co-occurrence analysis, visualized in a clustered network diagram. Three main themes in consumer behaviour research amid Covid-19 are identified: a) Food purchasing decisions and food wastage, b) Adoption of technology, and c) Intrinsic and extrinsic influence on consumer behaviour. The paper highlights the impact of Covid-19 pandemic on consumer behaviour and the need for marketeers to understand and adapt to these changes. It offers insights into emerging themes and subthemes in consumer behaviour research, contributing to the future expansion of this research domain.

With an emphasis on Indian languages, the paper [19] proposes a content-based approach called CBTM (Content-Based Text-Mining) for knowledge discovery of multilingual texts. This system used keywords and patterns stored as gif strings to enable extensions to other forms of data, and potential applications in a distributed environment are highlighted. This approach was found effective in extracting knowledge from multilingual texts, including ancient and out-of-print texts in different languages. The application of the Content-Based Text-Mining technique for knowledge extraction from newspaper advertisements is also demonstrated in this study.

3 TEXT MINING IN INDIAN LANGUAGES

India is a multilingual nation and its growing focus on internet services is being provided in regional languages. Within data mining, text mining is becoming more and more popular. If the content is provided in the local language on the internet, the users of the internet could be increased. Also the existing internet users, long for content in the language of their choice. Under Technology Development for Indian Languages (TDIL) projects Development of Corpora, Optical Character Recognization (OCR), Text-to-Speech (TTS), Machine Translation (MT), and Generic Software for Information processing, etc. were supported. So availability of the constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. The web content in Indian languages also has increased. Numerous portals have emerged that include huge volumes of content in Indian languages. However, the data is being under-utilized due to the unavailability of Indian language text mining methods. The huge number of available documents in digital media makes it difficult to obtain the necessary knowledge related to the needs of the user. Thus, text mining is necessary for regional languages spoken in India.

TM for Indian languages has several difficulties. Indian languages differ from English in several ways, which makes it challenging to use text-mining tools and techniques that were created for the English language. The great variety of Indian languages also makes it difficult to find linguistic resources for text mining in Indian languages. Additionally, a gap exists between the knowledge that can be created from stored data and a paucity of study and development done in Indian language text processing. The extraction of text data from photographs presents another difficulty because the text can have different styles, sizes, orientations, alignments, and lighting conditions. Moreover, as majority of the study in this topic is limited to English language, opinion mining in Indian languages also poses difficulties.

3.1 SPECIALITY OF MALAYALAM

The distinctive script used in Malayalam was inspired by old Brahmi scripts. A combination of vowels and consonants is represented by the characters of this alphasyllabic alphabet. It has a complex structure of derivations and inflections and is morphologically rich. As an agglutinative language, Malayalam represents grammatical information through the addition of affixes to root words. This property makes it possible to combine different morphemes to create complex words. Understanding and recognizing these structures is necessary for efficient processing. Malayalam exhibits regional variances that result in variations in vocabulary, pronunciation, and specific grammatical characteristics. Code-mixing, or inserting English or other languages into Malayalam text, is a common practice among Malayalam speakers. Text mining technologies must be able to handle various languages within a single corpus to address this linguistic phenomenon.

3.2 NEED FOR TEXT MINING TOOL FOR MALAYALAM

Malayalam has its unique script and linguistic characteristics. From the enormous quantity of Malayalam data that is available online, such as news articles, social media posts, and other textual data sources, Malayalam TM tools can assist in retrieving relevant data. Analyzing Malayalam text provides insights into various aspects such as sentiment, topics, and entities. This can be valuable for understanding public opinion, monitoring trends, and gaining knowledge from Malayalam-language content. For researchers, scholars, and academicians who are interested in Malayalam language studies, the TM tool aids in analysing and extracting meaningful information from large corpora of Malayalam texts. It encourages research in linguistics, literature, and cultural studies. These tools can be useful for media organizations and businesses to track mentions, sentiments, and trending subjects relating to their brands or interests in news stories and social media. The development and improvement of Malayalam TM tools contribute to the creation of language resources such as annotated corpora, lexicons, and language models.

3.3 CHALLENGES IN DEVELOPING MALAYALAM TM TOOL

Developing a text mining tool for Malayalam presents several formidable challenges. The absence of standardised scripts and the morphological complexity of Malayalam, characterized by intricate word formations, pose difficulties in tasks such as stemming and lemmatization. Limited language resources, including annotated corpora and lexicons, hinder the training and evaluation of natural language processing models. Code-mixing with English or other Indian languages, along with semantic ambiguity in sentence structures, adds complexity to language understanding. Named entity recognition faces challenges due to variations in naming conventions, and the diverse dialects and vernaculars across regions require adaptable tools. The evolving state of Natural Language Processing research for Malayalam, coupled with the lack of standard evaluation benchmarks, further complicates the development and assessment of effective text mining tools. These challenges underscore the need for collaborative efforts to advance linguistic resources and algorithmic solutions for the unique characteristics of Malayalam language.

4 METHODOLOGY AND IMPLEMENTATION

Large volumes of unprocessed data can be studied with the help of text mining to find valuable insights. To develop text analysis models that learn to categorize or extract certain information based on prior training can be integrated with machine learning. The basic steps involved in text mining are illustrated in Figure: 1.

The initial stage in text mining is to gather and prepare the text data to be analysed. This could include scrap-

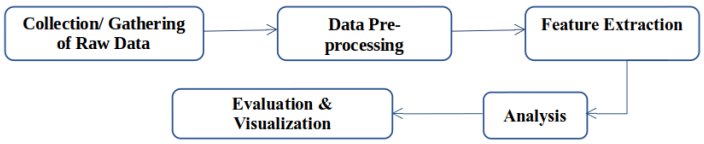

Figure 1: Basic Steps in Text Mining

ing content from web pages, extracting data, or gathering social media posts, documents, articles, or websites. Once collected, the data must be pre-processed before being analysed. This may include deleting special characters, stop words, and punctuation, as well as activities like tokenization, stemming or lemmatization. Tokenization is the process of breaking a text into individual words or phrases, or tokens. This step facilitates further analysis by providing a unit of meaning for the subsequent processing stages. Stemming involves reducing words to their root or base form by removing suffixes, while lemmatization aims to transform words to their base or dictionary form. Both processes help in standardizing word representations, reducing the complexity of the data. These are NLP techniques used by TM systems to produce inputs for machine learning models. This allows the text analysis to proceed. Depending on the linguistic properties of the content, language-specific processing may be used.

Extraction of pertinent features from text data comes next after pre-processing. Two popular methods are Bag-of-Words (BoW), in which documents are represented as collections of words together with their frequencies, and Term Frequency-Inverse Document Frequency (TF-IDF), which weighs words according to how important they are in a document with the corpus as a whole. Word embeddings that capture semantic relationships, like Word2Vec or GloVe, give dense vector representations of words in a continuous vector space. While named entity recognition, sentiment ratings, and part-of-speech tags enhance text comprehension, N-grams take into account word sequences to identify local patterns. Depending on the particular task at hand, the feature extraction technique selected will attempt to accurately capture the complex nature and context of the textual data for further analysis and machine learning.

After the text data has been pre-processed and features retrieved, the next step is to analyse it to extract insights and information. Clustering, classification, topic modelling, sentiment analysis, and entity identification techniques may be used to identify patterns, trends, and correlations in the data. After analysis, the results must be reviewed to determine their success in meeting the text mining project's objectives. This could include evaluating the accuracy of classification models or the coherence of topic models. Depending on the particular task, evaluation entails a quantitative assessment of a text mining model's performance using metrics like accuracy, precision, recall, F1 score, or others. It gives a numerical measure of the model's effectiveness in accomplishing its goals, assisting in identifying the benefits as well as drawbacks of their selected approaches. On the contrary, visualisation makes it possible to present complex textual data in an understandable way. Patterns, trends, and relationships within the data are visually conveyed through techniques such as word clouds, bar charts, and heatmaps. Results may be interpreted more easily and conclusions can be communicated to a wider range of users with the help of visualisation.

4.1 MALAYALAM TEXT MINING TOOL – IMPLEMENTATION

The task discussed here is to extract data from numerous online Malayalam newspapers and organize it into a well-structured corpus database. The work also focuses on tagging the extracted corpus for Parts of Speech (POS). POS Tagging is a process of assigning the best part of speech to the constituents of the sentence. Identification of the parts of speech, such as nouns, pronouns, verbs, adjectives, and adverbs helps in analysing the role of each constituent in a sentence. For this, approximately 15 online newspaper sites were examined, analysing the structure and determining the viability of mining data from these websites. When checked manually, majority of them do not satisfy the required criteria, i.e. they do not reveal the actual result or the required data or are not in legible format; some lack data information. Finally, it was decided to choose five newspapers that satisfied all the criteria. These include dailies such as MalayalaManorama, Mathrubhumi, Deshabhimani, Madhyamam, and Pravasi Express.

Text mining for Malayalam follows the same processes as TM for any other language. The basic architecture for developing TM tool for Malayalam along with POS tagging is depicted in Figure: 2.

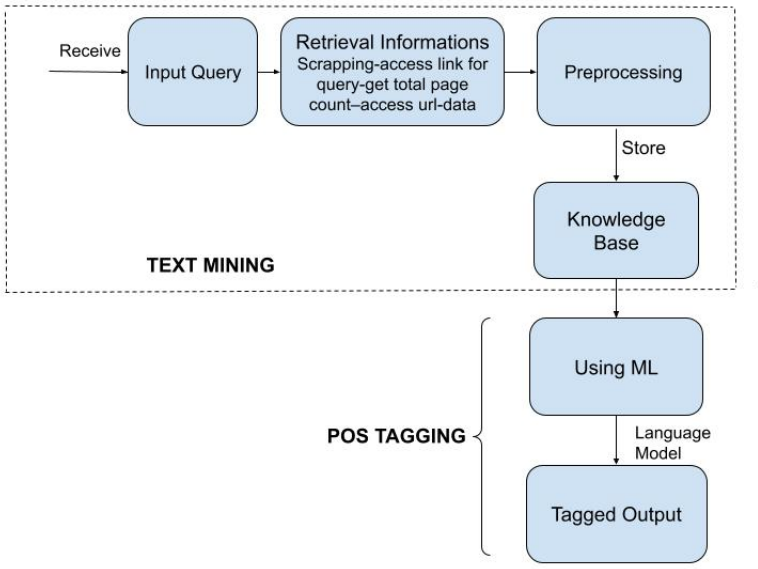

Figure 2: Architecture of TM tool for Malayalam

The first and foremost step in mining the data is to define a query and submit it to TM tool. The algorithm for extracting the text from different newspapers is given in Algorithm: 1.

As per the given query, the next stage is data gathering, in which useful information is obtained from online news sources using techniques like web scraping and API (Appli-

Algorithm 1 Extract Text For Query From Malayalam Online Newspapers

Input:
Query for searching
List of online Malayalam newspaper URLs

Initialize BeautifulSoup
Initialize an empty list to store textual data extracted
for each newspaper's URL **do**
  for each article page **do**
    Fetch HTML content from the constructed URL using the 'requests' library
    if the request is successful
      Parse the HTML content with BeautifulSoup
    else
      Handle the error
    endif
    Identify HTML Elements
      Inspect the HTML structure of the article webpage using browser developer tools
      Determine the HTML tags containing the article's text
    Extract Text
      Use BeautifulSoup methods to locate the identified HTML elements
      for each element **do**
        Extract the text content using the get_Text() method
        Append extracted text to the list for the current newspaper
      endfor
      Move to the next page based on the pagination template
  endfor
  Store the extracted text for the current newspaper in a separate list
endfor
Handle dynamic content, if applicable
Review continuously webpage structure and make changes in code accordingly.
Output:
A list containing the textual data related to the given query extracted from different Malayalam newspapers.

cation Programming Interface) utilization. Web pages are drastically different in style and form from text documents used for text mining. The majority of online pages show their content using HTML (Hyper Text Markup Language) rather than plain text. HTML readily allows a page to be a combination of textual content, known as the page payload; formatting information, such as tables and headers; multimedia features, such as images or video; and links to other HTML pages. The data preparation is the next step. The retrieved raw text is then put through pre-processing procedures to guarantee the quality of the data, including cleaning, removing HTML elements, and addressing noise. Once collected, the data must be pre-processed before being analysed. For Malayalam, this may involve employing tools such as Morphological Analysers, Sandhisplitter, and stopword lists, which can be used to remove common words that are unhelpful for analysis. After pre-processing, the data thus gathered are stored in the database in text format or CSV format.

POS is the task of assigning each word of a text to the proper parts of speech tag in its context of appearance in the sentences. This is considered to be the first step towards understanding any natural language. POS tags augment the information contained among the words alone that help in explicitly indicating the structure that is inherent in a language. CRFs are a probabilistic framework [22] that is used for labelling and segmenting structured data, such as sequences, trees, and lattices. The tool incorporates CRF based POS tagging to find entities or patterns in Malayalam text. When identifying a pattern, this model takes into account not just the dependence of features on one another but also future observations. It is thought to be the most effective approach for entity recognition in terms of performance. Since these models incorporate previous data, users feed CRF with features that are modelled from the data. These feature functions, such as the tag sequence noun - adverb - verb, describe particular characteristics of the sequence that the data point represents. Treebank Malayalam POS annotated corpus was used along with collected corpus. The collected corpus was annotated with BIS Tagset [20] for getting the POS tagged data.

5 RESULTS AND DISCUSSIONS

When a query related to any specific topic is given, the text mining tool provides a set of documents, sentences, or keywords that are most relevant to the given query. Thus obtained data is pre-processed thoroughly and then stored in the TXT, CSV format for future uses. The CSV file contains all the details of the data collected like, the title of the article, date of publication of the article, content, article link, newspaper details etc is shown in Figure: 3. The TXT file contains the plain raw data which can be used for many NLP applications.

One essential method for figuring out the linguistic variations present in the language is to use word clouds and frequency distribution. By utilizing these visualization tools, one can do a thorough investigation of the most frequently occurring terms in a corpus, providing a valuable understanding of the semantic landscape and major themes included in the given text. A simple and visually appealing method for quickly identifying key concepts is the word cloud, a graphical depiction of words sized according to how frequently they appear in the text. Through the creation of a Malayalam-specific word cloud, it is possible to effectively identify the frequently occurring terms in the mined data, providing insight into the contextual complexities and lexical richness. This helps to identify important trends and patterns and also contributes to the qualitative understanding of the text. The wordcloud for given query "മിഷൻ", mined from all five newspapers is shown in Figure: 4 and the frequency of the most commonly used words in the same is depicted in Figure: 5.

As per the research study [21], cloud technique is used to

Show 10 entries Search:

INDEX	SITE	URL	TITLE	POSTTIME	CONTENT
0	Mathrubhumi	https://www.math... net-december-2023-result-declared-1.9252619	യു.ജി.സി. നെറ്റ് ഡിസംബർ 2023 ഫലം പ്രസിദ്ധീകരിച്ചു ന്യൂഡൽഹി: ഡിസംബറിൽ നടത്തിയ യുജിസി നെറ്റ് (നാഷണൽ എലിജിബിലിറ്റി ടെസ്റ്റ്) ഫലം പ്രഖ്യാപിച്ചു. ആപ്ലിക്കേഷൻ നമ്പറും ജനനത്തീയതിയും ഉപയോഗിച്ച് ugcnet.nta.ac.in ...	2024-01-19	ന്യൂഡൽഹി: ഡിസംബറിൽ നടത്തിയ യുജിസി നെറ്റ് (നാഷണൽ എലിജിബിലിറ്റി ടെസ്റ്റ്) ഫലം പ്രഖ്യാപിച്ചു. ആപ്ലിക്കേഷൻ നമ്പറും ജനനത്തീയതിയും ഉപയോഗിച്ച് ugcnet.nta.ac.in വഴി ഫലമറിയാം. ജനുവരി 10-ന് ഫലം പ്രഖ്യാപിക്കുമെന്നായിരുന്ന നാഷണൽ ടെസ്റ്റിങ് ഏജൻസി നേരത്തെ അറിയിച്ചിരുന്നത്. എന്നാൽ മിഷോങ് ചുഴലിക്കാറ്റിന്റെ പശ്ചാത്തലത്തിൽ പ്രളയത്തിലകപ്പെട്ട ചെന്നൈയിലും ആന്ധ്രാപ്രദേശിലും വീണ്ടും പരീക്ഷ നടത്തിയതാണ് ഫലപ്രഖ്യാപനം നീളാൻ കാരണം.ജൂനിയർ റിസർച്ച് ഫെലോഷിപ്പോടെയുള്ള ഗവേഷണത്തിനും, മാനവിക വിഷയങ്ങളിൽ അസിസ്റ്റന്റ് പ്രൊഫസർ നിയമനത്തിനുള്ള യോഗ്യത പരീക്ഷയാണ് യുജിസി നെറ്റ്.
1	Mathrubhumi	https://www.... net-december-2023-result-on-january-17-says-nta-1.9227303	യുജിസി നെറ്റ് ഡിസംബർ 2023 പരീക്ഷാഫലം ജനുവരി 17 ന് ന്യൂഡൽഹി: ഡിസംബറിൽ നടത്തിയ യുജിസി നെറ്റ് (നാഷണൽ എലിജിബിലിറ്റി ടെസ്റ്റ്) ഫലം ജനുവരി 17 ന് പ്രഖ്യാപിക്കുമെന്ന് നാഷണൽ ടെസ്റ്റിങ് ഏജൻസി. ആപ്ലിക്കേഷൻ നമ്പറും ...	2024-01-10	ന്യൂഡൽഹി: ഡിസംബറിൽ നടത്തിയ യുജിസി നെറ്റ് (നാഷണൽ എലിജിബിലിറ്റി ടെസ്റ്റ്) ഫലം ജനുവരി 17 ന് പ്രഖ്യാപിക്കുമെന്ന് നാഷണൽ ടെസ്റ്റിങ് ഏജൻസി. ആപ്ലിക്കേഷൻ നമ്പറും ജനനതീയതിയും ഉപയോഗിച്ച് ugcnet.nta.ac.in വഴി ഫലമറിയാം. ജനുവരി 10ന് ഫലം പ്രഖ്യാപിക്കുമെന്നായിരുന്ന എൻടിഎ നേരത്തെ അറിയിച്ചിരുന്നത്. എന്നാൽ മിഷോങ് ചുഴലിക്കാറ്റിന്റെ പശ്ചാത്തലത്തിൽ പ്രളയത്തിലകപ്പെട്ട ചെന്നൈയിലും ആന്ധ്രാപ്രദേശിലും വീണ്ടും പരീക്ഷ നടത്തിയതാണ് ഫലപ്രഖ്യാപനം 17-ലേക്ക് മാറ്റിയതെന്ന് എൻ.ടി.എ അറിയിച്ചുള്ളനിയർ റിസർച്ച് ഫെലോഷിപ്പോടെയുള്ള ഗവേഷണത്തിനും, മാനവിക വിഷയങ്ങളിൽ അസിസ്റ്റന്റ് പ്രൊഫസർ നിയമനത്തിനുള്ള യോഗ്യത പരീക്ഷയാണ് യുജിസി നെറ്റ്.
2	Mathrubhumi	https://www.math... music/news/siva-karthikeyan-donated-10-lakhs-to-tamil-nadu-chief-minister-s-flood-relief-fund-1.9145877	ചെന്നൈ വെള്ളപ്പൊക്കം: ഉദയനിധിയെ കണ്ട് 10 ലക്ഷം രൂപയുടെ ചെക്ക് കൈമാറി ശിവ കാർത്തികേയൻ ചെന്നൈ: മിഷോങ് ചുഴലിക്കാറ്റിനെത്തുടർന്നുണ്ടായ മഴയിലും വെള്ളപ്പൊക്കത്തിലും തമിഴ്നാടിന്റെ വിവിധഭാഗങ്ങൾ അനുഭവിച്ച ദുരിതം കനത്തതാണല്ലോ. വെള്ളപ്പൊക്കം സൃഷ്ടിച്ച ...	2023-12-11	ചെന്നൈ: മിഷോങ് ചുഴലിക്കാറ്റിനെത്തുടർന്നുണ്ടായ മഴയിലും വെള്ളപ്പൊക്കത്തിലും തമിഴ്നാടിന്റെ വിവിധഭാഗങ്ങൾ അനുഭവിച്ച ദുരിതം കനത്തതാണല്ലോ. വെള്ളപ്പൊക്കം സൃഷ്ടിച്ച കെടുതികളിൽ നിന്ന് സംസ്ഥാനം പതിയെ കരകയറി വരികയാണ്. ഇപ്പോൾ ദുരിതബാധിതരെ സഹായിക്കാൻ മനോട്ടുവന്നിരിക്കുകയാണ് നടൻ ശിവ കാർത്തികേയൻ. സംസ്ഥാന ദുരിതാശ്വാസ നിധിയിലേക്ക് സംഭാവനയുംചെയ്തു താരം.കഴിഞ്ഞദിവസം ശിവ കാർത്തികേയൻ തമിഴ്നാട് മന്ത്രിയും നിർമാതാവുമായ ഉദയനിധി സ്റ്റാലിനെ സന്ദർശിച്ചിരുന്നു. ദുരിതാശ്വാസ പ്രവർത്തനങ്ങളിൽ പങ്കാളിയാവാൻ സന്നദ്ധത പ്രകടിപ്പിച്ച താരം മുഖ്യമന്ത്രിയുടെ ദുരിതാശ്വാസനിധിയിലേക്കുള്ള പത്ത് ലക്ഷം രൂപയുടെ ചെക്ക് ഉദയനിധിക്ക് കൈമാറി. എല്ലാ

Figure 3: CSV File format

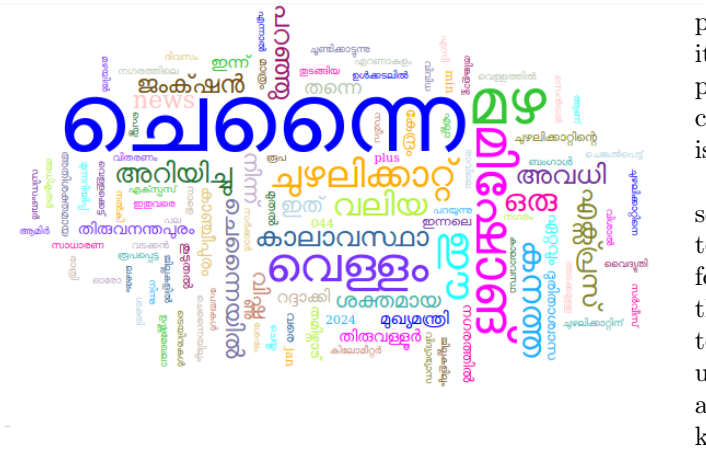

Figure 4: WordCloud for "മിഷോങ്"

represent the most frequent keywords collected from the articles. As shown in Figure:5, it is found that "ചെന്നൈ"is the most occurring term or keyword that is mentioned across all the collected newspapers articles. The next highest frequent keywords that occurred in the collected corpus includes "മഴ" and "മിഷോങ്" followed by "വെള്ളം", "ഈ", "ചുഴലിക്കാറ്റ്", "കനത്ത" keywords.

The analysis of frequency distribution offers a numerical viewpoint on the frequency of specific terms inside the Malayalam text. By displaying the frequency of each word and providing an organized output, this distribution enables more thorough statistical analysis. This distribution can be used by scholars and language lovers to determine the frequency of particular terms, which can help identify important linguistic traits and support further analytical work. In the realm of text mining, a confusion matrix is a valuable tool for assessing the performance of a classification model. It provides a comprehensive breakdown of the predicted and actual class labels, aiding in the evaluation of model accuracy. Precision measures the accuracy of positive predictions, while recall assesses the model's ability to capture all positive instances. The balance between precision and recall is given by F1 score. Figure:6 gives the classification report and the confusion matrix for the same is depicted in Figure:7.

When it comes to POS tagging, in Malayalam there are some words that may take different tags in different contexts. Good knowledge of Malayalam grammar is needed for tagging words. The tagger is trained using CRF with the respective annotated corpus consisting of around 210k tokens and thus the respective language model is generated using machine learning techniques. Testing was undertaken after learning, and the results obtained for POS tagging for known and unknown words are given in Table: 1.

Table 1: Known vs Unknown Accuracy

Known Data	98.89
UnKnown Data	84.67

The results obtained are very promising and further enhancements can be done by training the CRF with more annotated corpus.

TM makes it possible for a systematic study of huge amounts of Malayalam text, which could be helpful especially for academic research, policy makers, Malayalam communities, article writing, or teaching purposes. Instead of just searching for, linking to, and obtaining documents containing specific data, the goal of TM is to extract useful information. This tool can be used for collecting corpus for various Malayalam langauge processing applications like Named Entity Recognizer, Multiword Expressions, Sentiment Analysis, etc. Text mining, unlike searches, produces results based on the researcher's intended use of the content. Web searches in general might be similar to TM, but there are notable differences. Search is the process of re-

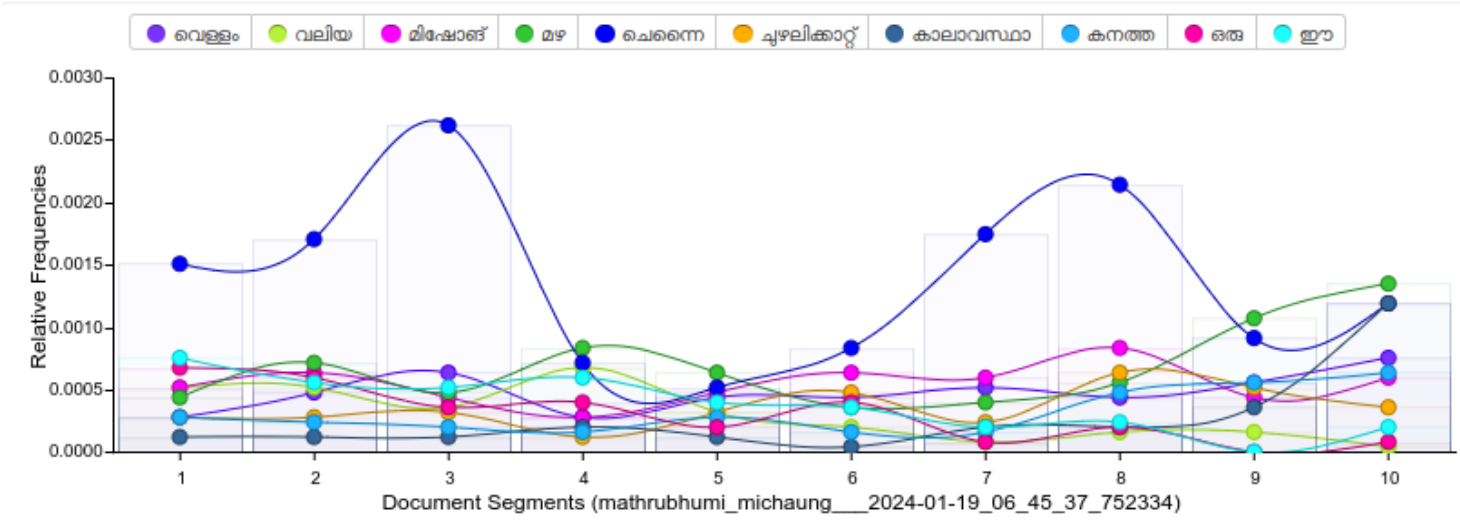

Figure 5: Frequency Distribution of most commonly used terms in Collected Newspaper Articles

	0	1	accuracy	macro avg	weighted avg
precision	0.0	1.000000	0.967742	0.500000	1.000000
recall	0.0	0.967742	0.967742	0.483871	0.967742
f1-score	0.0	0.983607	0.967742	0.491803	0.983607
support	0.0	93.000000	0.967742	93.000000	93.000000

Figure 6: Classification Report

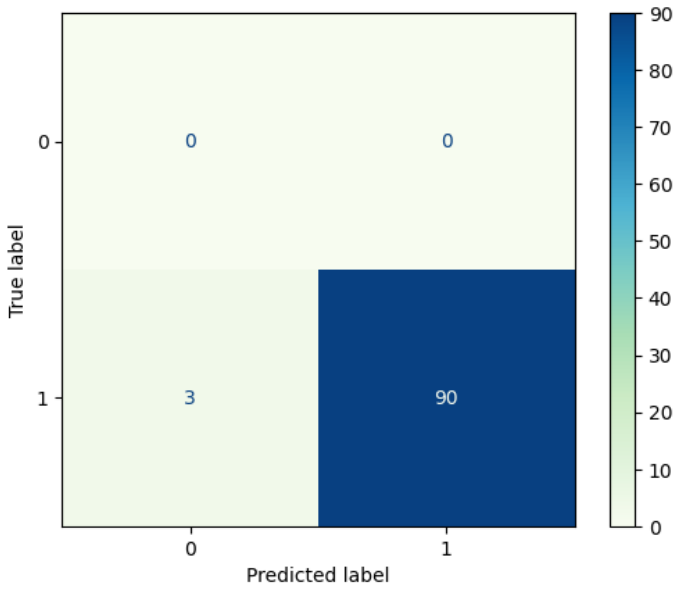

Figure 7: Confusion Matrix

trieving documents or other results based on specific search keywords. This kind of search is commonly performed using search engines. The output often consists of a hyperlink to text or information located elsewhere, as well as a summary of what is available at the other end of the link. The goal is to locate the entire current work so that its contents can be exploited.

6 LIMITATIONS AND FUTURE SCOPE

Malayalam has regional dialects and variations in vocabulary and syntax, making it difficult to create models that work well across different dialects. The performance of tools for Malayalam text mining may vary depending on the domain, such as news articles, literature, and social media, and may require specific adjustments. There are limited benchmark datasets and evaluation standards for Malayalam, which can make it difficult to assess the performance of text mining tools. POS taggers may have accuracy issues due to the quality and size of the annotated corpus used for training, and may struggle with disambiguating word meanings. Errors in pre-processing and feature extraction can lead to inaccuracies in downstream analyses.

As a future scope, by incorporating Government websites to this tool, public can easily access all relavent information related to specific query from this portal, instead of searching different Government portals. As a further enhancement to this work, information can be made available from pdf files other than textual data. For better understanding of the semantic landscape and important themes included in the retreived data, visualization tools can be incorporated. Future research could explore additional ML algorithms for comparison, paving the way for deeper understanding and more advanced applications of Malayalam text mining across various sectors.

7 CONCLUSION

In an era where information is abundant but often overwhelming, the ability to distill meaning from unstructured text is paramount. The creation of a Malayalam text mining tool is an important step towards enhancing linguistic technology for the Malayalam language. This endeavour smoothly integrates NLP and machine learning, demonstrating a strong interaction between algorithms and lin-

guistic characteristics. The tool not only enhances Malayalam computing but also demonstrates the ongoing progress of language-centric technologies. This tool helps to collect corpus from five different newspapers which is the core element for many NLP applications. The corpus collected can be annotated for POS and the accuracy in predicting the tags is quite promising. The comparison of POS results has been carried out for the known and unknown data. It is also found that the increase in the training data can help in increasing the accuracy of predicting tags. Other ML algorithms can be applied and the results can be compared with this. In trying to explore deeper into the complexities of Malayalam text mining, this research lays the way for future advancements, encouraging a better knowledge of the language's unique features and enabling more advanced applications in a variety of sectors.

References

[1] Afsha, Akkalkot. (2023). A Survey on Text Mining - Techniques, Application. International journal of scientific research in computer science, engineering, and information technology, doi: 10.32628/cseit2390391

[2] Sorensen, L. 2009. User managed trust in social networking comparing Facebook, MySpace and LinkedIn. In Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace Electronic System Technology, (Wireless VITAE 09), 427431

[3] Liu, F. Lu, X. 2011. Survey on text clustering algorithm. In Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS), 901904.Google Scholar

[4] G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. Commun. ACM 18, 11 (Nov. 1975), 613620. https://doi.org/10.1145/361219.361220

[5] Landauer, T.K. and Dumais, S.T. (1997) A Solution to Platos Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, Representation of Knowledge. Psychological Review, 104, 211-240.

[6] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[7] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning Research 3, no. Jan (2003): 993-1022.

[8] Textmining at EmoInt-2017: A Deep Learning Approach to Sentiment Intensity Scoring of English Tweets, Hardik Meisheri, Rupsa Saha, Priyanka Sinha, Lipika Dey, Harvard University1, Indian Institute of Technology Kharagpur

[9] Knowledge Discovery in Online Repositories: A TextMining Approach, I. T. Afolabi, G. A. Musa, C. K. Ayo, A. B. Sofoluwe

[10] Text Mining Analysis in Turkish Language Using Big Data Tools, Mehmet Ulas Cakir, Seren Guldamlasioglu, Vol. 1, pp 614-618, Computer Software and Applications Conference

[11] A review of methods for mining ideas from text, Mostafa Alksher, Azreen Azman, Razali Yaakob, Rabiah Abdul Kadir, Universiti Putra Malaysia1, National University of Malaysia, Sabha University, Ibb University

[12] Yu, Zhang., Mengdong, Chen., Lianzhong, Liu. (2015). A review on text mining. doi: 10.1109/ICSESS.2015.7339149

[13] Kanak, Sharma., Ashish, Sharma., Dhananjay, Joshi., Nikhil, Vyas., Arpit, Bapna. (2017). A Review of Text Mining Techniques & Applications.

[14] Sudha, Cheerkoot-Jalim., Kavi, Kumar, Khedo. (2021). A systematic review of text mining approaches applied to various application areas in the biomedical domain. Journal of Knowledge Management, doi: 10.1108/JKM-09-2019-0524

[15] Fatima, Almatrooshi., Sumayya, Alhammadi., Said, A., Salloum., Khaled, Shaalan. (2021). Text and Web Content Mining: A Systematic Review. Doi: 10.1007/978-3-030-82616-1_8

[16] M., Hanumanthappa., M., Narayana, Swamy. (2015). Indian language text mining.

[17] A review of recent advances in text mining of Indian languages, Prabin Kumar Panigrahi1, Nishikant Bele, Indian Institute of Management Indore, Vol. 23, Iss: 2, pp 175-193 : Business Information Systems

[18] Abu, Bashar., Brighton, Nyagadza., Neo, Ligaraba., Eugine, Tafadzwa, Maziriri. (2023). The influence of Covid-19 on consumer behaviour: a bibliometric review analysis and text mining. Arab Gulf Journal of Scientific Research, doi: 10.1108/agjsr-12-2022-0281

[19] Unified Parts of Speech (POS) Standard in Indian Languages https://tdil-dc.in/tdildcMain/articles/34692Draft%20POS%20Tag%20standard.pdf

[20] Shailaja Jayashankar and Sridaran R. "Superlative model using word cloud for short answers evaluation in eLearning." Education and Information Technologies 22, no. 5 (2017): 2383-2402.

[21] Wallach, Hanna M. Conditional random fields: An introduction. Technical Reports (CIS). 2004. pp 22.

[22] Chitrakala, S. and Manjula, D. Distributed multilingual content based text mining. In Proceedings of the third conference on IASTED International Conference: Advances in Computer Science and Technology. 2007. pp. 500-505.