International Journal of Science, Strategic Management and Technology
Volume **02** Issue **06** June-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

An Approach for Cloud-Based Framework for Multilingual Text Summarization and Sentiment Analysis in Indian Languages

[1] Adlakadi. Anand,

Ph.D. scholar & Assistant Professor in Computer Science and Engineering, Malla Reddy Technical Campus, Malla Reddy Vishwa Vidyapeeth (Deemed to be University) – Hyderabad. 2025506015001@mrvv.edu.in.

[2] Dr. Praveen Kumar P,

Professor in Computer Science and Engineering, Malla Reddy Technical Campus, Malla Reddy Vishwa Vidyapeeth (Deemed to be University) – Hyderabad.

https://doi.org/10.55041/ijsmt.v2i6.218

Cite this Article: Anand, A. (2026). An Approach for Cloud-Based Framework for Multilingual Text Summarization and Sentiment Analysis in Indian Languages. International Journal of Science, Strategic Management and Technology, *02*(6). https://doi.org/10.55041/ijsmt.v2i6.218

ABSTRACT: With the rise of online shopping and the abundance of information available online, the internet has become the go-to place for customers seeking unbiased reviews and different perspectives on a certain product, news story, issue, or market trend. Therefore, it is preferable to use systems that filter through the mountain of data and summarise the available ideas so that the seeker can easily grasp them in order for making this search simpler. Performing this kind of work is now a hotspot for academic inquiry into sentiment analysis. Businesses, data analysts, data scientists, and consumers may all benefit from sentiment analysis. A dearth of systems capable of analysing data in languages other than English exists, despite the fact that several ways have been developed to do this job on English data. A comprehensive examination of sentiment analysis techniques used on non-English languages is the goal of this work. Every method's efficiency, along with its tools, advantages, and downsides, are detailed. The difficulties that come with it are also covered. Methods for analysing data in both the source and target languages are discussed in the article.

Keywords: Topics covered include sentiment analysis, web data mining, text mining, natural language processing, deep learning, multi-lingualism, and cross-lingualism.

I. INTRODUCTION

An information explosion has spread over the internet as a result of the proliferation of online platforms & social media in our everyday lives. Therefore, there is a plethora of information available on any subject, including goods, companies, market trends, etc. Also publicly accessible in abundance are the thoughts of consumers of such sites. Anyone looking to make an educated choice has easy access to data such as movie reviews, evaluations of goods, feelings in the financial markets, political perspectives, etc. Such a setting calls for, and makes good use of, a system that sorts through mountains of data, evaluates them, and ideally classifies, quantifies, or scores them to help in decision making. Sentiment analysis, often called opinion mining, emerged as a result of

this. The term "sentiment analysis" refers to a method of categorising, rating, or quantifying the general tone and viewpoint of written materials. Getting a sense of how people generally feel about a specific issue is the main goal [1, 2]. A large quantity of unstructured & unlabelled data is the main focus of sentiment analysis. In addition, the data that is currently accessible is often subjective, lacks specificity, and does not comply rigidly to linguistic standards. Due to the multi-domain nature of sentiment analysis, experts in fields such as data analytics, computer intelligence, machine learning, the processing of natural languages, data mining, etc. are required to complete the work. Sentiment analysis, at its core, is a classification issue that seeks to assign overarching positive, negative, or neutral labels to various points of view. The benefits of in-depth research, however, include the ability to uncover more relevant features and a greater quantity of data [2, 3]. A lot of companies rely on customer reviews to inform their decisions. Methods like as opinion polls, questionnaires, and surveys of customers are used to get this data. The proliferation of internet access has allowed for more people to participate in these types of surveys, increasing the reliability of the responses. It doesn't take long at all and is very simple. Businesses, suppliers of services, e-commerce organisations, governments, and others might benefit from gathering diverse viewpoints and using them to supplement decision-making processes [4]. The massive volume of user-generated content on popular forums makes manual processing and evaluation of the material a tedious and time-consuming task. Methods for Automated Sentiment Analysis allow for quicker output, less human intervention, the removal of irrelevant data from large datasets, and the presentation of findings in applicable forms. For in-depth analytical operations on data extracted from a variety of sources (e.g., review sites, comment discussion boards, social media platforms, blogs, etc.), tools to analyse sentiment are invaluable [4, 5]. Though sentiment analysis has made great strides in recent years, the majority of studies have focused on English data. Languages beyond English have received much less attention from researchers. Analysing data that is not in English poses more serious issues. When it comes to resources, mechanisms that operate on various languages either have to make do with what they have or choose to translate to English and use what is widely accessible. This endeavour, called multi-lingual sentiment analysis, has a lot of room for improvement and is currently a relatively unexplored field of study. Using sentiment analysis on non-English languages, the authors survey the existing research. Because of their widespread use, machine learning techniques are the subject of this comprehensive review, which also includes information on the methods, tools, processes, and results. Given that the majority of sentiment analysis is conducted in English and that research into other languages is still in its early stages, this is meant to give readers a sense of the breadth and depth of the field's analysis and study while still leaving room for future exploration.

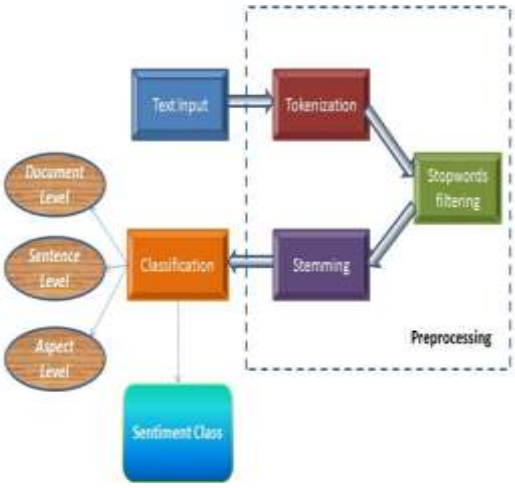

Figure 1 – The general approach for text summarization.

II. RELATED WORK

Conventional ML Methods for Summarising Text When it comes to text summarisation, traditional machine learning approaches mostly use extractive methods. These methods aim to find and pull out the most crucial lines or words from a document. In order to examine the structure and content of the text, these techniques depend substantially on statistical & feature-based methodologies. In order to determine how significant and relevant a sentence is, we employ characteristics like TF-IDF, word frequency, phrase position, & n-grams. To categorise or rank sentences using these attributes, algorithms like as Naive Bayes, Support Vector Machines (SVM), random forest algorithms, and K-Means clustering are often used. These approaches work well with smaller datasets in situations where computing resources are restricted because they are economically efficient and easy to apply. Nevertheless, they have a hard time deciphering the text's underlying semantic meaning and instead concentrate on superficial patterns, which might result in summaries that aren't always coherent or relevant to the context. More sophisticated, deep learning-based approaches have emerged to handle complicated and unstructured text data, whereas classical ML techniques have proved successful for extractive summarisation tasks [5]. The extraction of significant sentences or phrases from the original text is the backbone of conventional machine learning (ML) approaches to text summarisation. This allows for the condensing of information. These techniques don't rely on sentence generation but rather on extracting the most representative elements from the text. Sentence length, word frequency, position inside words (e.g., beginning sentences often being more relevant), presence of keywords or named entities, and frequency-inverse document frequency, or TF-IDF, are some of the statistical and linguistic features used to evaluate sentence importance. Machine learning methods like Naive Bayes, Support Vector Machines (SVM), Decision Trees, & Random Forests are used to rank or categorise phrases according to their relevance using these extracted attributes. Sentences are grouped into clusters using clustering methods such as K-Means. The purpose of this is to discover key sentences that reflect the document's primary ideas. TextRank and LexRank are graph-based algorithms that use measures of phrase similarity (like cosine similarity) to model sentences as a network with nodes representing sentences and edges representing the semantic links between them [9, 10]. After that, the most crucial phrases are chosen using centrality metrics like PageRank. Despite their computational efficiency and suitability for tiny datasets or structured language, these approaches often fail to grasp semantic subtleties or context. One example is when they don't pay attention to the text's logical progression or ignore the implied connections between concepts. The fact that these approaches usually rely on characteristics that have been hand-engineered and on established rules makes them less flexible when it comes to dealing with different domains or languages. Although classic ML approaches are easy to understand and work with, the summaries they provide aren't always as useful or coherent as those produced by more sophisticated models based on deep learning [6]. Lightweight, interpretable, & resource-efficient summarisation is still a desirable situation for these approaches. B. A method for text summarisation using deep learning Because of their superior ability to capture deeper semantic linkages and produce more coherent & contextually relevant summaries, deep learning techniques to text summarisation constitute a huge improvement over older methods. Included in these approaches are both abstractive and extractive techniques; a particular focus is on neural network topologies such as RNNs, LSTMs, GRUs, and, more recently, transformers models. For extractive summarisation, deep learning models prioritise phrases using embeddings or hierarchical

attention mechanisms like Word2Vec or GloVe, or contextualised representations like BERT. Seq2Seq models with attention processes have been extensively used for abstractive summarisation because they can generate new sentences can paraphrase the original text while keeping its meaning. By using pre-trained language models that have been enhanced for summarisation tasks, transformer-based models like BERT, GPT, T5, and BART have established state-of-the-art standards [7].

III. METHODOLOGY

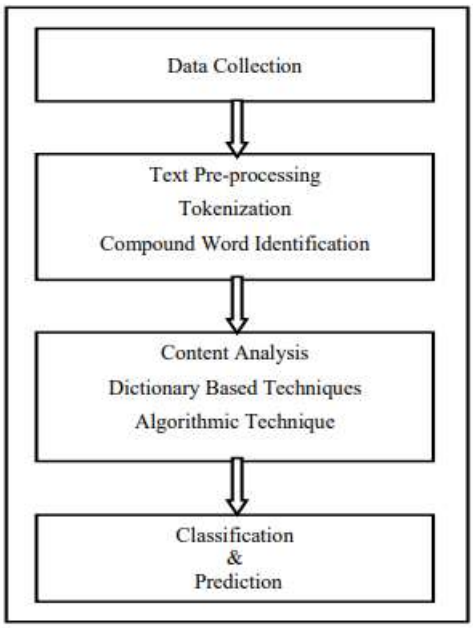

Figure 2 – Process flow of text processing

Figure 2 depicts the standard procedure for classifying unstructured text input using machine learning (ML) techniques. Gathering relevant data is the first and most important stage of any NLP (natural language processing) job. This process begins with the collection of raw data, usually in textual form, from a variety of sources such as social networks, websites, books, & articles. It might be difficult to handle this data since it is not always in a specified format (unstructured). Tasks such as sentiment analysis, text categorisation, and named entity identification rely on data generated in several NLP applications. Due to the fact that poor model performance may be caused by noisy or irrelevant data, the task's success depends on the data's quality and relevance. Web scraping, APIs, and open datasets are just a few examples of data collecting techniques that could change based on the use case. Text preprocessing is a set of procedures used to clean and standardise text prior to analysis after data collection. To make sure the text is machine-readable and to filter out any unwanted noise, pre-processing is a must. Some common steps in pre-processing texts include tokenisation, which involves splitting the text towards words or paragraphs, lowercasing, which means to standardise the text, removing stop words, stemming, which means to reduce words to their root form, and lemmatisation, which means to convert words to the underlying or dictionary form. Because it simplifies the text and improves the execution of NLP models, pre-processing is essential. Identifying and working with compound words, which are created by combining a number of phrases to establish a new meaning, is known as compound word identification. Compound words are widespread and need thorough processing in some languages. This is particularly true in German and Finnish, where word combinations are long and complicated. The problem is that regular tokenisers and models may not be able to tell these complex terms apart. Methods such as statistical modelling, morphological analysis, and dictionary search may be used to identify compound words. The accuracy of activities like text categorisation or machine learning is greatly enhanced by this phase, which treats the compound word as a single entity. Many natural language processing applications rely on methods derived from dictionaries. These methods find relevant spoken language, entities, or phrases in texts by consulting pre-existing dictionaries or lexicons. A sentiment lexicon may be used to categorise words as either neutral, negative, or positive in a sentiment analysis activity. Named entity recognition is one such job that makes use of a dictionary of well-known names, locations, or organisations to locate and categorise textual items. There

may be coverage restrictions with dictionary-based techniques, despite their efficiency and interpretability, since they can only identify words or things that have been already in the dictionary. To solve this problem, we may use hybrid approaches that combine dictionary-based methodologies with machine learning models, or we can dynamically extend dictionaries.

Lastly, text is sorted into predetermined classifications or labels via the categorisation process. This is a typical problem in natural language processing (NLP) and it entails training an ML model using labelled data, where the right category for each word is known. Decision trees, which are support vector machines (SVMs), and neural networks are some of the most used techniques for classification. The objective is to teach the model to spot trends in the data and properly classify hidden content. Many things go into making a classification model work, including picking the right features, having high-quality labelled data, and using the right algorithm. Some examples of classification problems in natural language processing include subject categorisation, sentiment analysis, and spam detection.

IV. RESULTS

A graphic illustration of written information in which words are shown with different sizes reflecting to their occurrence or importance within a certain corpus. You can easily identify prominent lexemes by looking at the size of a word, which represents its prominence, and how often or relevant it is in the dataset. To better understand the main ideas or concepts presented in a piece of literature, word clouds are often used in the analysis of texts, sentiment analysis, and topic modelling. They are a visually appealing and easily understandable way to condense large amounts of text, which is why you see them used a lot in presentations and analytical reports to quickly communicate key points (see picture 3). Word clouds are great for quick exploration but not for deep, interpretative research since they could hide complex semantic linkages or contextual subtleties between words.

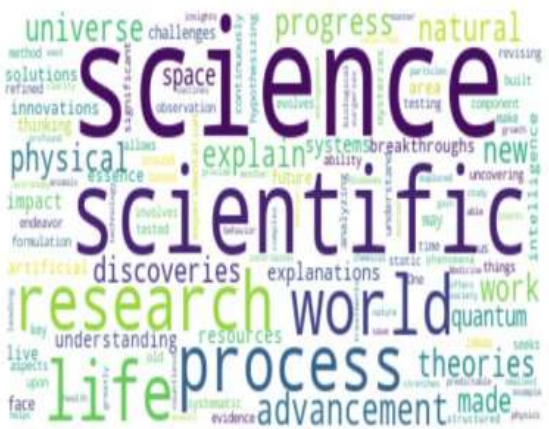

Figure 3 - Word cloud before text pre-processing.

A Top words graph is a way to show the most important or common words in a dataset or text corpus visually. In order to provide a clear picture of the text's main ideas, this graph usually shows the frequency or significance of keywords or concepts. NLP, or natural language processing, often employs it for activities like content analysis and topic modelling. Bar charts and word clouds are two examples of the graph types that may be used, with the size of each phrase being determined by how often it appears or weighted importance. Figure 4 shows how Top Terms graphs may help with information retrieval, text summarisation, and content categorisation by highlighting the most important ideas in a set of texts. To get textual data ready for research or model training, cleaning and preparation is the next step. Preprocessing sometimes involves deleting URLs, which gets rid of links that may not add anything useful to the research. Similarly, to make sure the text only includes essential words, it is frequently required to remove special characters like punctuation, symbols, or emojis. Standardising the text and removing noise from the findings is what this stage is all about. Step two is cutting out unnecessary words. These are terms like "the," "is," "and" and

" and "of" that are used a lot but don't really add anything. Most natural language processing (NLP) tasks are better served by ignoring stop words and concentrating on nouns, verbs, and adjectives. These preprocessing methods, when combined with others like as stemming, tokenisation, or lemmatisation, simplify text data, allowing computers to more easily extract useful insights or patterns.

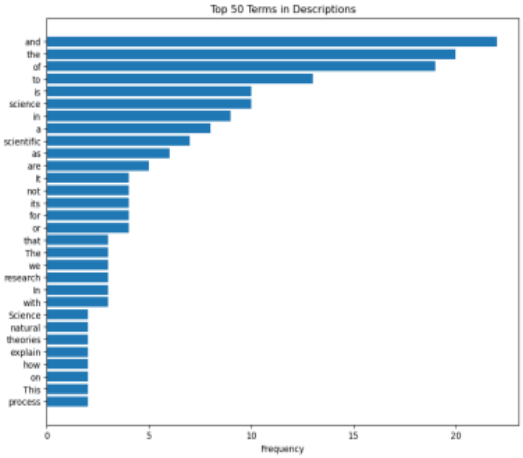

Figure 4 - Top 50 most frequent terms before text preprocessing.

V. CONCLUSION

In conclusion, the approach for summarising unstructured text data is fundamentally dependent on efficient preprocessing to convert raw, chaotic material into a structured format amenable to analysis and summarisation. Preprocessing employs methods such as tokenisation, stop word deletion, stemming, or lemmatisation to mitigate the obstacles posed by unstructured data, therefore diminishing complexity and enhancing the quality of input for summarisation models. Refining the text enables the extraction of significant information, identification of important topics, and the generation of cohesive, short summaries that maintain the core of the original content. The preprocessing step is essential for ensuring accuracy and relevance, while the selection of summarisation technique—either extractive or abstractive—significantly influences the efficiency and the quality of the output. As NLP technologies mature, improvements in preprocessing and summarisation techniques will become more vital for managing extensive unstructured data and facilitating more efficient content analysis.

REFERENCES

[1] Z. Li, X. Yu, T. Wei and J. Qian, "Unstructured Big Data Threat Intelligence Parallel Mining Algorithm," in Big Data Mining and Analytics, vol. 7, no. 2, pp. 531-546, June 2024, doi: 10.26599/BDMA.2023.9020032.

[2] S. Liu et al., "Multimodal Data Matters: Language Model PreTraining Over Structured and Unstructured Electronic Health Records," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 504-514, Jan. 2023, doi: 10.1109/JBHI.2022.3217810.

[3] Y. Seo, J. Park, G. Oh, H. Kim, J. Hu and J. So, "Text Classification Modeling Approach on Imbalanced-Unstructured Traffic Accident Descriptions Data," in IEEE Open Journal of Intelligent Transportation Systems, vol. 4, pp. 955-965, 2023, doi: 10.1109/OJITS.2023.3335817.

[4] Gowda, Madhu Belur Gopala, Naveen Kumar Boraiah, Varun Eshappa, and Gopala Krishna Chandra Shekara. "Classification of Epileptic EEG Signals Using Improved Atomic Search Optimization Algorithm." International Journal of Intelligent Engineering & Systems 16, no. 6 (2023).

[5] K. Adnan, R. Akbar and K. S. Wang, "Towards Improved Data Analytics Through Usability Enhancement of Unstructured Big Data," 2021 International Conference on Computer & Information Sciences (ICCOINS), Kuching, Malaysia, 2021, pp. 1-6, doi: 10.1109/ICCOINS49721.2021.9497187.

[6] Y. Asim, A. K. Malik, B. Raza, A. R. Shahid and N. Qamar, "Predicting Influential Blogger's by a Novel, Hybrid and Optimized Case Based Reasoning Approach With Balanced Random Forest Using Imbalanced Data," in IEEE Access,

vol. 9, pp. 6836-6854, 2021, doi: 10.1109/ACCESS.2020.3048610.

[7] M. Elsayed, A. Abdelwahab and H. Ahdelkader, "A Proposed Framework for Improving Analysis of Big Unstructured Data in Social Media," 2019 14th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2019, pp. 61-65, doi:10.1109/ICCES48960.2019.9068154.

[8] A. Moldagulova and R. B. Sulaiman, "Document Classification Based on KNN Algorithm by Term Vector Space Reduction," 2018 18th International Conference on Control, Automation and Systems (ICCAS), PyeongChang, Korea (South), 2018, pp. 387-391.

[9] E, V., & Pushpa Ravikumar, D. (2018). Attribute Selection for Telecommunication Churn Prediction. International Journal of Engineering & Technology, 7(4.39), 506- 509. https://doi.org/10.14419/ijet.v7i4.39.24364

[10] E. Varun and P. Ravikumar, "Community Mining in Multi-relational and Heterogeneous Telecom Network," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, 2016, pp. 25-30, doi: 10.1109/IACC.2016.15.