Pakistan Journal of Scientific Research, PJOSR
ISSN (p): 0552-9050
Volume: 4, Number: 1, Pages: 75- 83, Year: 2024

Topic Modeling for Urdu Articles Using Unsupervised Learning Approaches

Muhammad Ashir[1], Ali Saeed[2], Muhammad Farhat Ullah[1,3], Syed Nisar Ali, Muhammad[4], Sauood[5,6], Mehmood Anwar[1], Naveed Hussain[2] and Shaukat Ali[7]

[1] Department of Software Engineering, FOIT, University of Lahore, Punjab, Pakistan
[2] Department of Software Engineering, University of Central Punjab, Lahore, Pakistan
[3] School of Software, Dalian University of Technology, Dalian, Ganjingzi District, Liaoning Province, China
[4] Department of Computer Science, FOIT, University of Lahore, Punjab, Pakistan
[5] Department of Computer Science, Bahria University, Lahore, Pakistan
[6] Faculty of Computer Science & Mathematics, Universiti Malaysia Terengganu, Malaysia
[7] Office of the Registrar, GC University Lahore, Pakistan

Corresponding author: Ali Saeed (e-mail: ali.saeed@ucp.edu.pk).

Received: 02/07/2024, Revised: 07/08/2024, Accepted: 13/08/2024

Abstract- **Topic modelling is a commonly used text-mining tool for discovering hidden semantic structures within a text corpus. This paper introduces an unsupervised learning-based topic modelling approach for Urdu documents, a language with limited resources. Specific and accurate topics are extracted from Urdu texts using unsupervised learning techniques such as Latent Dirichlet Allocation (LDA) and Unsupervised Latent Semantic Indexing (ULSI). The experimental results illustrate our recommended ULSI and LDA models' dominance, achieving 99% and 98% accuracy and 44% and 37% coherence values in LDA and ULSI, respectively. The experimental results demonstrate the superiority of the proposed ULSI and LDA models, which achieve high accuracy and coherence values.**

Index Terms-- **Local Dirichlet Allocation (LDA), Urdu Latent Dirichlet Allocation (ULSI), Prediction, Natural Language Processing (NLP).**

I. INTRODUCTION

Urdu is a language with rich morphology but limited resources, as noted by [1]. It serves as the national language of Pakistan and is used by over 300 million people worldwide for communication [2] [3] [4]. Processing Urdu poses challenges due to its complex grammar, extensive word derivations, and inflections. Urdu has received relatively little research attention in the fields of NLP and information retrieval (IR) [5], given its relatively recent entry into these domains. Many models and techniques developed for other languages cannot be directly applied to Urdu due to its distinct language structure, as discussed by [3]. However, these existing approaches primarily focus on document classification and lack emphasis on unsupervised learning techniques, which is the primary objective of this research in exploring unsupervised approaches for Urdu topic modeling.

Topic modeling [6] is a technique used in natural language processing to uncover latent themes within a collection of documents. It helps to identify underlying patterns and associations among words, enabling a deeper understanding of the data [7]. The primary goal of topic modelling is to provide machine-intelligible conceptual explanations for text contents in order to extract knowledge rather than separate relevant data.

Topic modeling involves analyzing a specific set of textual data to uncover underlying issues. It utilizes feature extraction and selection techniques based on correlations between subjects and words, as well as in between words, to reveal hidden topics within a document [8].

To perform topic modeling and annotation, a diverse range of models are employed. These models include Latent Semantic Analysis (LSA) [9], which explores the relationships between words and documents to discover underlying topics. Probabilistic Latent Semantic Analysis (PLSA) [10] extends LSA by introducing a probabilistic framework for topic modeling. The k-means algorithm [11] clusters documents based on similarity, aiding in the identification of topic groups. Ridge Classifier [12] and Linear SVC models utilize machine learning techniques for document classification and topic prediction. Lastly, Latent Dirichlet Allocation (LDA) [13] is a popular generative probabilistic model that assigns topics to documents based on the distribution of words. The utilization of

these models empowers researchers and practitioners to unravel hidden topics and annotate textual data effectively.

The researcher has conducted an extensive study on various aspects of topic modeling [6] with the aim of achieving the highest possible accuracy in result prediction. Other researchers are also exploring different models and frameworks for extracting specific topics from text documents, focusing on individual models or frameworks. In this particular study, the researcher applied unsupervised learning techniques, namely ULSI [10] and LDA, to predict topics with the objective of maximizing accuracy. The primary goal of this study is to propose a new framework for topic modeling, leveraging unsupervised learning techniques to optimize result accuracy. Subsequently, the study intends to compare these results with the techniques used and the findings of other researchers.

In this study proposed a new framework specifically designed for extracting topics from multiple Urdu language documents, with a particular focus on Urdu news text. To enhance result accuracy, this study employs unsupervised learning algorithms, namely ULSI (Urdu Latent Semantic Indexing) and LDA (Latent Dirichlet Allocation). Fig. 1 illustrates the scope of the research, providing a visual representation of the domain under investigation.

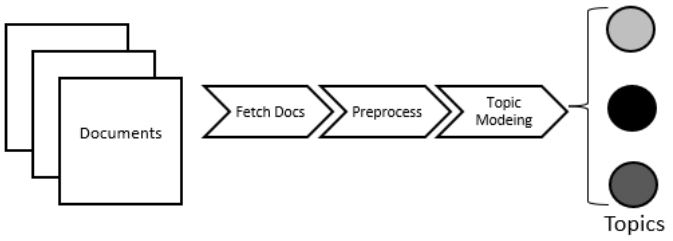


FIGURE 1. Scope of this research article

II. RELATED WORK

This literature encompasses languages such as English, Chinese, Arabic, Hindi, Romanian, and Urdu, which have been developed for topic modeling. Primarily, topic modeling has been utilized to delve into unknown article topics, while document clustering, a widely used technique in data mining literature, is also explored in this study. Various techniques including K-means, LDA (latent Dirichlet allocation), SVM, ULDA, Gibbs Modeling, and others have been employed for clustering purposes.

In a research work on topic modeling using both unsupervised and supervised approaches to obtain results for topic modeling [14]. They obtained data from various platforms, extracted through the Urdu corpus, comprising a total of 129 documents. The authors utilized the LDA and Naïve Bayes algorithms to extract the data and predict the most suitable topic for each document. The results obtained through these techniques showcased an accuracy of over 90%, as reported by the authors.

The authors conducted an empirical study on topic modeling [15]. They collected data from newspaper articles as the primary source, with the dataset for topic extraction comprising 6000 documents from newspapers. The authors utilized an unsupervised technique, specifically LDA, to analyze the data. The reported results in this study demonstrated an accuracy of 92% using the F1 measure. This research focuses on the Urdu language and contributes to its increasing reliability in forensic analysis.

Another research [16] employed unsupervised learning approaches to explore topic modeling and obtain results. The authors collected data from various sources, specifically Urdu tweets. The dataset utilized for their experiments consisted of 0.8 million words from this dataset. The researcher implemented LDA, NMF (non-negative matrix factorization), and LSA (latent semantic analysis) models to predict the most appropriate topic for each section of the document. The results obtained by the author using these techniques showcased 71% accuracy for LDA (mix-up) and 80% accuracy for NMF in Urdu tweet documents.

Another study proposed a framework for Urdu topic modeling using latent Dirichlet allocation (LDA) [17]. The researcher collected data from various sources such as BBC, Nawa-i-Waqt, Jang, and added additional news sources from Express-News in the second phase of the experiment. In the first experiment, the dataset consisted of 13,642 articles, while in the second experiment, the dataset increased to 21,341 articles. LDA and Gibbs modeling techniques were applied for topic modeling. The results showed 76% accuracy in LDA for the first experiment and a significant improvement to 78% accuracy in ULDA for the second experiment, indicating enhanced performance in topic modeling compared to the previous one.

This research [18] is centered on the Urdu language[1], and data is gathered from BBC Urdu[2], Urdu Point[3], and Urdu Geo[4]. The dataset comprised 10,000 general documents. The author extracted the data and applied the top-ranked algorithm along with TF-IDF (term frequency-inverse document frequency) for topic modeling. The results obtained by the author using these techniques were reported to be 0.88 or 88% accuracy.

In [19], the authors introduce an Urdu News Article Recommendation Model leveraging Natural language Processing Techniques. Their research utilizes unsupervised learning techniques to delve into topic modeling and obtain results for their study.

To conclude, it is stated that researchers have focused on various topic modeling research areas but have not given special attention to topic modeling using unsupervised techniques, particularly LDA and ULSI. To fill this gap, this study provides an empirical evaluation for topic modeling using LDA and ULSI and achieved state of the art results.

[1] express.pk Last visited 12 July 2023

[2] bbc.com/Urdu/ Last visited 12 July 2023

[3] urdupoint.com Last visited 12 July 2023

[4] Urdu.geo.tv Last visited 12 July 2023

III. Research Methodology

This study addresses three research questions, and the selection of techniques is based on their suitability in answering these questions. The first research question (RQ1) focuses on investigating previous studies to obtain existing results. The second research question (RQ2) pertains to developing a new framework aimed at achieving maximum accuracy in topic modeling. The third research question (RQ3) aims to improve the accuracy of topic modeling specifically for title predictions, utilizing a proposed model. The study is conducted through experiments and evaluation of selected datasets, which have been collected from various sources.

This section provide a description of the suggested unsupervised learning approach for topic prediction. The proposed model, as depicted in Fig. 2, consists of several steps. The first step, pre-processing of the collected text data. In the second step, the pre-processed text data is fed into the models for training. The models used for training are ULDA and LDA. The third step involves the topic prediction process, where the goal is to select the most important and relevant text from a document. In the fourth step, topics are identified and organized in a clustering form. Finally, in the fifth step, a topic is selected for the target document.

In software engineering, there are typically four types of research: scientific, engineering, empirical, and analytical [20]. During the first stage, it is crucial to acknowledge the significance of the scientific method in order to recognize the importance of the problem at hand before proposing a model. In the second stage, the proposed model is validated using appropriate methodologies to test and prove the hypotheses. In the final stage, the process is replicated to the best extent possible, ensuring its reproducibility and reliability.

Alternatively, in engineering methods, existing solutions are analyzed to develop a new solution, which is then tested to validate hypotheses [20] [21]. In the empirical method, the initial step involves model development, followed by the establishment of statistical or qualitative methods to validate the given hypotheses. Finally, the proposed model is applied to case studies for evaluation, and this process is repeated iteratively. On the other hand, in the analytical method [22], a formal theory is formulated first, and then the proposed theory is used to derive results. If feasible, these findings are compared to empirical observations to establish their validity.

Based on the aforementioned types, this study has opted for the engineering method. In the engineering approach, the focus lies on constructing solutions that are more accurate and effective compared to existing ones, with a strong emphasis on design. Therefore, applying the engineering technique is a suitable choice for this research. The objective of this study is to develop a model and enhance the accuracy of existing results in the field of topic modeling.

In the first step of this study is to conduct a thorough review of existing literature to examine their solutions. This involved investigating developed models, techniques, and methods for topic modeling of Urdu articles. In the second step, after an extensive study, it proposed a new model to assess the accuracy of topic modeling using different evaluation metrics. In the third step, researcher evaluated the proposed model using the given data to predict the document titles in topic modeling. This study applied unsupervised learning approaches to maximize the improvement in accuracy. In the final step, the comparison of this research solution with existing ones.

In this research article the dataset is of one million words contain news article from various domains like politics, sports, commode, and finance etc. Complete dataset and code can be downloaded from link[5].

In natural language processing tasks, having a benchmark dataset is crucial. It is important to have a substantial amount of pre and post-labeled data to effectively train the models. Additionally, for Urdu language processing, it is necessary to have linguistic resources specific to Urdu. However, Urdu lacks significant linguistic resources for NLP tasks [23]. For this experiment, the dataset was extracted from different sources, including BBC Urdu news, Urdu Point, and various social media platforms. The dataset comprises 1 million words and encompasses diverse classes such as political news, international news, sports, arts, and others.

Data analysis in this study involved the use of quantitative and statistical methods to analyze the data. These analyses were performed after applying data preprocessing techniques to ensure the data was in a suitable format for analysis.

A. PROPOSED METHODLOGY

In the proposed model, various aspects are utilized for data pre-processing. Data pre-processing is an essential part of natural language processing (NLP) operations, and it involves following important steps to process the data effectively. An example is shown in Fig. 3 from dataset.

It aims to convert the text into the most authentic format, enabling the algorithms to produce better results at each stage. To accomplish this, part of speech (POS) tagging is employed, which allows for syntax filtering to extract nouns, verbs, and sentence structures from the given text [16] [24] [25]. For this specific purpose, a well-known POS tagger [26] for the Urdu language is utilized. The dataset undergoes several different processes during the tag generation phase, including stop word removal, stemming, filtering out invalid characters, and segmentation, which will be discussed in the following subsections.

Natural language processing and text analysis heavily depend on data stemming and tokenization techniques. Stemming involves reducing words to their root form [3], which allows treating words with the same stem as identical. This process effectively reduces the dimensionality of the data and eliminates redundancy.

Tokenization plays a crucial role in text data processing as it involves breaking down text into individual words, phrases, or symbols. This step is essential because it enables the analysis of

[5] https://drive.google.com/drive/folders/1vSTSIIpWS8dPBrSWPELcWGj4j_l2uK5v?usp=sharing

individual characters and facilitates the examination of relationships between them.

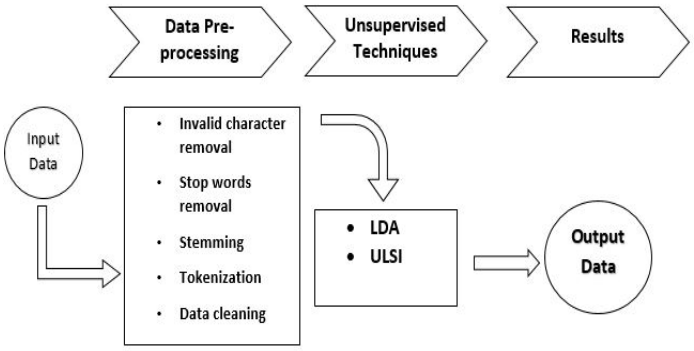


FIGURE 2. Proposed Model

Stemming and tokenization are essential processes in preparing text data for analysis and model development. This is used in applications for various scenarios, including sentiment analysis [27], text classification [28], and topic modeling. By reducing the complexity of text data, stemming and tokenization enable more accurate analysis and extraction of insights from large volumes of text data.

Moreover, in model development, unsupervised learning techniques such as Unsupervised Latent Semantic Indexing (ULSI) and Local Dirichlet Allocation (LDA) are implemented. Researchers can utilize evaluation measures such as Coherence value and F_1 score to assess the accuracy. These approaches assess the data and enable the extraction of the most accurate results. The following subsection discuss the model implementation.

B. IMPLEMENTATION

Model development is a crucial step in the realm of machine learning and artificial intelligence. Its objective is to design and test models that can perform specific tasks such as prediction, classification, and clustering. The primary aim of model development is to create a model that can generate accurate results based on input data and generalize well to new, unseen data. Once a model is created, it undergoes evaluation using various performance metrics like accuracy, precision, recall, and F_1 score. This evaluation helps to identify any inconsistencies in the model and, if necessary, make improvements. Model development is an ongoing process that involves careful consideration of the problem statement, available data, the selection of algorithms and hyper parameters. With the advancements in machine learning and artificial intelligence,

Urdu News	English Translation
اسلام باد عالمی بینک خیبرپختونخوا کے قبائلی اضلاع میں عسکریت پسندی سے پیدا ہونے والے بحران سے متاثرہ خاندانوں کی جلد بحالی بچوں کی صحت کی بہتری اور شہری مراکز ترسیل میں معاونت کے لیے فنڈز فراہم کرے گا اس منصوبے کے لیے عالمی بینک ملٹی ڈونر ٹرسٹ کے تحت ایک کروڑ 20 لاکھ ڈالر فراہم کرے گا جس کے تحت شہریوں کی سہولت کے مراکز قائم کیے جائیں گے جو پوری قبائلی بادی اور اس سے منسلک اضلاع کو خدمات فراہم کرے گا ذان اخبار کی رپورٹ کے مطابق یہ جولائی 2019 میں ایک کروڑ 50 لاکھ ڈالر کی منظوری کے بعد سے منصوبے کی تیسری فنانسنگ ہوگیہ بھی پڑھیں ئندہ سال تک پاکستان میں غربت میں اضافے کا امکان ہے عالمی بینکان سہولیات کے مراکز کے ذریعے منتخب خدمات فراہم کی جائیں گی جس میں وائٹل رجسٹریشن سروس وی ایس سول رجسٹریشن منیجمنٹ سسٹم سی ایم ایس اور نادرا ای سہولت شامل ہے جو اس سے مستفید ہونے والوں کی مزید سرکاری خدمات تک رسائی کو بڑھائے گا اس کے علاوہ ئندہ ماہ عالمی بینک کی اضافی فنانسنگ کی منظوری دی جائے گی جو گاہی سیشن میں حاضری سے منسلک نقدی کی منتقلی کے ذریعے بچوں کی صحت کو فروغ دے گیمنصوبے کی دستاویز کے مطابق اضافی سروسز مثلا وی ایس سی ایم ایس نادرا ای سہولت پلیٹ فارم اور دیگر سہولیات کے ساتھ اضلاع بنوں ڈیرہ اسمعیل خان لکی مروت اور ٹانک میں 16 نئے سہولت مراکز قائم کیے جائیں گےوی ایس میں کمپیوٹرائزڈ قومی شناختی کارڈ اور سی سیز کی تجدید اور اجرا سے متعلق تمام خدمات شامل ہوں گیمزید پڑھیں عالمی بینک سندھ میں چھوٹے ڈیموں کی تعمیر کے لیے 20 کروڑ ڈالر قرض دے گا اس کے علاوہ بلدیاتی حکومت یاکمشنر کے دفاتر کے تعاون سے سی ایم ایس متعارف کروائے سے شہری بالخصوص خواتین پیدائش شادی اور موت کے سرٹیفکیٹ حاصل کرسکیں	Islamabad The World Bank will provide funds for the early rehabilitation of families affected by the crisis caused by militancy in the tribal districts of Khyber Pakhtunkhwa, the improvement of children's health and support for delivery to urban centers. 2 million dollars will be provided under which citizen convenience centers will be established. According to the report of Dawn newspaper, this will be the third financing of the project after the approval of 15 million dollars in July 2019. Also read: Poverty is likely to increase in Pakistan by next year. which includes Vital Registration Service VS Civil Registration Management System CMS and NADRA e-facility which will increase the beneficiaries' access to more government services. Which promotes children's health through cash transfers linked to attendance at regular sessions. According to the project document, additional services such as VSCMS, NADRA e-facilitation platform and other facilities will be established along with 16 new facilitation centers in Bannu Dera Ismail Khan Lucky Marwat and Tank districts, computerized National Identity Cards and CCs in VS. All services related to the renewal and issuance of the game are included. Read the World Bank will lend 20 million dollars for the construction of small dams in Sindh. In addition, by introducing CMS with the support of the local government or commissioner's offices, the citizens, especially women, births and marriages. And get death certificates

FIGURE 3. A News article example obtained from dataset

Gensim is a Python library for topic modeling and natural language processing tasks, while Keras is a user-friendly interface to TensorFlow, enabling easy development of deep learning models. TensorFlow, a comprehensive machine learning framework, provides a wide range of functionalities for efficient model development and deployment.

model development is becoming increasingly intelligent, leading to progress in various fields.

This research developed a model to predict the correct topics of given Urdu documents. The model incorporates data preprocessing techniques and unsupervised learning techniques. These techniques are combined in a single framework. Data preprocessing involves tokenization, stemming, removal of

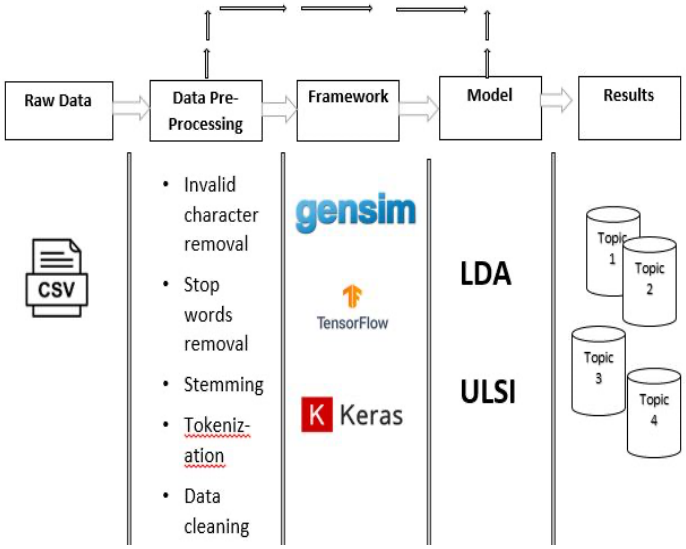


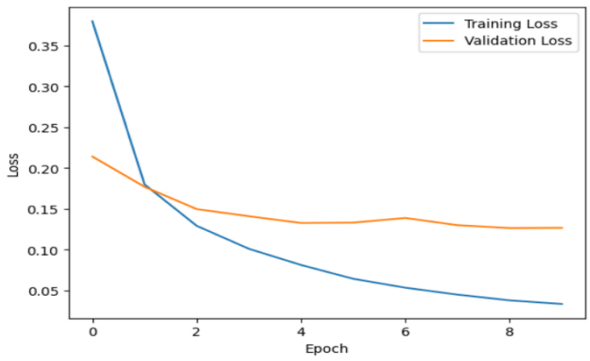
FIGURE 4. Process of Model Development and Implementation

invalid characters, data cleaning, and stop word removal. For topic prediction model development, unsupervised learning techniques such as LDA and ULSI are provided for researchers to implement. The complete process of model development and implementation for predictions is illustrated in Fig. 4.

Fig. 5.a and Fig. 5.b respectively display the loss functions graphs of the LDA and ULSI approaches. The purpose of loss function is to find the optimal model parameters that minimize this loss, which, in turn, results in a well-performing model capable of making accurate predictions on new, unseen data.

In this research observed a significant turning point in our models (LDA and ULSI) training process at "epoch 2 and loss 0.2." After just two complete iterations through our training data, the model achieved a remarkably low loss of 0.2, signifying a close alignment between its predictions and the actual target values. This rapid convergence is promising, suggesting that our model is learning quickly and holds the potential for accurate predictions.

FIGURE 5a. Loss function graph of LDA approach



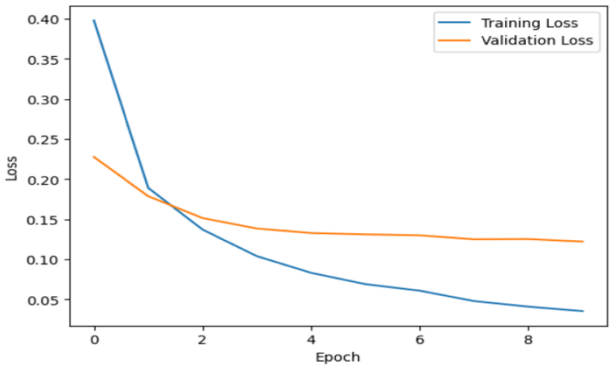


FIGURE 5b. Loss function graph of ULSI approach

However, further evaluation is needed to assess its performance on unseen data and its long-term training behavior. This finding motivates us to delve deeper into this researcher's model capabilities as progress with our research.

This study have obtained a substantial and diverse dataset consisting of over one million words, all composed in the Urdu language. This corpus encompasses news articles from a wide spectrum of domains, including politics, sports, commerce, finance, and many more. This rich and extensive collection of textual data not only reflects the linguistic richness of Urdu but also offers a valuable resource for research and analysis across multiple domains. With such a comprehensive dataset at this study disposal, we are well-positioned to explore and understand the nuances of language and content in Urdu, facilitating a deeper comprehension of the topics and themes that are of significant societal interest and relevance.

C. PREPROCESSING

Prior to the implementation of the actual unsupervised approaches, namely LDA (Latent Dirichlet Allocation) and ULSI (Unsupervised Latent Semantic Indexing), a critical phase of data preprocessing was carried out to optimize the quality and relevance of our input data. This preprocessing pipeline encompassed a series of essential steps, including the removal of invalid characters, the elimination of stop words, stemming, lemmatization, and comprehensive data cleaning procedures. These measures were vital in ensuring that the text data was in its most refined and standardized form, ready for analysis. This study harnessed well-established Python frameworks such as Keras, Gensim, and TensorFlow to facilitate this process, leveraging their robust capabilities in natural language processing. Subsequently, the deployed LDA and ULSI topic modeling approaches on the meticulously preprocessed data to extract and identify meaningful and contextually relevant topics, thereby enriching researcher's understanding of the underlying patterns and themes within the corpus.

The findings presented in Table 5 underscore the remarkable progress achieved in our study when compared to previously reported results in the domain of Urdu topic modeling. This investigation achieved notably superior outcomes, exemplified by

an impressive 99% F_1 score. This achievement stands in stark contrast to the comparatively lower results reported by other researchers in the field. For instance, Shakeel [17] reported a mere 78%, Munir reached 13.8% [29], Latif attained 71% [16], and Anwar reported 92% [15] F1 scores in their respective studies. It is worth highlighting that our study harnessed the power of both LDA and ULSI approaches, yielding exceptional results of 99% and 98%, respectively. This substantial improvement of approximately 7% in performance compared to prior studies is a testament to the effectiveness of our methodology and the advancement it brings to Urdu topic modeling research.

IV. RESULTS AND ANALYSIS SECTION

In this article, the implemented two distinct topic modeling algorithms LDA and ULSI. As mentioned in the model development, all the selected approaches were applied to a pre-processed dataset. The dataset used was collected from internet, comprising news headlines and descriptions, containing 1 million diverse Urdu news items encompassing topics like entertainment, sports, politics, and business. Both algorithms were implemented in an unsupervised manner, where the pre-processed text was fed into LDA and ULSI, generating topic words from it.

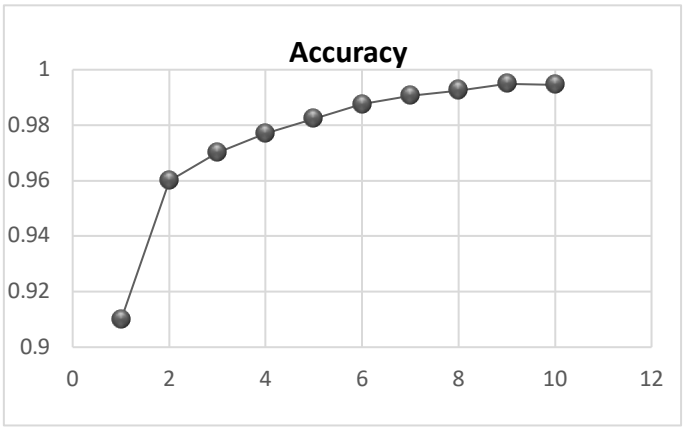


FIGURE 6. Result using LDA approach

A. LDA MODEL

After applying LDA to generate topic words, in this study researcher observed that each topic corresponds to a specific class, such as sports, entertainment, business, or politics. Table 1 illustrates the results obtained from LDA, where each topic number contains topic words from 10 different topics, along with their accuracy in various categories. Additionally, the generated topic words demonstrate strong semantic relationships among them. Fig. 6 presents the accuracy graph obtained from the LDA model using the F_1 score.

As a result, the accuracy results from the LDA model using the coherence value and F_1 score as evaluation measures. The results demonstrate significant improvement in accuracy when incorporating the F_1 measure. The overall results for the LDA model are presented in Table 2, providing detailed insights into its performance.

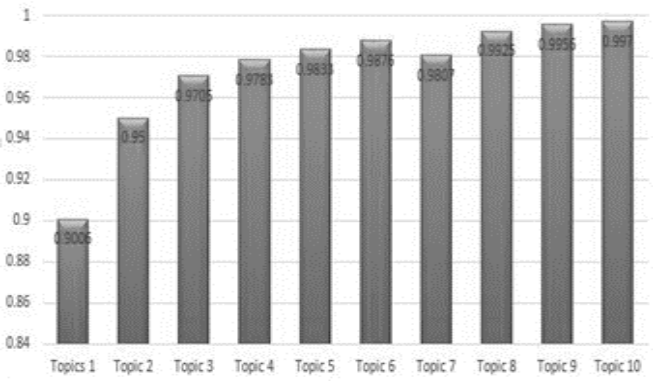


FIGURE 7. Result using ULSI approach

A. ULSI MODEL

Fig. 7 displays the graphical representation of the results obtained from the ULSI model applied to the Urdu news dataset. The findings indicate that LDA performed better than ULSI. The topic words generated by LDA exhibit stronger semantic relationships compared to those produced by ULSI. Furthermore, the generated topics preserve semantic relationships among their respective topic words. Specifically, ULSI topic 10 exhibits keywords obtained from various news categories.

TABLE 1. TOP TEN ACCURACIES ACHIEVED USING LDA TECHNIQUE

Topics	Accuracy
Topics 1	0.9097
Topic 2	0.9601
Topic 3	0.9701
Topic 4	0.9769
Topic 5	0.9822
Topic 6	0.9876
Topic 7	0.9907
Topic 8	0.9925
Topic 9	0.9949
Topic 10	0.9945

TABLE 2. LDA OVERALL RESULTS

Result of LDA technique	
F_1 score	99%
Coherence value	44%

This study emphasizes the significantly superior values depicted in the provided figures. These findings shed light on the relative words that occur most frequently within a specific topic related to news domains. Table 3 presents the estimated accuracy for the top 10 themes using ULSI, while Table 4 showcases the overall results obtained from the ULSI model.

The ULSI model presents results based on coherence and F_1 score values to demonstrate the high reliability of the topics extracted from the given documents. By showcasing these values, this study aims to highlight the strong performance of this model. The Fig. 6 and Fig. 7 provided illustrate the superior performance metrics obtained from these models. These results focus on specific news-related topics and reveal the most prominent and relevant words.

TABLE 3. SHOWING TOP TEN ACCURACIES ACHIEVED USING ULSI APPROACH

Topics	Accuracy
Topic 1	0.9006
Topic 2	0.95
Topic 3	0.9705
Topic 4	0.9783
Topic 5	0.9833
Topic 6	0.9876
Topic 7	0.9807
Topic 8	0.9925
Topic 9	0.9956
Topic 10	0.997

TABLE 4. SHOWING TOP TEN ACCURACIES ACHIEVED USING ULSI APPROACH

ULSI model results	
F_1 score	98%
Coherence value	37%

B. COMPARISON

This study obtained the best results from the LDA model and was able to achieve higher coherence value and F_1 measure compared to previously reported results. The following results are compared with those of other studies. Table 5 presents the current study results with previous study results.

Techniques/ Evaluation Measure	LDA	ULSI
F_1 score **(current study)**	99%	98%
Coherence value **(current study)**	44%	37%
Shakeel, K., Tahir, G. R., Tehseen, I., & Ali, M. (2018) (Shakeel, Tahir, Tehseen, & Ali, 2018)	78%	-
Munir, Wasi, and Jami (2019) (Munir, Wasi, & Jami, 2019)	13.8%	37.8%
Latif, Shafait, and Latif (2021) (Latif, Shafait, Latif, & others, 2021)	71%	-
Anwar, Bajwa, Choudhary, and Ramzan (2018) (Anwar, Bajwa, Choudhary, & Ramzan, 2018)	92%	-

TABLE 5. COMPARISON OF RESULTS FOR TOPIC MODELING USING TWO ALGORITHMS LDA AND ULSI

Fig. 8 shows the detailed results of both approaches LDA and ULSI. In this figure, the developed LDA and ULSI model with 10 topics. Here you can observe printed the top words and their weights for each topic. The numbers on the top represent the topic number, and the words on the below the topic are the most important words for that topic along with their associated weights. The weights indicate the strength of the association between each word and the topic.

V. CONCLUSION

In conclusion, our article delved into applying two unsupervised learning approaches, LDA and ULSI, for the task of Urdu topic modelling. This study experimentation yielded compelling results, with both LDA and ULSI demonstrating their efficacy in capturing meaningful topics within the Urdu text data. Specifically, this study achieved the highest F_1 score of 99% and a Coherence value of 44% using LDA. In comparison, ULSI also delivered strong results with the highest F_1 score of 98% and a Coherence value of 37%. Based on these findings, it is evident that LDA exhibits a slight advantage in terms of F_1 score and Coherence, making it the more suitable choice for Urdu topic modelling in this study. However, it is important to note that the choice of approach may depend on various factors, including the specific characteristics of the dataset and the research objectives. As we look towards the future, we plan to further enhance our topic modelling efforts in Urdu by exploring semi-supervised approaches. These endeavours will contribute to a deeper understanding of Urdu content and its applications across various domains.

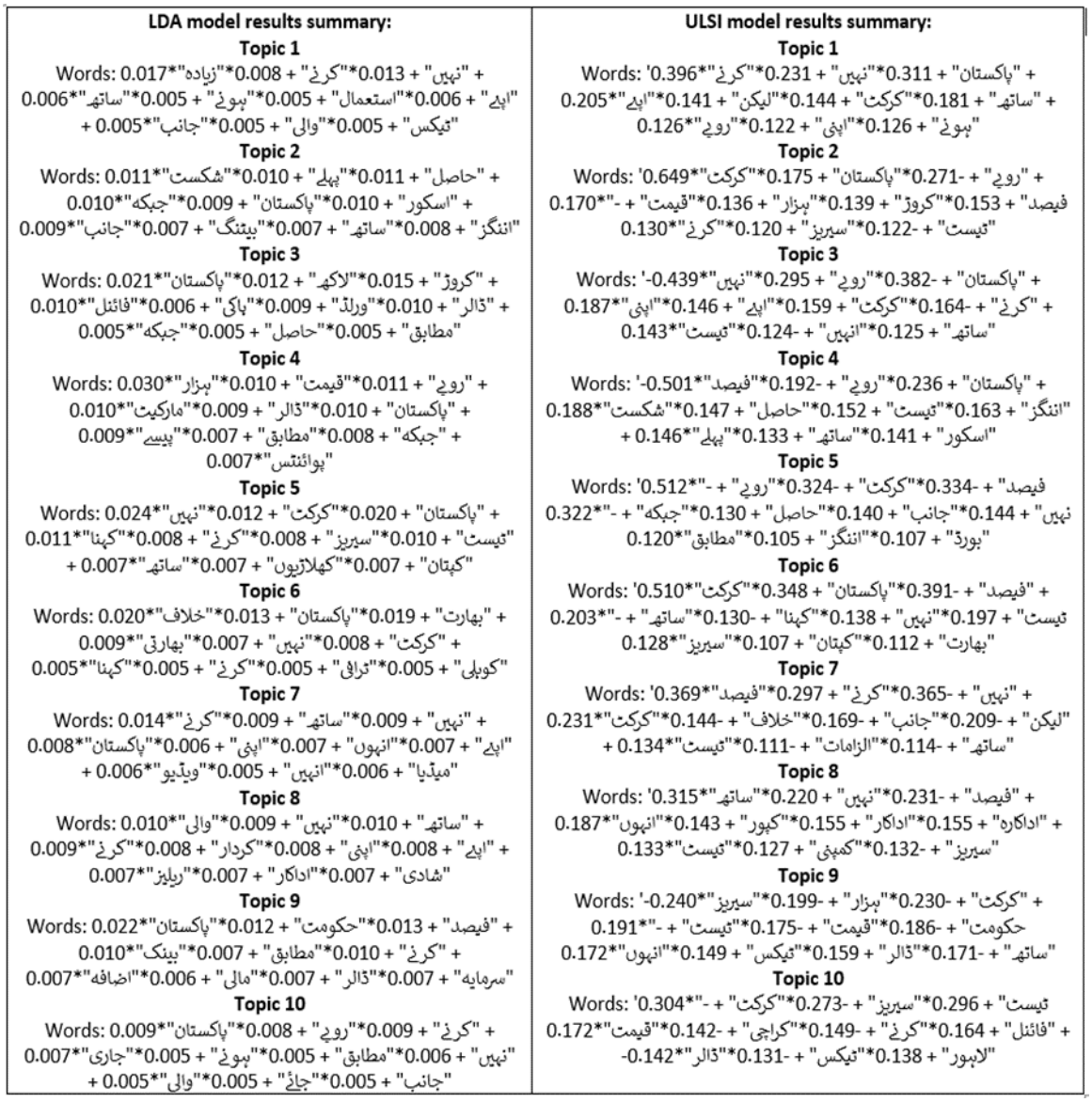


FIGURE 8. LDA and ULSI models results summary for all 10 topics

REFERENCES

[1] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic topic modeling: an experimental study on BERTopic technique," Procedia computer science, vol. 189, pp. 191--194, 2021.

[2] A. Saeed, R. M. A. Nawab, M. Stevenson and P. Rayson, "A word sense disambiguation corpus for Urdu," Language Resources and Evaluation, vol. 53, pp. 397--418, 2019.

[3] A. Daud, W. Khan and D. Che, "Urdu language processing: a survey," Artificial Intelligence Review, vol. 47, pp. 279--311, 2017.

[4] U. Hayat, A. Saeed, M. H. K. Vardag, M. F. Ullah and N. Iqbal, "Roman urdu fake reviews detection using stacked lstm architecture," SN Computer Science, vol. 3, no. 6, pp. 470--479, 2022.

[5] S. Shaukat, A. Shaukat, K. Shahzad and A. Daud, "Using TREC for developing semantic information retrieval benchmark for Urdu," Information Processing \& Management, vol. 59, no. 3, p. 102939, 2022.

[6] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," Int. J. Adv. Comput. Sci. Appl.(IJACSA), vol. 6, no. 1, pp. 147--153, 2015.

[7] L. Yue, L. Xueqiang, X. Shibin, W. Tao, "Topic detection based on keyword," in International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), USA, 2011.

[8] W. Sriurai, "Improving text categorization by using a topic model," Advanced Computing, vol. 2, no. 6, pp. 21--27, 2011.

[9] S. T. Dumais, "Latent semantic analysis," Annual Review of Information Science and Technology (ARIST), vol. 38, pp. 189--230, 2004.

[10] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Machine learning, vol. 42, pp. 177--196, 2001.

[11] A. Likas, V. Nikos, J. Verbeek, "The global k-means clustering algorithm," Pattern recognition, vol. 36, no. 2, pp. 451--461, 2003.

[12] A. Singh, B. S. Prakash and K. Chandrasekaran, "A comparison of linear

discriminant analysis and ridge classifier on Twitter data," in 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2016.

[13] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993--1022, 2003.

[14] T. Ehsan and A. H. Shahzad, "Finding Topics in Urdu: A Study of Applicability of Document Clustering in Urdu Language," Pakistan Journal of Engineering and Applied Sciences, vol. 23, pp. 77-85, 2018.

[15] W. Anwar, I. S. Bajwa, M. A. Choudhary and S. Ramzan, "An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution," IEEE Access, vol. 7, pp. 3224--3234, 2018.

[16] S. Latif, F. Shafait, R. Latif and others, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," IEEE Access, vol. 9, pp. 127531--127547, 2021.

[17] K. Shakeel, G. R. Tahir, I. Tehseen and M. Ali, "A framework of Urdu topic modeling using latent dirichlet allocation (LDA)," in 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, 2018.

[18] A. Amin, T. A. Rana, N. A. Mian, M. W. Iqbal, A. Khalid, T. Alyas and M. Tubishat, "TOP-rank: a novel unsupervised approach for topic prediction using keyphrase extraction for Urdu documents," IEEE Access, vol. 8, pp. 212675--212686, 2020.

[19] S. A. Zain, A. B. Mughal and S. M. Haider, "Urdu News Article Recommendation Model using Natural Language Processing Techniques," arXiv preprint arXiv:2206.11862, 2022.

[20] C. Wohlin and A. Aurum, "Towards a decision-making structure for selecting a research design in empirical software engineering," Empirical Software Engineering, vol. 20, pp. 1427--1455, 2015.

[21] V. R. Basili, "The experimental paradigm in software engineering," in Experimental Software Engineering Issues: Critical Assessment and Future Directions: International Workshop Dagstuhl Castle, Germany, September 14--18, 1992 Proceedings, Castle, Germany, 2005.

[22] R. L. Glass, I. Vessey and V. Ramesh, "Research in software engineering: an analysis of the literature," Information and Software technology, vol. 44, no. 8, pp. 491--506, 2002.

[23] S. Hussain, "Resources for Urdu language processing," in Proceedings of the 6th workshop on Asian Language Resources, IJCNLP, 2008.

[24] M. Mustafa, F. Zeng, H. Ghulam and H. Muhammad Arslan, "Urdu documents clustering with unsupervised and semi-supervised probabilistic topic modeling," Information, vol. 11, no. 11, pp. 518--534, 2020.

[25] M. S. Bhatti, A. Ullah, R. Latip, A. Sohail, A. Riaz and R. Hassan, "Benchmarking Performance of Document Level Classification and Topic Modeling," CMC-COMPUTERS MATERIALS \& CONTINUA, vol. 71, no. 1, pp. 125--141, 2022.

[26] T. Ahmed, S. Urooj, S. Hussain, A. Mustafa, R. Parveen, F. Adeeba, A. Hautli and M. Butt, "The CLE urdu POS tagset," in LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 2014.

[27] A. Ghafoor, A. S. Imran, S. M. Daudpota, Z. Kastrati, S. Shaikh and R. Batra, "SentiUrdu-1M: A large-scale tweet dataset for Urdu text sentiment analysis using weakly supervised learning," Plos one, vol. 18, no. 8, pp. 1--22, 2023.

[28] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed and M. Fayyaz, "Exploring deep learning approaches for Urdu text classification in product manufacturing," Enterprise Information Systems, vol. 16, no. 2, pp. 223--248, 2022.

[29] S. Munir, S. Wasi and S. I. Jami, "A comparison of topic modelling approaches for Urdu text," Indian Journal of Science and Technology, vol. 12, pp. 45--52, 2019.