DOI: 10.51153/kjcis.v7i2.233

Topic Modeling and Identification in a Resource-Scarce Roman Urdu Language

Afsa Riaz,[a] Mohsin Ali Memon,[b] Amirita Dewani,[c],* Sania Bhatti,[d] Fariha Naeem [e], Memoona Sami [f]

aMehran University of Engineering and Technology, Pakistan (afsa.riaz123@gmail.com)
bMehran University of Engineering and Technology, Pakistan (mohsin.memon@faculty.muet.edu.pk)
cMehran University of Engineering and Technology, Pakistan (amirita@faculty.muet.edu.pk)
dMehran University of Engineering and Technology, Pakistan (sania.bhatti@faculty.muet.edu.pk)
eMehran University of Engineering and Technology, Pakistan (farihanaeem2206@gmail.com)
fMehran University of Engineering and Technology, Pakistan (memoona.sami@faculty.muet.edu.pk)

Corresponding Author: *Amirita Dewani (amirita@faculty.muet.edu.pk)*

Abstract

In an age dominated by the Internet revolution, companies are making their businesses available to diverse groups of customers by leveraging the usage of e-commerce. To keep track of customer satisfaction and a competitive edge in the market, e-commerce businesses need to scrutinize their customer reviews. The manual approach to analyzing customer reviews is time and effort consuming. Automated product review analysis exists but resource-poor languages like Roman Urdu lack such resources. To overcome this problem, this research presents a solution by incorporating Topic Modeling for Roman Urdu product reviews. A dataset of 8,000 Roman Urdu product reviews was curated from an online shopping platform. Various language-specific data cleaning steps were applied to data in the pre-processing step before experimentation. Different algorithms for Topic Modeling were implemented, out of them BERTopic produced outstanding results leaving the others behind. The results were also evaluated with an open-source dataset to check model generalization and reliability. Utilizing the power of machine learning and recent approaches, this study is a step forward to automated review analysis in the Roman Urdu language.

Keywords: Natural language processing, Topic modeling, Low resource language processing, Roman Urdu Language, BERT Model.

1. Introduction

After the internet revolution, people got accustomed to internet usage in their daily lives. Similarly,the e-commerce industry got a hike as people started to do online shopping. There are various international and national online shopping sites available. Many of these websites provide customers with the opportunity to provide feedback about the purchased products. Feedback can be in the form of ratings and reviews. In most e-commerce platforms, the user is given the option of giving a star-based rating. Ratings alone cannot define

the actual opinion of the customers about the product because different levels of rating have different interpretations for each individual.

Customer reviews play a vital role in understanding the customer's opinion related to the products. They provide businesses with the opportunity to analyze different types of information about their products s. But going through customer reviews manually is labor-intensive, time-consuming, and sometimes infeasible. The use of Machine Learning and Natural Language Processing techniques can make review analysis much easier and more efficient. For automated examination of product reviews, various resources and systems are available in resource-rich languages. However, the Roman Urdu language being a low-resource language lacks such resources. The Roman Urdu language is an informal way of writing the Urdu language using the Latin script or Roman Script, and it is widely used in online communication in the South Asian region. Roman Urdu is one of the challenging languages in automatic language processing due to its lexical variations [1].

To address the shortage of automatic review analysis resources in Roman Urdu, a topic modeling-based product review analysis of Roman Urdu product reviews is proposed in this study. To our knowledge, limited research is available in topic modeling for the resource-poor Roman Urdu language. Topic Modeling is a Natural Language Processing technique used to perform an exploratory analysis of textual data and helps to uncover the latent topics present in the textual data [2]. With the help of topic modeling, online businesses can scrutinize their product reviews, uncovering the hidden topics and themes of interest from them.

This research focuses on the development of a framework for identifying hidden themes and aspects in the low-resource Roman Urdu language. In summary, this research study makes the following contributions:

1. Curated a valuable linguistic resource for the Roman Urdu language regarding a product reviews dataset.
2. Performed a comparative analysis of various topic modeling techniques to identify the most efficient and feasible approach for the Roman Urdu language.
3. Finally, the study cross-validated the results using an open-source dataset.

As per our knowledge, this is the first time that a topic modeling-based product review analysis of Roman Urdu product reviews has been performed.

The rest of the paper is organized as follows:

In Section 2, a thorough review of the relevant literature is provided to contextualize the work. Section 3 sheds light on the research gap and defines the problem statement. Section 4 outlines the methodology employed in the research, detailing the data collection and analysis procedures. The results of the study are presented in Section 5, followed by a comprehensive discussion of the findings in Section 6. The paper concludes in Section 7 with a summary of key findings, implications, and directions for future research.

2 Literature Review

The framework outlined in [3] is the initial endeavor to suggest an approach based on topic modeling for product comparison. Online reviews and ratings of two different computer mice and oil diffusers from Amazon were collected. Based on ratings, positive and negative reviews were separated first, and then LDA was applied for topic modeling. After conducting various experiments, the study concluded with 10 dimensions of satisfaction for mouse reviews and 6 dimensions for oil diffusers.

Meanwhile, the study discussed in [4,5] utilized the approach of topic modeling for identifying the opinions of airline passengers. A dataset of 4933 reviews was collected from airlinequality.com. After various data cleansing steps were applied, the reviews were transformed into Bag of Words (BoW) format, and three algorithms—LDA, LSA, and NMF—were employed with 6 to 20 topics to determine the optimal one. The best topic model was identified as LDA with 10 topics, showcasing a coherence score of 0.455. In addition, the work presented by [2] proposed a topic modeling-based framework for teaching evaluation using students' written comments. They collected a dataset of 1,10,420 undergraduate surveys from a research institute located in Texas, United States. They chose LDA for topic modeling and compared the coherence scores from 5 to 65 topics. They found the highest coherence score of 0.59 with 8 dimensions of satisfaction.

The work in study by [6] proposes an unsupervised aspect identification framework based on topic modeling. For this purpose, five datasets were obtained from Kaggle, and all the data pre-processing steps were executed, followed by the conversion of the data into Bag of Words (BoW) format. Subsequently, topic modeling was carried out using LDA, LSA, and the Hierarchical Dirichlet Process, and the optimal number of topics was determined by applying the coherence score. Ultimately, a comparison of the algorithmic results was conducted using the coherence score, revealing that LDA outperformed LSA and the Hierarchical Dirichlet Process in all five datasets.

The study proposed by [7] investigated the power of BERTopic in analyzing recurrent social science surveys. They extracted the dataset from 318 questionnaires from 1946 to 2020. The overall rows in the final dataset were 42008 with several columns. They applied BERTopic with different sentence embeddings

and observed the coherence scores between 3 to 25 topics. For the evaluation of their results, they compared their best performing BERTopic model with the Top2vec and Contextualized Topic Model and their model outperforms both. Furthermore, the work described in [8] presents a topic modeling and sentiment analysis based study on global climate change tweets. A dataset of 390,016 tweets was collected, and then all the text preprocessing steps were applied along with the reverse geocoding of the address of the tweets to perform a geospatial analysis. Volume analysis was first performed to uncover the time and geographical information from the tweets. Subsequently, author-pooled LDA with 5, 20, and 80 topics was applied, and sentiment analysis using VADER was performed. It was found that LDA with 20 topics generated more interpretable topics, and most of the reviews in that dataset were negative, especially related to political and extreme weather conditions.

The research outlined by [9] proposed a framework based on topic modeling and sentiment analysis to analyze online passenger reviews. They collected more than 14,000 reviews of 27 Asian airlines from SKYTRAX, the world's largest airline website. For implementation, they chose R programming language and started their work by cleaning those reviews. Afterwards, they performed a frequency analysis through which they observed that most of the words from the top 100 words were positive. Through the process of trial and error, they found LDA with 6 topics best with their dataset and then labeled those topics with the help of top words in the topics. For sentiment analysis, they used the Opinion Lexicon and concluded that most of the reviews were positive.

Likewise, the study presented by [10] put forward a model for the detection of aspects and sentiments in a collection of literary texts. For this research, a dataset of 28 Iranian literary books was gatheres. After data collection, data cleaning was performed which included text normalization and stopwords removal. Then different visualizations like word clouds and bar charts were generated to understand the themes from the dataset. LDA, LSA, NMF, and HDP were applied for topic modeling, and coherence scores from 10 to 30 topics were observed for each algrithm. LSA was identified as the best topic model with the highest coherence score of 0.57. VADER and TextBlob were utilized for sentiment detection, and it was observed that the results of VADER were more comprehensive and accurate.

The framework outlined by [11] identified the interest of social media users through Roman Urdu tweets and Google reviews. A dataset of 15000 tweets and 6000 Google reviews was collected. After data collection, different data-cleaning steps were applied to make the dataset clean and balanced. DistilBERT and VADER with TextBlob were utilized to perform sentiment analysis, resulting in the separation of positive and negative tweets. Afterward, three different topic modeling algorithms LDA, LSA, and

BERTopic were implemented. The study concluded that the BERTopic model with 6 topics and a coherence score of 0.58 outperformed the other two models and Random Forest combined with DistilBERT with an accuracy of 76% outperformed the other models. On the other hand, [12] conducted a comparative study of different topic modeling techniques for analyzing customer reviews. The first dataset was of 29,200 reviews and was collected from the UAE Ministry of Economy. The second dataset was taken from Kaggle consisting of 1600 customer reviews. After applying data preprocessing steps, they applied six different topic modeling algorithms including BERTopic. They observed the values of coherence score for 5 and 10 topics in both datasets. The BERTopic outperformed all the other topic models in both datasets for both 5 and 10 topics.

The research work by [13] discussed a topic modeling and sentiment analysis based study of Bangladesh Airlines' reviews. The main objective of the study was to identify aspects of passenger satisfaction and dissatisfaction from airline reviews. For data collection, SKYTRAX (airlinequality.com) and TripAdvisor were used as data sources, and a total of 1095 airline reviews were scraped. Following extensive data cleaning, various types of analysis and techniques, including frequency analysis, word cloud generation, topic modeling, and sentiment analysis, were performed. Six distinct topics, including flight schedule, food and beverages, luggage, and staff service, were identified. It was concluded that most of the services, especially the food, were well-received by the passengers. The research conducted by [14] presents an innovative exploration of COVID-19 vaccine discussions through topic modeling and sentiment analysis. The study aimed to systematically support government officials and policymakers by identifying the concerns of the general public. A dataset of 78,827 tweets covering the 7 most popular vaccine brands from Dec 2020 to July 2021 was collected. Plots were generated to understand the distribution of tweets according to geography and vaccine brands. VADER was employed for sentiment analysis, and LDA was applied for topic modeling on positive and negative reviews separately. The optimal number of topics was found to be 11, and the conclusion was drawn that the vaccines were mostly trusted by the public, who expressed willingness to get vaccinated. Negative tweets primarily focused on concerns related to vaccine shortage and side effects.

The work presented by [15] identified user interests in social data over time using the technique of topic modeling. They collected reviews of electronic products from the websites of different popular brands. For sentiment analysis, they applied different techniques, out of which VADER gave the most promising results. To identify user interests, they performed topic modeling using LDA, BERT and a hybrid approach by combining LDA and BERT. The hybrid approach outperformed the LDA and BERT by giving the highest

coherence of 52%. For profiling user interests over time, they employed Top2Vec and performed human-based validation for the evaluation of results. The study proposed by [16] identified the rapid growth of digital libraries and performed a comparative study on e-books to categorize them based on their content. A dataset of 300 books consisting of 23 million words was collected. Various cleaning techniques and steps were applied to prepare the dataset and then popular topic modeling algorithms LDA and LSA were applied to identify the optimal topics based on coherence score. The most promising results were obtained with LDA for 20 topics, yielding a coherence score of 0.59. The model could be enhanced by increasing the size of the dataset. Additionally, the research outlined by [17] focused on the social media comments of the learners and proposed a topic modeling based framework for the identification of topical information from learners' comments on social media. A dataset of 120,000 tweets was collected from 12,187 learners using Twitter API and Netlytic. After the implementation of various cleaning and data pre-processing steps, LDA, LSA and BERTopic were applied for topic modeling. BOW and TFIDF feature extraction techniques were used with LDA and LSA, whereas BERTopic has built-in BERT embeddings. The study concluded that BERTopic performed superior to the other two algorithms.

To contextualize our work, Table 1 presents a summary of the entire literature review.

Table 1 Summary of the Literature Review.

Reference	Dataset Description	Algorithms/Techniques	Best Algorithm	Coherence Score
[2]	1,10,420 undergraduate surveys	LDA	LDA with 8 dimensions	0.59
[4]	4933 reviews from airlinequality.com	LDA, LSA, NMF	LDA with 10 topics	0.455
[7]	extracted the dataset from 318 questionnaires from 1946 to 2020	BERTopic, CTM and Top2Vec	BERTopic	–
[8]	collected a dataset of 390,016 tweets	LDA	LDA with 20 topics	–
[10]	collected data from 28 Iranian literary books	LDA, LSA, NMF	LSA	0.57
[11]	15000 tweets and 6000 Google reviews	LDA, LSA, BERTopic	BERTopic with 6 dimensions	0.58
[12]	29,200 reviews from the UAE Ministry of Economy	LDA, LSA, NMF, Top2Vec, BERTopic	BERTopic with 10 dimensions	0.60
[18]	10,000 randomly sampled bank customers' tweets	LDA, LSA, HDP, BERTopic	BERTopic with 3 topics	0.84
[14]	78,827 tweets covering 7 popular vaccine brands	LDA	LDA with 11 topics	–
[15]	15,231 Roman Urdu reviews collected from different sites	LDA, BERT, hybrid(LDA + BERT)	Hybrid	0.52
[16]	300 books comprising 23 million words	LDA, LSA	LDA with 20 topics	0.59
[17]	120,000 tweets from 12,187 learners	LDA, LSA, BERTopic	BERTopic with 40 topics	0.61

The approach of performing topic modeling in various application areas for the identification of latent themes is very common in resource-rich languages. However, Roman Urdu, a low-resource language, has very limited work in various areas of NLP, including topic modeling. Taking it as a motivation, this research work puts efforts to propose a feasible topic modeling approach for the Roman Urdu reviews dataset and provide cross domain evaluation for the validity of the produced results.

3 Problem Statement

Topic Modeling in NLP is an important research area that aims to extract the underlying themes and aspects from a collection of documents. It is an efficient textual analysis technique that has automated the analysis of vast amounts of textual unstructured data. Research efforts made till so far in this context are mainly oriented towards Resource-rich languages that have several resources for automated analysis of reviews to uncover hidden patterns. Roman Urdu language, being a resource-poor language, lacks such resources and has various unexplored areas in the field of NLP. Given these limitations, various NLP problems in Roman Urdu language tend to be more challenging such as Topic Modeling, Named Entity Recognition, Question Answering systems, and Sentiment Analysis. To fulfil the above stated gap, in this study we have implemented topic modeling and extraction techniques along with language specific preprocessing methods to extract significant information from documents comprising of product reviews collected in Roman Urdu language from vast number of sources. We conducted various experiments including LSA, LDA, NMF and fine-tuning BERT-Topic, which has proven to set new standards in transformer-based models.

4 Proposed Methodology

Fig.1 depicts the major phases of the proposed methodology followed to accomplish this research work, initiating the data collection process. All the implementation work has been accomplished using Python programming language. Python is a general-purpose programming language as it supports programming for multiple domains including web development, data science, artificial intelligence, machine learning, game development, and more. Various libraries are used to develop the models including Numpy, Pandas, Scikit-learn, Gensim, BERTopic, and Matplotlib. Reviews were scraped from an online shopping platform [19] (Online Shopping in Pakistan, 2024) and subsequently, Roman Urdu reviews were segregated. Following this, various pre-processing and cleaning procedures were applied to the Roman Urdu reviews. To align these reviews with Machine Learning algorithms, they were converted into vector formats.

Subsequently, the vectorized reviews were input into the topic modeling algorithms to derive results. Finally, the last step involved the evaluation and interpretation of the obtained results and cross dataset evaluation phase. The significant phases of research methodology are detailed in subsequent sections.

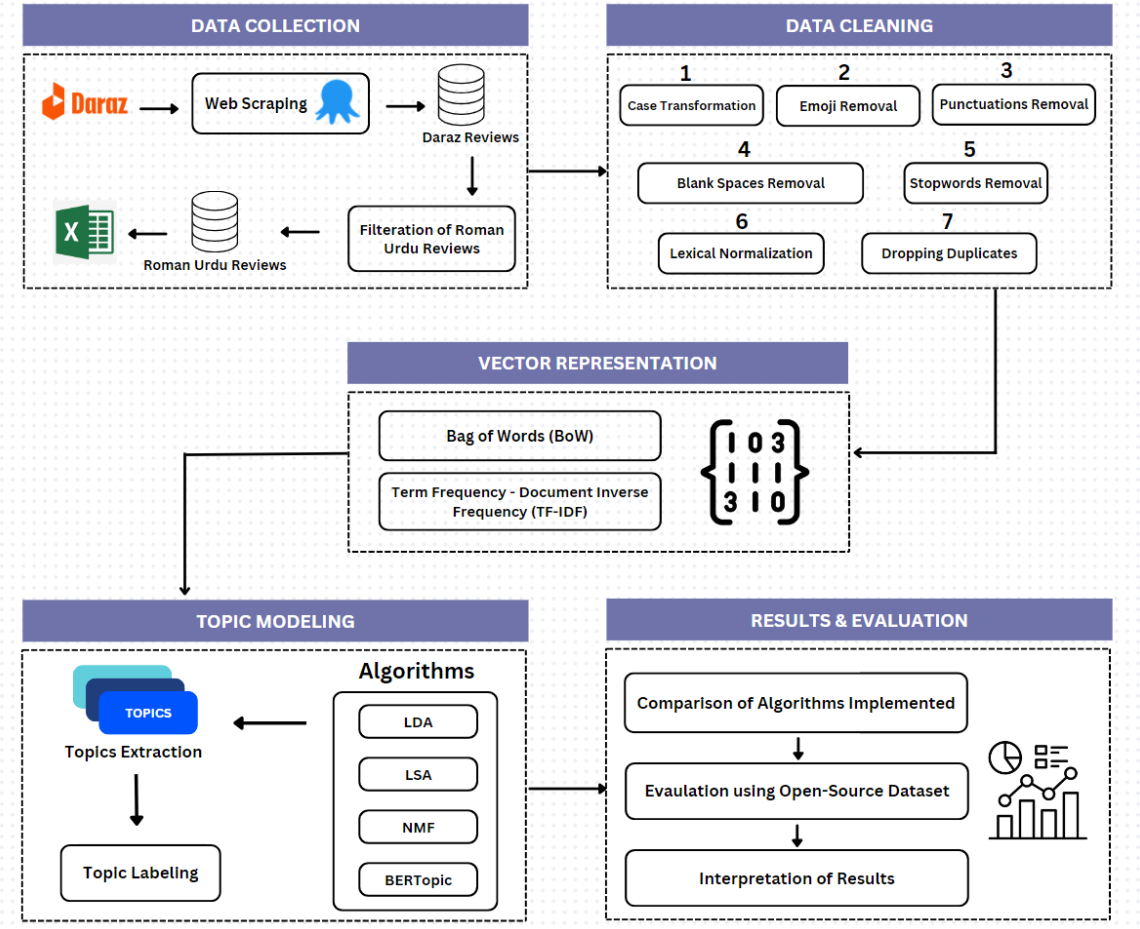

Fig. 1 Proposed Methodology of Roman Urdu Topic Modeling.

4.1 Data Collection

Roman Urdu, being a low-resource language, lacks sufficient linguistic resources, and a similar scarcity is observed in datasets for Roman Urdu product reviews. The open-source product reviews datasets exist as in [20], but they are insufficient in size and are not enough for getting appropriate results in an ML-based study. Therefore, we decided to create our dataset for this study. The process of data collection consisted of two phases. During the initial stage, reviews were scraped, followed by the subsequent phase dedicated to filtering out Roman Urdu reviews.

4.1.1 Web Scraping

Daraz is an online marketplace and e-commerce platform that operates in Pakistan. This platform allows the customers to give their feedback about the purchased products. We have extracted the product reviews from this online shopping platform Daraz using the Octoparse tool [21]. The version of Octoparse used for web scraping was Octoparse 8.0. The categories from where the reviews were collected are clothing, footwear, and watches.

4.1.2 Extraction of Roman Urdu Reviews

The data extracted from Daraz encompassed reviews in Roman Urdu, English, and Urdu languages. This study specifically concentrated on conducting topic modeling for Roman Urdu reviews. As a result, we manually isolated the Roman Urdu reviews from the scraped dataset, resulting in a final dataset size of 8,000 reviews. We stored the final dataset in a CSV Excel file. Table 2 presents some instances of the Roman Urdu reviews from the final dataset along with their English translation.

Table 2 Roman Urdu Reviews from the final dataset along with their English translation.

S.No.	Roman Urdu Review	English Translation
1.	price k hisab se shoes ache hain.satisfied....	The shoes are good considering the price satisfied.
2.	Dress acha h par design or color different bheja hai	The dress is nice, but a different design and color were sent.
3.	bkws quality ese plastic material h..	The quality is poor; it's made of plastic -like material.
4.	Thora sa damage piece mila ha baqi quality best ha Jo pic ma ha same wo he ha	Received a slightly damaged piece, but otherwise, the quality is excellent. It's the same as shown in the picture.

4.2 Data Preparation and Cleaning

Online product reviews, a type of social text, display a lack of structure and informality. Consequently, rigorous pre-processing and cleaning procedures are essential before implementing any machine learning algorithm. Therefore, we have applied the following steps to clean the Roman Urdu product reviews.

- Case Transformation: We transformed all the upper-case letters into lower case letters for dimensionality reduction.
- Removal of Non-ASCII Characters: In this phase, we eliminated the emojis because they are treated as non-ACSII characters.
- Removal of Punctuations and insignificant characters: In this step, we removed punctuations and unnecessary characters such as full stops, commas, exclamation marks etc. This helped for cleaning the data and better data readability.
- Removal of Digits: We eradicated all the digits present in the Roman Urdu reviews.
- Roman Urdu Stop Words Removal: This phase involves elimination of Roman Urdu domain specific stop words. Stop words are those words in a language that are used very frequently in sentences but do not contribute significantly to the meaning of the sentence [22]. We have taken Roman Urdu stop words from a GitHub Repository [20] (Roman Urdu Stop Words, 2024) and extended this list of stop words by adding more stop words according to our dataset.
- Lexical Normalization: The purpose of this step was to reduce the lexical variations of the Roman Urdu language. As the nature of Roman Urdu is highly unstructured, some lexical normalization rules were applied to the data like replacing the 'ay' in the last of a word with 'e'. So, the words like achay, kapray, jesay will become ache, kapre, jese. Also, the 'ia' in the last of a word will be replaced by 'i' and some other rules were also referred from study [23].
- Dropping Duplicates: This phase includes the elimination of any duplication in reviews caused by human error or resulting from applying the above steps.

Fig. 2 represents the visual illustration of the data cleaning process.

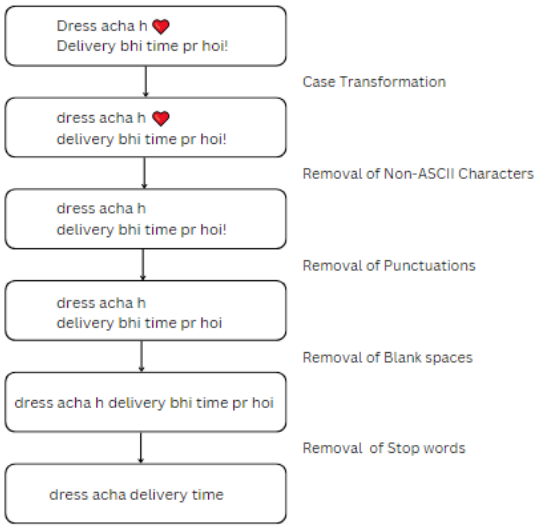

Fig. 2 Visual Illustration of Data Cleaning Process.

4.3 Vector Representation

As Machine Learning algorithms only learn from numerical or vector representation of data [24]. Therefore, it's significant to map the data to vector representation. In this research study, we have converted our dataset into the following vector formats.

4.3.1 Bag of Words

The Bag of Words or BoW is a commonly used technique in various areas of Artificial Intelligence including image classification, object detection, text classification, event recognition and distance determination [25]. In Natural Language Processing, the Bag of Words technique is used to convert data to a vector of numbers representation. In this vector representation, each element corresponds to a unique word in the corpus and the value of each element is calculated by the frequency of the representative word in the corresponding sentence sentence [26]. As a result, the Bag of Words technique gives a Document Term Matrix commonly called DTM.

We have used the BoW format of vector representation with the Topic Modeling algorithm LDA. As this algorithm works with word co-occurrence patterns and does not consider their order and semantics, the BoW is a suitable choice for LDA.

4.3.2 Term Frequency-Inverse Document Frequency

The TF-IDF is also a popular text vectorization technique in NLP. It scores a term in the document based on its relative frequency in the whole corpus. It consists of two parts: Term Frequency (TF) and Inverse Document Frequency (IDF) [24]. TF is calculated by dividing the number of occurrences of a term t by the total number of terms in a document. IDF is calculated by dividing the total number of documents in the corpus by the total number of documents containing the term t and taking the logarithm of this division result. TF and IDF are then multiplied to get the TF-IDF score for a particular term t.

We have used the TF-IDF format of vector representation with the Topic Modeling algorithms LSA and NMF as these algorithms also consider the relative importance of terms in the corpus.

4. 4 Algorithm Selection Process

There are many popular algorithms and techniques used for topic modeling. In resource-rich languages,

BERTopic, the most advanced addition to topic modeling techniques, has proven to outperform traditional topic modeling algorithms [7,11,12,17,18].

Moreover, the relatively recent surge of transformer-based models has made them suitable for various NLP tasks and language understanding problems. As compared to the traditional Topic modeling techniques, fine tuning BERT has proven to be more efficient by different research studies since it captures contextual similarity between words and considers the connection and inner semantics between phrases and sentences. In this study we investigated whether BERTopic, which outperforms traditional methods in resource-rich languages, would have a similar edge in Roman Urdu language. Therefore, we selected three of the most popular traditional topic modeling algorithms, namely, Latent Dirichlet Allocation, Latent Semantic Analysis, and Non-Negative Matrix Factorization, along with BERTopic and compared the results.

4.4.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, which comes under the category of unsupervised algorithms, is a probabilistic generative technique that is widely used for information retrieval, text summarization and topic modeling [27]. The LDA algorithm works on two main assumptions. It assumes that the documents or sentences are a mixture of different topics, and these topics are a mixture of different words. LDA takes textual data in the form of DTM and breaks down this DTM into two other smaller matrices. One is the Document Topic matrix which shows the topics present in each document and the second is the Topic Word matrix which represents the words or terms present in a topic.

The most important parameter of LDA is the number of topics to be generated and the user defines this parameter. LDA works in a repetitive manner and in each iteration, it assigns one of the k topics to each word present in the corpus and calculates some probabilities. Based on these probabilities, it reassigns a topic to each word in the next iteration [28]. In this way, the algorithm works and converges to a point when the values of both submatrices become optimal.

4.4.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a widely accepted and implemented technique in the fields of Artificial Intelligence, Psychology, Education, Cognitive Science, Genetic Science, and Information Systems [29]. Similar to LDA, the LSA algorithm also takes input in the form of DTM and then applies a popular matrix factorization technique called Singular Value Decomposition (SVD) to it [30]. The SVD decomposes the DTM into three smaller matrices: U, S, and Vt. U and Vt are column orthonormal matrices and S is a

diagonal matrix. U is called the left singular vector, and it represents the topics present in each document. Vt is called the right singular vector, and it represents the words present in each topic. S contains singular values that denote the importance of each topic, and we can reduce the number of topics by defining it before LSA execution. By defining the number of topics, user can the control the number of topics generated by LSA.

4.4.3 Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization is a popular unsupervised Machine Learning technique that is widely used in language processing, image analysis and speech recognition [31].

The NMF algorithm takes input in the form of a DTM and then reduces this matrix into two submatrices W and H [32]. The W matrix represents the importance of each topic related to each document and the H matrix represents the importance of each word related to each topic. The number of topics is a critical parameter in the NMF technique, and it can be controlled by the user. Initially, when the NMF starts to work, it assigns random non-negative values to both submatrices. The product of these smaller matrices can never be equal to the actual DTM. The objective of NMF is to optimize the values of W and H matrices such that their product will be much closer to the actual DTM. In each iteration, the values of W and H are updated so that the divergence will become closer to zero [32].

4.4.4 BERTopic

BERTopic is an advanced topic modeling algorithm that leverages the use of BERT embeddings and c-TF-IDF to create high-quality topics from unstructured textual data [33]. The BERT embeddings are state-of-the-art embeddings that are highly contextualized and not only store the semantics of individual words but also store the context in which the words are used. BERTopic does not need any kind of DTM as an input. It simply takes the documents in the form of a list. First, it internally converts the textual data to BERT embeddings. The next thing it does is dimensionality reduction because BERT embeddings are of 384 dimensions, so it converts the embeddings to lower dimensions. After that, a clustering algorithm is applied to cluster similar documents which can be interpreted to a single topic. In the end, a class-based TF-IDF is applied to these clusters to get the most important terms in each cluster, and through them, cluster labeling can be completed [34].

4.5 Experimental Setup for Implementation

For implementing all the pre-processing steps and topic modeling algorithms, the Google Colab environment was used. The programming language used was Python along with its most popular libraries including Numpy, Pandas, Gensim, Sklearn, BERTopic, Matplotlib and WordCloud.

The most crucial parameter for most topic modeling algorithms is the number of topics to be generated. This parameter needs to be chosen carefully. We implemented each of the four topic modeling algorithms with the number of topics ranging from 2 to 20 to select the model with the optimal number of topics. For the implementation of BERT, the topic representation technique that we used was KeyBERTInspired, as it is more likely to reduce insignificant words such as stop words from the resulting topic representations and improve the overall coherence score. For creating dense clusters and identifying more interpretative topics, we incorporated C-TFIDF which is a class-based format of vector representations. The value of k was in the range of 5, 10, 15, and 20 topics.

For implementing all the pre-processing steps and topic modeling algorithms, the Google Colab environment was used. The programming language used was Python along with its most popular libraries including Numpy, Pandas, Gensim, Sklearn, BERTopic, Matplotlib and WordCloud.

The most crucial parameter for most topic modeling algorithms is the number of topics to be generated. This parameter needs to be chosen carefully. We implemented each of the four topic modeling algorithms with the number of topics ranging from 2 to 20 to select the model with the optimal number of topics. For the implementation of BERT, the topic representation technique that we used was KeyBERTInspired, as it is more likely to reduce insignificant words such as stop words from the resulting topic representations and improve the overall coherence score. For creating dense clusters and identifying more interpretative topics, we incorporated C-TFIDF which is a class-based format of vector representations. The value of k was in the range of 5, 10, 15, and 20 topics.

5 Results & Evaluation

The results of topic modeling can be interpreted in two ways: quantitatively and qualitatively. Qualitative evaluation involves examining the top terms for each cluster and determining their coherence. However, relying solely on qualitative evaluation is insufficient for identifying a topic model with the optimal number of topics. A quantitative value is also necessary to assess the quality of the generated topics. For this purpose, there are several scores available but the quantitative measure that we selected to find the best topic model is the coherence score.

The coherence score measure for topic modeling defines and interprets the degree of semantic similarity between high scoring words in a topic The reason for choosing coherence score is that it is easy to interpret and at the same time it can be used for selecting the optimal number of topics and it enables us to compare different topic models. Coherence score is the measure of semantic similarity between the top words of a topic cluster. It is computed by averaging the pairwise similarity scores of the top terms within a topic cluster [35]. The higher the coherence score, the better the quality of topics; conversely, the lower the coherence score, the worse the quality of topics.

We have implemented all four topic models from 2 to 20 topics and recorded respective coherence scores to find the optimal number of topics for each algorithm. The line chart of coherence scores of LDA plotted against the number of topics can be seen in Fig. 3. It can be observed from this chart that we got the maximum coherence score for 6 topics and after that coherence scores decreased. So, the optimal number of topics for LDA is 6.

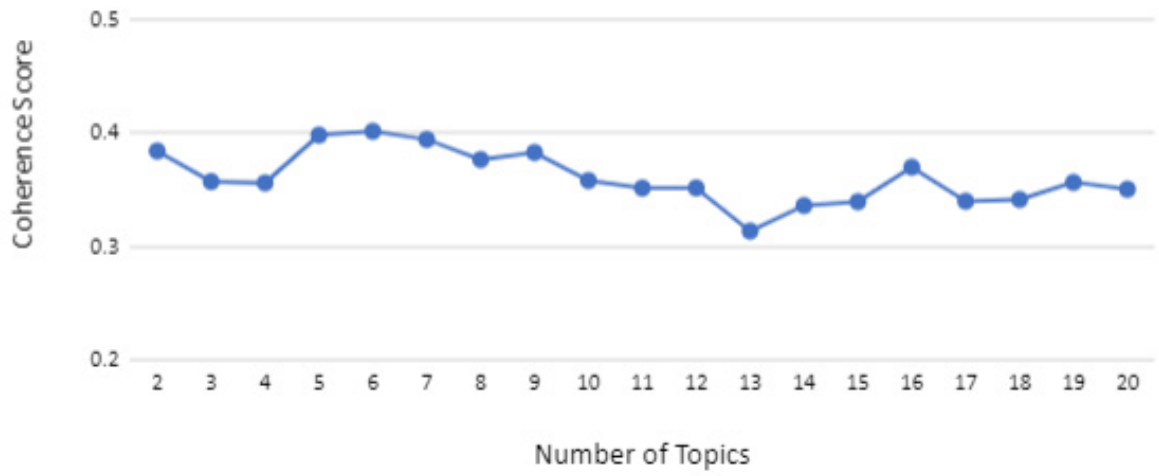

Fig. 3 Coherence Scores of LDA.

The line chart of coherence scores of LSA plotted against the number of topics can be seen in Fig. 4. It can be observed from this chart that we have got maximum coherence score for 3 topics and after that coherence scores are decreasing. So, the optimal number of topics for LSA is 3.

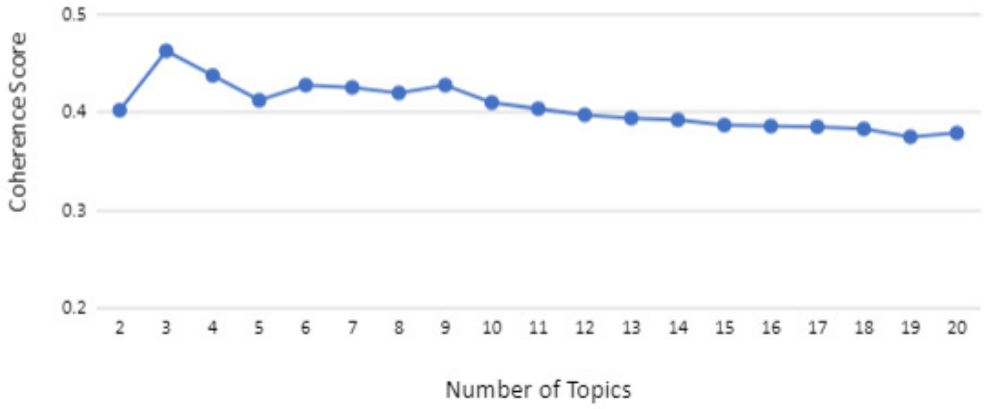

Fig. 4 Coherence Scores of LSA.

The line chart of coherence scores of NMF plotted against the number of topics can be seen in Fig. 5. It can be observed from this chart that we have got a maximum coherence score for 9 topics. So, the optimal number of topics for NMF is 9.

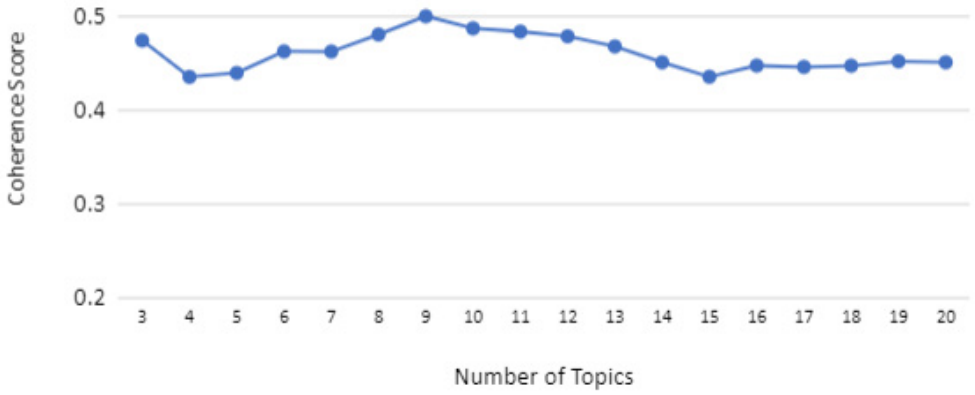

Fig. 5: Coherence Scores of NMF

The line chart of coherence scores of BERTopic plotted against the number of topics can be seen in Fig. 6. It can be observed from this graph that we have got a maximum coherence score for 9 topics. The coherence score then decreases slightly indicating that the optimal number of topics for BERTopic is 9.

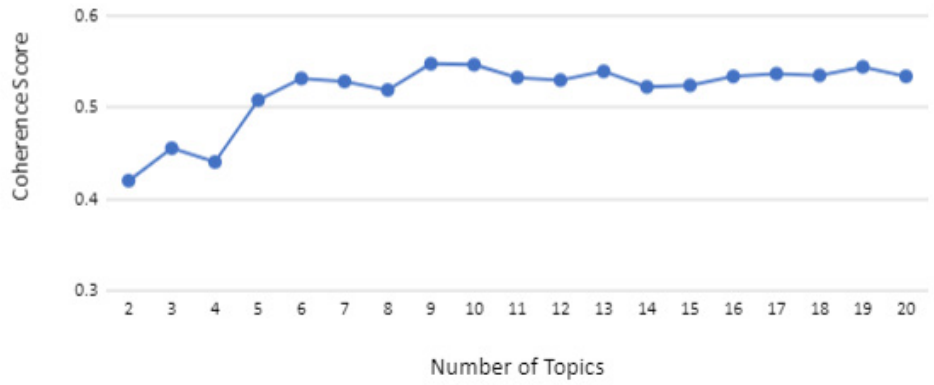

Fig.6 Coherence Scores of BERTopic

Fig. 7 presents a comparative line chart of the coherence scores for all the implemented algorithms. It can be observed that, overall, at different numbers of topics, BERTopic outperformed the traditional topic models—LDA, LSA, and NMF.

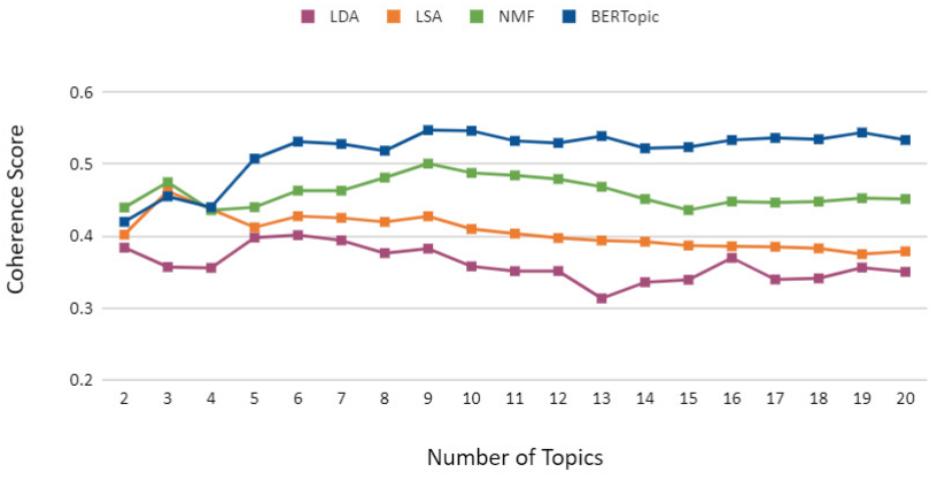

Fig. 7 A Comparison of Coherence Scores of Implemented Algorithms.

Table 3 presents a summary of the optimal number of topics along with the corresponding coherence score for each of the implemented algorithms. The highest coherence score for LDA was found to be 0.401 with 6 topics. For LSA, the highest coherence score was 0.462 with 3 topics. For NMF, the highest coherence score was 0.5 with 9 topics. Finally, BERTopic achieved the most promising coherence score of 0.547 with 9 topics, leaving the traditional topic modeling algorithms behind.

Table 3 Comparison of Different Topic Models.

Topic Model	Number of Topics	Coherence Score
LDA	6	0.401
LSA	3	0.462
NMF	9	0.5
BERTopic	9	0.547

This comparison of the implemented algorithms provides a clear indication that BERTopic also performed well for the low-resource Roman Urdu language. To further verify the efficiency of BERTopic, an open-source dataset was obtained from Kaggle [36,37]. We implemented all four algorithms with 5, 10, 15, and 20 topics for this dataset as well, aiming to find the best topic modeling algorithm.

Table 4 Topic Modeling Evaluation on Open-Source Dataset.

Topic Model	K=5	K=10	K=15	K=20
LDA	0.52	0.54	0.48	0.48
LSA	0.46	0.44	0.45	0.42
NMF	0.51	0.47	0.46	0.46
BERTopic	0.60	0.63	0.58	0.57

The analysis of data presented in Table 4 reveals that BERTopic consistently outperforms traditional topic modeling algorithms across various numbers of topics in the open-source dataset. This assessment suggests that like its effectiveness in resource-rich languages, BERTopic demonstrates a comparable advantage over conventional topic modeling approaches in the context of the Roman Urdu language. Hence, BERTopic can be considered an effective approach for topic modeling in Roman Urdu as well.

6 Discussion & Analysis

This section presents a brief discussion and analysis of the insights gained after implementing the process of topic modeling. Fig. 8 represents the nine optimal topics as produced by the BERTopic in the form of topic clusters. Each cluster represents one of the nine unique topics and within each cluster, every single dot represents a unique review of the dataset belonging to that cluster. The clusters are visually clear and well separated from each other, indicating their high quality and semantic similarity.

Fig. 8 Visual Representation of Topic Clusters.

Each of the nine topics extracted from the reviews represents an underlying theme or interest of the customers. Fig. 9 illustrates the top words of each topic cluster using bar charts to depict the frequency of these words. Analysis of these top words allows for the assignment of labels to each of the nine topics. For instance, in Topic 0, the prominent words include 'watch,' 'achi,' 'daraz,' 'thanks,' 'battery,' 'quality,' 'bohat,' and 'time.' Notably, the high frequency of 'watch' and 'achi' suggests that Topic 0 is associated with watches and has received positive feedback from customers.

Fig. 9 Top Words of the Topics Extracted.

Through a thorough analysis of the prominent words within each topic (as shown in Fig. 9), distinct names have been assigned to characterize each theme. The topics are labeled as follows: "Watch Reviews" (Topic 0), "Dress Reviews" (Topic 1), "Product Quality" (Topic 2), "Product Order" (Topic 3), "Product Display Picture" (Topic 4), "Shoes Reviews" (Topic 5), "Product Color" (Topic 6), "Shirt Reviews" (Topic 7), and "Product Size" (Topic 8). To assess the distribution of reviews within each topic, we assigned the corresponding topic name as the label to reviews associated with that specific topic.

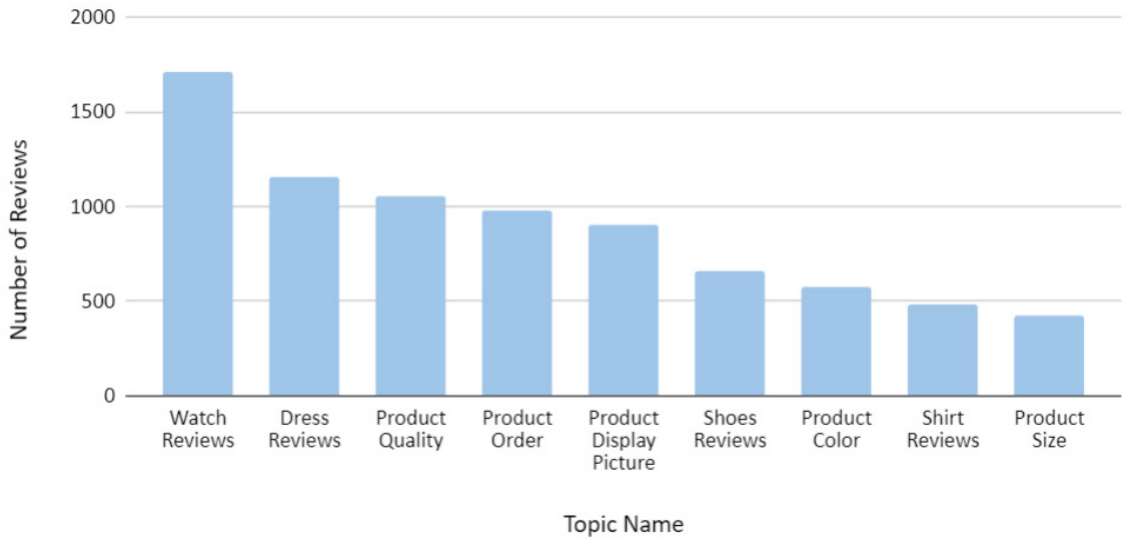

Fig. 10 Bar chart of the Number of Reviews in each Topic.

Fig. 10 depicts a bar chart representing the count of reviews for each of the nine topics, with the y-axis indicating the review count and the x-axis showing the topic names. The analysis reveals that the majority of reviews in the dataset are related to watches. Similarly, the second-largest category of reviews pertains to dresses. Conversely, the topic with the fewest reviews concerns product size.

Without manual review, the application of topic modeling allows for the analysis of major topics and themes within reviews. This methodology proves advantageous for the routine examination of extensive review datasets. Moreover, this approach bears significant managerial implications. By categorizing reviews according to their respective topics or major themes, it facilitates the allocation of related reviews to relevant departments for further analysis of customer feedback.

7 Conclusion and Future Work

The automation of review analysis is a dire need in this tech-oriented world. Resource-rich languages have several resources for automated analysis of reviews. However, Roman Urdu, being a resource-poor language, lacks such tools and has numerous unexplored areas in Natural Language Processing (NLP). Recognizing this significant gap in the literature, this study was conceived to conduct topic modeling on Roman Urdu product reviews. A dataset comprising 8,000 Roman Urdu reviews was collected from a popular online shopping website. Multiple data preprocessing and cleaning steps were applied to refine the data. For topic modeling, various algorithms, including LDA, LSA, NMF, and BERTopic, were employed. Among these, BERTopic demonstrated superior performance over other algorithms. The results were further evaluated using an open-source dataset, and insights derived from the original dataset were systematically analyzed. This evaluation gives a generalized framework based on BERTopic for efficiently performing Topic Modeling on text data in Roman Urdu language. This framework can be employed by online businesses to automate their review analysis process. Since the presented research attempt is based on a specific scope, future extensions can be made for further improvements. Firstly, to enhance the efficacy of this framework, incorporating additional product categories into the dataset could provide a more comprehensive understanding. Secondly, the current study has only focused on Roman Urdu reviews, but real-world e-commerce sites have a blend of Roman Urdu, English, and Urdu reviews. Our framework is not efficient for multilingual reviews as we have only trained our models on Roman Urdu data. Therefore, one more vital improvement that can be incorporated is making a more sophisticated framework that will be efficient for analyzing multilingual reviews. The proposed framework can be extended by adapting it to a multilingual context. Additionally, human-based validation of the generated topics can be carried out to provide qualitative insight into the model's interpretability.

8 Acknowledgments

We would like to thank the Department of Software Engineering, Mehran University of Engineering & Technology, for all the support provision of facilities necessary to accomplish this research work.

9 Availability of data and materials

The used raw data in this research is not publicly available. The data that support the findings of this research work are available from the corresponding author, on valid request due to privacy and ethical restrictions.

References

1. Khan, A. R., Karim, A., Sajjad, H., Kamiran, F., & Xu, J. (2022). A clustering framework for lexical normalization of Roman Urdu. Natural Language Engineering, 28(1), 93–123.
2. Sun, J., & Yan, L. (2023). Using topic modeling to understand comments in student evaluations of teaching. Discover Education, 2(1), 25.
3. Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. Electronic Commerce Research and Applications, 29, 142–156.
4. Farzadnia, S., & Vanani, I. R. (2022). Identification of opinion trends using sentiment analysis of airlines passengers' reviews. Journal of Air Transport Management, 103, 102232.
5. Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. Expert Systems with Applications, 116, 472–486.
6. Pathan, A. F., & Prakash, C. (2021). Unsupervised aspect extraction algorithm for opinion mining using topic modeling. Global Transitions Proceedings, 2(2), 492–499.
7. Sharifian-Attar, V., De, S., Jabbari, S., Li, J., Moss, H., & Johnson, J. (2022). Analysing longitudinal social science questionnaires: topic modelling with BERT-based embeddings. 2022 IEEE International Conference on Big Data (Big Data), 5558–5567.
8. Dahal, B., Kumar, S. A. P., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. Social Network Analysis and Mining, 9, 1–20.
9. Kwon, H.-J., Ban, H.-J., Jun, J.-K., & Kim, H.-S. (2021). Topic modeling and sentiment analysis of online review for airlines. Information, 12(2), 78.
10. Chu, K. E., Keikhosrokiani, P., & Asl, M. P. (2022). A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts. Pertanika Journal of Science & Technology, 30(4), 2535–2561.
11. Tahir, R., & Naeem, M. A. (2022). A Machine Learning based Approach to Identify User Interests

from Social Data. 2022 24th International Multitopic Conference (INMIC), 1–6.

12. Krishnan, A. (2023). Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. ArXiv Preprint ArXiv:2308.11520.

13. Hasib, K. M., Towhid, N. A., & Alam, M. G. R. (2021). Topic modeling and sentiment analysis using online reviews for bangladesh airlines. 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 428–434.

14. Yin, H., Song, X., Yang, S., & Li, J. (2022). Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web, 25(3), 1067–1083.

15. Ali, I., & Naeem, M. A. (2022). Identifying and Profiling User Interest over time using Social Data. 2022 24th International Multitopic Conference (INMIC), 1–6.

16. Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 353–362.

17. Zankadi, H., Idrissi, A., Daoudi, N., & Hilal, I. (2023). Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques. Education and Information Technologies, 28(5), 5567–5584.

18. Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. Applied Sciences, 13(2), 797.

19. Online shopping in Pakistan. (2024). https://www.daraz.pk

20. Daraz Code Mixed Product Reviews. (2024). https://www.kaggle.com/datasets/yrrebeere/daraz-code-mixed-product-reviews

21. Roman Urdu Stop Words. (2024). https://github.com/haseebelahi/roman-urdu-stopwords

22. Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. Procedia Computer Science, 152, 341–348.

23. Chandio, B., Shaikh, A., Bakhtyar, M., Alrizq, M., Baber, J., Sulaiman, A., Rajab, A., & Noor, W. (2022). Sentiment Analysis of Roman Urdu on E-Commerce Reviews UsingMachine Learning. CMES - Computer Modeling in Engineering and Sciences, 131(3), 1263–1287. https://doi.org/10.32604/cmes.2022.019535

24. Tusar, M. T. H. K., & Islam, M. T. (2021). A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data. 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 1–4.

25. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019). An overview of bag of words; importance, implementation, applications, and challenges. 2019 International Engineering Conference (IEC), 200–204.

26. Thavareesan, S., & Mahesan, S. (2019). Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. 2019 14th Conference on Industrial and Information Systems (ICIIS), 320–325.

27. Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. ACM Computing Surveys (CSUR), 54(7), 1–35.

28. Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic modelling twitter data with latent dirichlet allocation method. 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 386–390.

29. Wagire, A. A., Rathore, A. P. S., & Jain, R. (2020). Analysis and synthesis of Industry 4.0 research landscape: Using latent semantic analysis approach. Journal of Manufacturing Technology Management, 31(1), 31–51.

30. Shen, C., & Ho, J. (2020). Technology-enhanced learning in higher education: A bibliometric analysis with latent semantic approach. Computers in Human Behavior, 104, 106177.

31. Hamamoto, R., Takasawa, K., Machino, H., Kobayashi, K., Takahashi, S., Bolatkan, A., Shinkai, N., Sakai, A., Aoyama, R., & Yamada, M. (2022). Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. Briefings in Bioinformatics, 23(4), bbac246.

32. Fathi Hafshejani, S., & Moaberfard, Z. (2023). Initialization for non-negative matrix factorization: a comprehensive review. International Journal of Data Science and Analytics, 16(1), 119–134.

33. Samsir, S., Saragih, R. S., Subagio, S., Aditiya, R., & Watrianthos, R. (2023). BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory. JURNAL MEDIA INFORMATIKA BUDIDARMA, 7(3), 1514–1520.

34. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv Preprint ArXiv:2203.05794.

35. Chehal, D., Gupta, P., & Gulati, P. (2021). Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. Journal of Ambient Intelligence and Humanized Computing, 12, 5055–5070.

36. Hussain, N., Mirza, H. T., Ali, A., Iqbal, F., Hussain, I., & Kaleem, M. (2021). Spammer group detection and diversification of customers' reviews. PeerJ Computer Science, 7, e472.

37. Hussain, N., Mirza, H. T., Iqbal, F., Hussain, I., & Kaleem, M. (2021). Detecting spam product reviews in roman Urdu script. The Computer Journal, 64(3), 432–450.