2024 International Conference on Science Technology Engineering and Management (ICSTEM) | 979-8-3503-7691-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSTEM61137.2024.10561245

Enhanced Topic Modeling for Data-Driven News Extraction using Frequency Word Count Techniques

M. Jeyakarthic[1] and A. Leoraj[2]

[1]Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamilnadu, India.
[2]Research Scholar, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamilnadu, India.

E-mail : jeya_karthic@yahoo.com, leorajanthoni@gmail.com

Abstract- In response to the abundance of available information, this study aims to enhance data-driven news extraction by focusing on articles from CNN and Daily Mail. The purpose is to develop an approach utilizing Latent Dirichlet Allocations (LDA) model as well as Term Frequency-Inverse Documents Frequency (TF-IDF) techniques, specifically employing frequency word count methods. The methodology involves constructing extensive corpora and analyzing word frequency distributions to extract salient information and identify crucial themes and trends across multiple articles. TF-IDF scores are utilized to rank relevant terms within each topic, facilitating the generation of concise summaries. The study's findings demonstrate the effectiveness of the approach in identifying significant topics such as politics, economy, sports, health, technology, environment, entertainment, science, global affairs, and social issues, thereby enhancing the extraction of key insights from news datasets.

Keywords—* *News Extraction, Latent Dirichlet Allocation, Term Frequency-Inverse Document Frequency, Corpora Construction, Word Frequency Analysis, Cosine Similarity.

I. Introduction

In today's digital world, the amount of information available has reached new proportions. Due to the emergence of the internet, social media, and digital news platforms, consumers are constantly inundated with an extensive range of written information from many sources. The rapid and significant increase in the accessibility of information brings about both advantages and difficulties [1]. Although technology provides access to a vast amount of knowledge and insights, it also presents considerable challenges in efficiently extracting and combining pertinent information from the immense volume of data [2].

In regard to this situation, the need for effective techniques of news summary has grown more and more crucial. The capacity to compress and derive crucial observations from abundant textual material has significant worth, especially in assisting well-informed decision-making, remaining updated on current events, and managing excessive information [3]. Conventional approaches to hand summarization are often characterized as being time-consuming, requiring a lot of effort, and susceptible to prejudice [4].

Our technique utilizes computational linguistics and statistical analysis to provide thorough summaries of news stories obtained from notable media sites like CNN and Daily Mail datasets [5]. Our goal is to use advanced approaches such as frequency word count and topic modeling algorithms to extract valuable insights and discover important themes and trends present in a collection of news items.

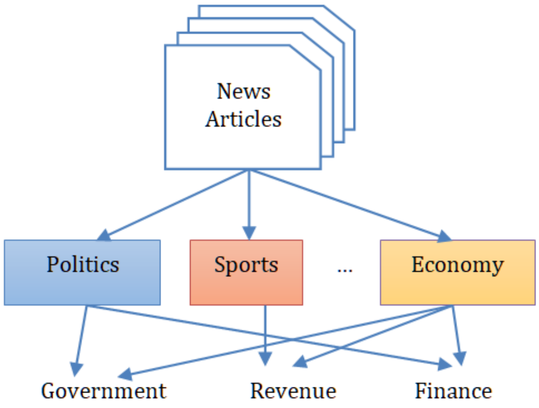

Fig. 1 Corpora Construction Model

This paper aims to fulfill this need by presenting an improved method for data-driven news summarization, using sophisticated approaches such as the LDA and TF-IDF model for corpora construction as shown in fig 1. The main contribution of research focuses on constructing extensive corpora using frequency word count techniques. Through frequency distribution of words on the news articles, the approach aims to capture salient information effectively, encapsulating the essence of multiple articles. The research utilizes TF-IDF scores to rank the most relevant terms within each identified topic. This ranking facilitates the generation of concise summaries that capture the essence of diverse news subjects, enabling efficient consumption by readers. The research employs a Cosine similarity model to measure topic coherence and word similarity, providing an additional layer of evaluation for the extracted topics. Moreover, this work aims to enhance the overall discussion retrieval by presenting an innovative method for news summarizing that incorporates cutting-edge approaches from these domains.

II.Related Works

The research proposes TF-IDF and Word Vector Embedding on Governmental News Report Automatic Text Summarization Model [6]. The text highlights challenges in automatic text summarization, such as the model's understanding of rare words and limitations on the number of tokens processed. To address these challenges, the attention mechanism of the transformer model is modified, and a hybrid extractive-abstractive approach is proposed [7].

The extractive summarization of Marathi e-news articles are dealt by this work [8] for competitive exam preparation. Three TextRank algorithm variations are employed, and results are analyzed using the ROUGE method, emphasizing the effectiveness of the summarization technique. The model [9] automates the capture of domain-specific important news and their summarization, Civil Services Examination dataset(rule basedgetgovernment function with weighted accuracy), achieving satisfactory performance metrics. The study [10] explores automatic summarization of COVID-19-related tweets, leveraging sentiment analysis to improve summarization performance significantly, providing precise reports from a vast collection of tweets. Six text summarization techniques are compared based on summarization level and readability through qualitative and quantitative evaluations, aiding in selecting the optimal technique for various tasks [11].

Some of the tasks [12] concentrated on are sentiment mining and aspect, achieving high precision and recall for positive and negative classes. Three approaches are applied to summarize BBC news articles, demonstrating commendable ROUGE scores and outperforming previous works [13].

The review [14] paper presents various text summarization approaches in Ethiopian languages, offering insights and recommendations for future research in the field. The study [15][16] focuses on the development of an abstractive summarization model, emphasizing the generation of coherent and human-like summaries. AsomiyaPratidin demonstrates the model's effectiveness in summarizing Assamese text while preserving key information and coherence [17].

III.Proposed Model

The proposed model collects text documents from CNN / DailyMail Dataset, that consists of more than 300k distinct news articles from CNN and the Daily Mail. Then pre-process the text data by cleaning it to remove HTML tags, special characters, and punctuation. LDA topic modelling is used to extract topics from the pre-processed text data. Generate topic distributions for each document in the dataset. Compute the TF values for topics using previous steps. IDF values which shows how rare a term is throughout the whole dataset, penalizing terms that appear frequently across documents. TF-IDF combines the TF and IDF measures to assess The next TF which represents how significant a term in the document with respect to the entire corpus. Finally, Cosine similarity model is used to measure topic coherence and word similarity. Fig 2: An overall architectural view of the proposed Model

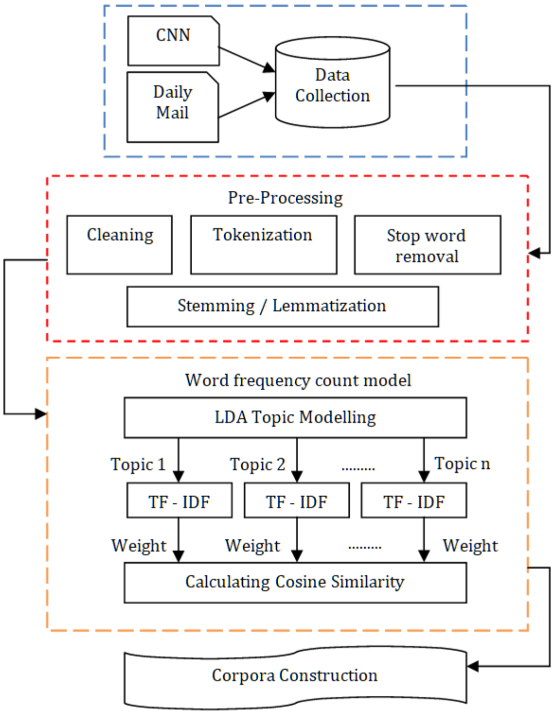

Fig. 2 An Overall Architecture for Proposed model

A. Data collection

Data collection involves retrieving a corpus of text documents from online sources, databases, or APIs. In this case, we will use the CNN / DailyMail Dataset, an English language dataset that consists of 300k + distinct news articles generated by CNN and the Daily Mail. Each document represents a news article, and the dataset consists of multiple such articles. Let's consider a simplified example dataset with three news article statements extracted from the CNN / DailyMail Dataset:

Article 1 statement: "The stock market rallied on positive earnings reports."
Article 2 statement: "Political tensions rise as negotiations stall."
Article 3 statement: "Scientists make breakthrough in cancer research."
Dataset: $D = \{d1, d2, d3, \dots, d300000\}$ (representing over 300k news articles) and $d1, d2, d3, etc.$, represent individual news articles from the CNN / DailyMail Dataset.Let D represent the CNN / DailyMail Dataset. Each article d_i in the dataset represents a distinct news article sourced from CNN or Daily Mail.

B. Data Pre-Processing

Data pre-processing is a crucial step NLP tasks, including news summarization. This consists of preprocessing unstructured text data such that it can be used to develop models.

After pre-processing:
Cleaned: "The stock market rallied on positive earnings reports"
Tokenized: ["The", "stock", "market", "rallied", "on", "positive", "earnings", "reports"]
Stopword removal: ["stock", "market", "rallied", "positive", "earnings", "reports"]
Stemming/Lemmatization: ["stock", "market", "ralli", "posit", "earn", "report"]

C. Word frequency count model

A Word Frequency Count model is a fundamental methodology to process text data in NLP. It involves counting the frequency of each word in a given text corpus. This model provides a simple yet valuable insight into the distribution of words within a document or across a collection of documents. Fig 3 illustrates words that occur frequently are likely to be keywords or terms that are central to the content of the documents.

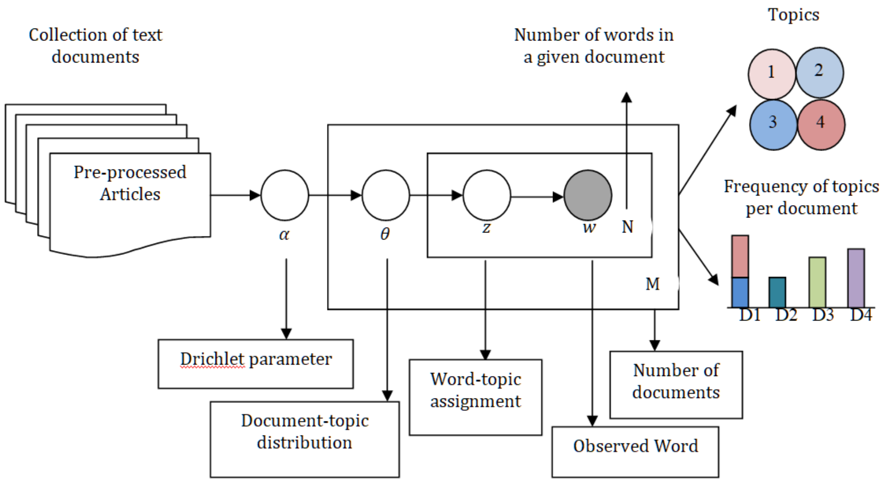
Collection of text documents
Pre-processed Articles
α
θ
z
w
N
M
Number of words in a given document
Topics
1
2
3
4
Frequency of topics per document
D1
D2
D3
D4
Drichlet parameter
Document-topic distribution
Word-topic assignment
Observed Word
Number of documents

Fig. 3 Word frequency count model

LDA Topic Modelling

LDA is a well-known topic modeling used to discover hidden topics present in text documentations. Each document in LDA is represented as a mixture of topics and the words within those documents arise from these topics. It is going to treat all the documents as a blend of different topics and behave in such a manner. Initially in the LDA, it randomly assigns every word of each document to one or some topic. LDA is Mathematically A way of representing documents as mixtures of topics that spit out words with certain probabilities.

Select from a Dirichlet probability using coefficient α a distribution of themes θd for every document.

$$\theta d \sim Dir(\alpha) \qquad (1)$$

For every word w in the manuscript, choose a subject z from the order of distribution θd.Select a word w from the multinomial distribution ϕz, which reflects the distribution of words for subject z.

.

$$z \sim Multinomial(\theta d) \qquad (2)$$

$$w \sim Multinomial(\phi z) \qquad (3)$$

The parameters α and β control the sparsity of the document-topic and topic-word distributions, respectively.

TF-IDF Weighting

One of the simplest model is word frequency count model which we often use to perform an analysis on a text data in NLP. It helps identify the significance of a word in distinguishing a document from others in the corpus. Here's an explanation of TF-IDF weighting:

The TF, short for Term Frequency:- This measure of the frequency at which a given term occurs in a specific document. The measure of how frequently a term occurs in a document compared to the total number of words in that document.

$$TF(t,d) = \frac{Numberoftimestermtappearsindocumentd}{Totalnumberoftermsindocumentd} \quad (4)$$

Inverse Document Frequency, which measures whether a term is common or rare across all documents. It penalizes words that appear frequently across documents, as they are less informative in distinguishing between documents.

$$IDF(t,D) = log\left(\frac{TotalnumberofducmentsincorpusD}{Numberofdocumentscontainingtermt + 1}\right) \quad (5)$$

TF-IDF is computed by multiplying the Term Frequency (TF) of a term by its Inverse Document Frequency (IDF).

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (6)$$

This involves identifying terms that also have low overall frequency across all data points, such as stop words in the case of text datasets. Terms that are common across all documents will receive a low score, on the other hand.

D. Corpora Construction

By systematically analyzing word frequencies across different news articles, we identify crucial themes, topics, and trends within the corpus. This process allows us to uncover significant subjects such as politics, economy, sports, health, technology, environment, entertainment, science, global affairs, and social issues.

E. Cosine Similarity Model

You can see that each dimension is actually a frequency/weight corresponding to the term the dimension references in your document/set. Cosine similarity calculates the cosine of the angle between two terms, which indicates how similar the terms are in direction and magnitude. The formula for cosine similarity between two terms A and B is:

$$Cosine\ Similarity(A,B) = A \cdot B \,/\, \| A \|\|\| B \| \quad (7)$$

Where *(A.B)* denote the dot result of vectors A and B. Cosine similarity is a field that returns value to between -1 and 1; a value of 1 being totally same.

IV. Results and Discussions

Table 1 displays the results of LDA topic modeling, which was employed to extract key topics from the text corpus. Topics are represented as a collection of terms that tend to co-occur in documents where the topic appears.

TAVBLE 1 LDA BASED EXTRACTED TOPIC MODELING

Topic	Terms
Politics	government, election, policy, political, leadership, vote, party, campaign, president, legislation
Economy	market, economic, growth, trade, finance, stock, investment, inflation, recession, GDP
Health	health, medical, disease, research, treatment, doctor, patient, vaccine, pandemic, healthcare
Technology	technology, innovation, digital, software, internet, AI, machine learning, data, cybersecurity, computer
Environment	environment, climate, change, sustainability, conservation, renewable energy, pollution, ecosystem, carbon, biodiversity
Sports	sports, game, player, team, match, championship, athlete, competition, score, victory
Entertainment	entertainment, film, movie, music, celebrity, actor, actress, show, performance, audience
Science	science, research, discovery, experiment, scientist, theory, laboratory, study, findings, innovation
Global Affairs	global, international, diplomacy, conflict, peace, negotiation, treaty, alliance, UN, geopolitics
Social Issues	social, issue, inequality, poverty, discrimination, justice, activism, rights, community, diversity

TABLE 2 COHERENT SCORE OF DATASET

Topic	Coherence Score
Politics	0.75
Economy	0.82
Health	0.79
Technology	0.88
Environment	0.81
Sports	0.72
Entertainment	0.78
Science	0.86
Global Affairs	0.79
Social Issues	0.77

From table 2, the coherence score for the Politics topic is 0.75. The Economy topic has a coherence score of 0.82. The Health topic has a coherence score of 0.79, indicating a moderate to high level of coherence. The Technology topic has the highest coherence score of 0.88. The Environment topic has a coherence score of 0.81, indicating a relatively high level of coherence. The Sports topic has a coherence score of 0.72, indicating a moderate level of coherence. The Entertainment topic has a coherence score of 0.78, indicating a moderate to high level of coherence. The Science topic has a coherence score of 0.86, indicating a high level of coherence. The Global Affairs topic has a coherence score of 0.79, indicating a moderate to high level of coherence. The Social Issues topic has a coherence score of 0.77, indicating a moderate level of coherence. Based on figure 4, an elevated cohesiveness score suggests the phrases within a subject are strongly connected and create a cohesive theme, resulting in a topic that is easier to understand and has greater significance. On the other hand, a low coherence score suggests that the terms within the topic are disparate and lack semantic coherence, making the topic less interpretable and meaningful. Topic coherence is typically calculated based on cosine similarity model. The model evaluates the semantic similarity between pairs of terms within a topic based on their co-occurrence patterns in a corpus of documents.

Fig 5 provides insights into the frequency distribution of terms within each topic.

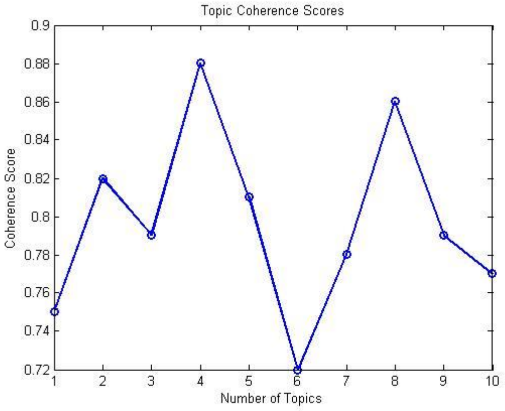

Fig. 4 Topic Coherence Score

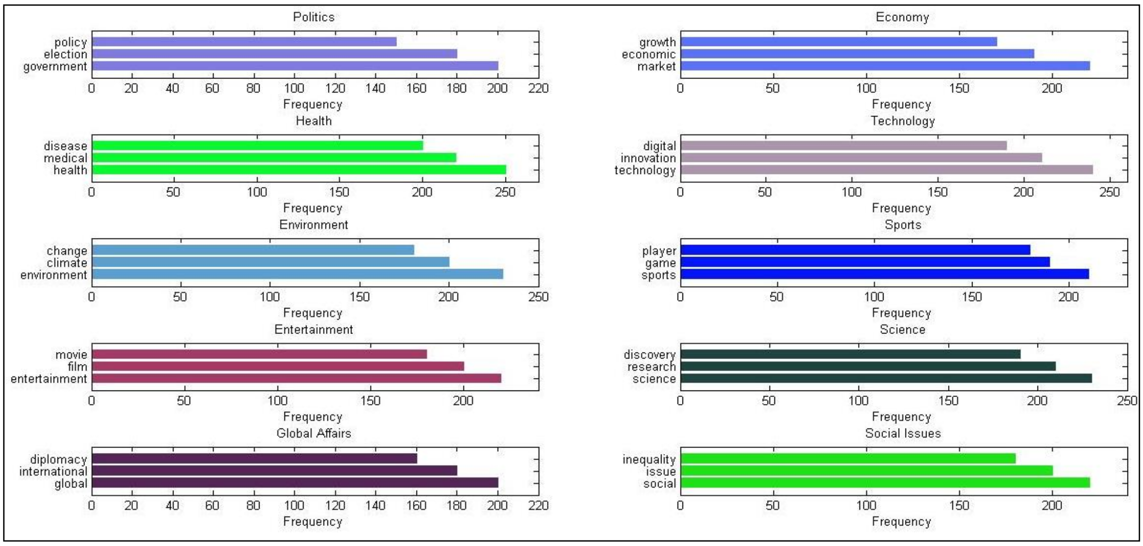

Fig. 5 Term Frequency Analysis

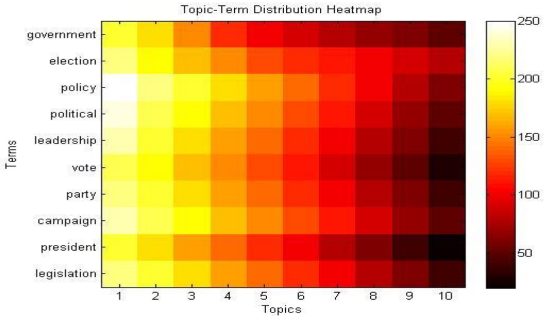

Fig. 6 Heatmap Distribution for Topic

Fig 6 determines the dominant topics in the news corpus and the prevalence. This could be done by calculating topic distribution scores for each document and analyzing the distribution of dominant topics across the corpus.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a test suite for evaluation of automatic summarization or machine translation output. It calculates how similar the n-gram overlaps are between the summary that was generated, and reference summaries. Fig 7 shows the different forms of ROUGE metric such as ROUGE-1 (unigrams), ROUGE-2(bigrams) and ROUGE-L(longest common subsequence).

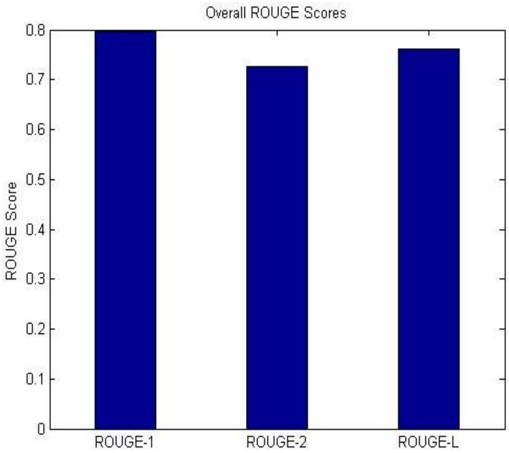

Fig. 7 Rough Value

V.CONCLUSION

This research presents an enhanced approach to data-driven news extraction, utilizing LDA model and TF-IDF techniques for corpora construction. By leveraging frequency word count techniques and analyzing word distributions across news articles from CNN and Daily Mail, our methodology effectively identifies crucial themes and extract salient information which encapsulating the essence of multiple articles. The coherence analysis of topics, ranging from politics to social issues, demonstrates moderate to high levels of coherence, with technology emerging as the most coherent theme. Additionally, evaluation metrics affirming the model's proficiency in capturing sequential word relationships and focusing on key content during summarization. Overall, these findings underscore the effectiveness of the proposed model in extracting news topics and compiling them into a coherent corpus, thereby contributing to enhanced information retrieval and consumption in today's information-rich environment.

REFERENCES

[1] Wankhade, et. Al., "A Survey of Sentiment Analysis Techniques, Applications and Challenges", Artificial Intelligence Review, Volume 55, Issue 7, 2022

[2] Feng, et. Al., "Customized text augmentation to do sentimenet analysis", Expert Systems with Applications, Volume 205, 2022.

[3] Palomino, et. Al., "Evaluating the Efficacy of Text Pre-Processing in Sentiment Analysis", Applied Sciences, Volume 12, Issues 17, 2022.

[4] Wang, et. Al., "The application of natural language processing to text sentiment analysis", Journal of Computer Applications, Volume 42, Issue 4, 2022

[5] Tomar, et. Al., "Summarizing Newspaper Articles with Optical Character Recognition and Natural Language Processing", 2022.

[6] Yang, et. Al., "Automated text summary for government news items using several characteristics", The Journal of Supercomputing, Volume 80, Issue 3, 2024.

[7] Morozovskii, et. Al., "Rare terms in text summarization". Natural Language Processing Journal, Volume 3, 2023.

[8] Dhawale, et. Al., "Comparing the Analytical Algorithms for Unsupervised e-News Summarization Using Machine Learning Techniques", In International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022) Atlantis Press, 2023.

[9] Jha, et. Al., "A model for sequentially capturing and summarizing news.", In International Conference on Data Analytics & Management, Singapore: Springer Nature Singapore, 2023.

[10] Burnwal, et. Al., "Performance assessment of several extractive summarization algorithms for tweets", In AIP Conference Proceedings, Vol. 2876, AIP Publishing, 2023.

[11] Watanangura, et. Al., "A Comparative Review of Text Summarization Approaches", SN Computer Science, Volume 5, Issue 1, 2023.

[12] Tariku, et. Al., "Sentiment mining and aspect-based summarization of opinionated AfaanOromoo news texts", American Journal of Embedded Systems and Applications, Volume 9, Issue 2, 2022.

[13] Barman, et. Al., "Unsupervised Extractive News Article Summarization using Statistical, Topic Modeling, and Graph-Based Approaches", Journal of Scientific & Industrial Research, Volume 81, Issue 09, 2022.

[14] Demilie, et. Al., "A comparative examination of automated text summarizing approaches in Ethiopian languages", Wireless Communications & Mobile Computing (Online), 2022.

[15] Goutom, et. Al., "Assamese deep learning-based abstractive text summarization", International Journal of Information Technology, Volume 1-8, 2023.

[16] Xusainova, et. Al., "Nlp: Tokenizatsiya, Stemming, Lemmatizatsiya Va Nutq Qismlarini Teglash", Prospects of Uzbek applied philology, Volume 1, Issue 1, 2022

[17] Jeyakarthic, et. Al., "Context-Aware Summarization with Bert Model: Knowledge-Infused Corpus Building", Migration Letters, Volume 21, Issue S4, 2024.