# URDU TOPIC MODELING USING TRANSFORMER-BASED ATTENTION NETWORKS

SABIA KHAN

01-241211-008

BAHRIA UNIVERSITY ISLAMABAD

# URDU TOPIC MODELING USING TRANSFORMER-BASED ATTENTION NETWORKS

SABIA KHAN
01-241211-008

A thesis submitted in fulfillment of the
Requirements for the award of the degree of master of
Science (software engineering)

Department of Software Engineering

BAHRIA UNIVERSITY ISLAMABAD

MARCH 2023

# Approval for Examination

Scholar's Name: Sabia khan Registration No.: 01-241211-008 Program of Study: MS. (Software Engineering)

Thesis Title: Urdu Topic Modeling Using Transformer-Based Attention Networks

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for evaluation. I have also conducted a plagiarism test of this thesis using HEC-prescribed software and found a similarity index of 16% which is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor’s Signature: ____________________
Date: ______________________
Name: ______________________

# Author's Declaration

I, Sabia Khan, hereby state that my MS thesis titled "Urdu Topic Modeling Using Transformer-Based Attention Network" is my work and has not been submitted previously by me for taking any degree from this university Bahria University Islamabad or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Sabia Khan (01-241211-008)

Date: _______________

# Plagiarism Undertaking

I, Sabia Khan, solemnly declare that the research work presented in the thesis titled "Urdu Topic Modeling Using Transformer-Based Attention Network" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above-titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred to/cited.

I undertake that if I am found guilty of any formal plagiarism in the above-titled thesis even after the award of my MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Sabia Khan (01-241211-008)

Date: _______________

# Dedication

To my beloved parents.

# ACKNOWLEDGEMENTS

.

First and foremost, I am deeply grateful to Dr. Raja Suleman for his guidance towards completing my thesis when it seemed not possible at all. I would like to thank him for patiently guiding me throughout the past year. Despite his busy schedule, he accepted me into his supervision, and his oversight has resulted in the timely completion of this thesis.

I would like to thanks my parents, for their love and support, and for encouraging me in every situation, especially studies. It is my pleasure to dedicate this degree to my mother. A special thanks go to my friends, for helping me with their technical expertise in software development.

Sabia khan (01-241211-008)

Date: _______________

# ABSTRACT

Urdu, the national language of Pakistan and the one of the most widely spoken languages of the Indian sub-continent, is considered a Low-Resourced Language owing to the lack of digital resources available. There are different tasks that can be performed on a language using Natural Language Processing (NLP) which can help automate the understanding and generation of text in these languages. Topic Modeling is one such task that aims at discovering Topics (themes of discussion) within unstructured text. For Topic Modeling most of the researchers have focused on Latent Dirichlet Allocation (LDA) which is a statistical topic modeling technique to generate topics for Urdu language. Such techniques are quite useful for low-resourced languages as they require less amounts of data to train such models. However, Transformer-based models have become the recent state-of-the-art for many NLP tasks including Topic Modeling. The transformer-based modeling techniques have seen wide adoption because of the availability of a large number of pre-trained multi-lingual models. To the best of our knowledge, no research has exploited the benefit of using these Transformer-based models to perform topic modeling for Urdu language. Through this research we analyze and compare two Topic Modeling techniques LDA and Transformer-based (BERT multilingual) on the basis of their performance, coherence scores and topic generation. Our results show that Transformer-based models return a higher coherence score than the LDA model which means that the topics generated through such models are more interpretable by humans.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| NLP | Nature Language Processing |
| LDA | Latent Dirichlet Allocation |
| BERT | Bidirectional Encoder Representations from Transformers |
| TM | Topic Modeling |

# CHAPTER 1

# INTRODUCTION

Urdu, the national language of Pakistan and the one of the most widely spoken languages of the Indian sub-continent, is considered a low-resourced language [1]. This arises from the fact that the Urdu language suffers from a lack of digital resources. Such resources are crucial for successful implementation of Natural Language Processing (NLP) capabilities for any language. NLP is a sub-domain of Artificial Intelligence which is concerned with the machine enabled understanding and generation of natural languages. Natural Language Processing involves the process of parsing, extracting, and transforming natural language into chunks of information, which can provide valuable insights. NLP has found extensive application in text analysis and can simplify various tasks. For instance, chatbots, spellcheckers, voice controls, translation software, marketing automation, recruitment, encryption, and many more are some examples of areas where NLP is being employed. NLP can help with tasks such as sentiment classification, intellectual assistance, spam filtering, identifying fake information, and language translation [5][6].

Topic modeling is a significant aspect of Natural Language Understanding (NLU) that utilizes statistical data analysis to identify words that represent a particular or selective content when text data is utilized as input. Unlike association rules, which divide information into different segments, topic modeling identifies or extracts subjects by identifying patterns [3]. This technique is particularly useful when dealing with numerous files to determine the type of data they contain. When done manually, it is time-consuming, but Topic modeling streamlines and simplifies the process.

Natural Language Understanding (NLU) tasks involve analyzing language to comprehend its meaning, enabling it to be used for various tasks such as Sentiment Analysis, Intent Classification, Text Classification, and Topic Modeling [4]. Topic modeling is an example of unsupervised machine learning, where the algorithm is capable of identifying patterns without the need for labeling or tagging. Machine learning is a technique used to teach computers how to perform specific tasks effectively. In many real-world scenarios, machine learning has proven to be an excellent solution.

Text and speech are the two most common types of unstructured data that are rich in information but can be challenging to extract. NLP can help with tasks such as sentiment analysis, topic detection, language detection, and text classification. Topic Modeling is an NLP task that clusters documents into topics based on their contextual and semantic similarities. Topic Modeling involves different models such as Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), and Hierarchical Dirichlet Process (HDP) [7]. LDA is a probabilistic LSI (PLSI) that is parametric Bayesian and has been widely used to generate topics for the Urdu language, LSI is a topic modeling approach that uses low-rank approximation over Single Value Decomposition (SVD), whereas Hierarchical Dirichlet Process (HDP) is a Bayesian nonparametric model that is used for topic modeling, which is a subfield of natural language processing (NLP). It is an extension of the Latent Dirichlet Allocation (LDA) model that allows for the creation of an arbitrary number of topics from a given corpus of documents without having to specify the number of topics beforehand [9].

In recent times, Transformer-based models have emerged as state-of-the-art models for many NLP tasks, including Topic Modeling. Transformer-based models employ self-attention mechanisms, enabling them to assign different weights to input data based on its significance. Due to the availability of numerous pre-trained models with accurate representations of words and sentences, Transformer-based models have gained widespread adoption [2].

## 1.1 Research Gap

This thesis identifies a research gap in the area of transformer-based attention networks for topic modelling in Urdu. Despite recent advancements in natural language processing and deep learning, there has been limited research in applying these techniques to Urdu language data. This research gap presents a significant challenge in automatic analysis and understanding of Urdu text data, which is widely available in various domains.

## 1.2 Problem Statement

The problem addressed in this thesis is the lack of research on Urdu topic modeling using transformer-based attention networks. With the increasing amount of digital text data available, there is a growing need for automatic analysis and understanding of such data. One of the key challenges in this regard is topic modeling, which involves the identification of the underlying themes or topics within a large corpus of text data. The objective of this study is to investigate the effectiveness of transformer-based models in identifying and analyzing the topics within Urdu text data.

## 1.3 Research Questions

**RQ1:** What are different Topic Modeling techniques in Natural Language Processing?

**RQ2:** How do statistical and Transformer-based techniques generate Topic Models?

**RQ3:** How does the performance of statistical Topic Modeling (LDA) compare to Transformer-based Topic Modeling (BERTopic) in terms of accuracy and computational efficiency?

## 1.4 Research Contribution

This research compares the performance of statistical analysis-based LDA models and Transformer-based models for Topic Modeling. The key contributions of the research are to analyze the performance of each technique based on different evaluation criteria and provide a detailed discussion of the findings. This experiment is carried out on benchmark dataset Urdu news 1M dataset.

## 1.5 Thesis Organization

This document has been structured in the following chapters. Chapter 2 present the literature and effort of the topic modeling for Urdu language. Chapter 3 introduced the proposed methodology, which intended the dataset, utilization of Bertopic and gensim. Chapter 4 includes results and discussion section and chapter 5 contains conclusion and perspectives in which describe the future direction and research on this and similar topics.

# CHAPTER 2

# LITERATURE REVIEW

Natural Language Processing (NLP) models are automated models that primarily analyze text data [11]. NLP comprises two main subcategories: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU aims to comprehend language on various levels of complexity from syntax to semantics, while NLG is concerned with generating new text. Topic modeling is a popular NLP task that clusters documents into topics based on contextual similarity. Latent Dirichlet Allocation (LDA) is a statistical topic modeling technique commonly used to generate topics in the Urdu language. Although LDA has been successful, recent years have seen Transformer-based models emerge as the state-of-the-art for many NLP tasks, including topic modeling. Transformers leverage Attention Mechanisms, which provide context for input sequence items and improve training times.

In [12], the author used LDA to analyze 51,346 abstracts from 23 prestigious structural engineering journals published between 2000 and 2022. The LDA analysis resulted in 50 distinguishable word clouds centered on individual research themes and assigned unique topic names. Advanced metrics were used to analyze research similarity and progress across different journals and regions/countries, enabling community stakeholders to explore the state of research and develop effective strategies.

In [13], the main objective was to determine the coherence and interpretability of LDA generated topics. Due to the COVID-19 pandemic, Twitter users sought different types of information, necessitating the application of unique techniques to analyze their messages' length, style, and slang use. The study aimed to determine the most suitable topic related to vaccination using topic modeling techniques. The result concluded that LDA offers a very good interpretation of topics.

In [14], the researcher used a probability-topic modeling approach to identify the core topics in finance by extracting approximately 15 coherent topics from 5,942 academic studies from
1990 to 2020 using LDA. They classified the topics into four categories based on their content.
This study provides a structured topography for finance researchers seeking to incorporate machine learning research approaches in their survey of finance phenomena.

In [15], the author proposed an effective topic modeling technique to extract topics from Urdu documents' morphological structure. They proposed a topic model for Urdu languages and named it Urdu Latent Dirichlet Allocation (ULDA), using the standard LDA model. The proposed model showed greater efficacy than other similar works.

In [16], the author proposed a semi-supervised framework for Urdu document clustering, combining pre-processing techniques, a seeded-LDA model (Seeded-ULDA), and Gibbs sampling. The proposed model was tested on the Urdu news dataset under two conditions: a dataset without overlapping and a dataset with overlapping. The result indicated that unsupervised models like LDA, NMF, and k-means performed well on the dataset without overlapping, but not on the dataset with overlapping. The proposed semi-supervised model, Seeded-ULDA, performed well on both datasets, instructing topic models to find topics of specific interest.

In [17], the author proposed using a restructured classification system and LDA-based topic modeling to categorize and modify search results for academic research papers. By examining the distribution of document weight in themes, system efficiency can be increased.

In [18], the study focused on text classification in low-resource languages like Urdu. Six medium-sized datasets with six categories were collected, and various methods were utilized, including Chi-2, linear discriminant analysis, and term frequency-inverse document frequency (TFIDF). Training-test data comprised 70% of the test data, and machine learning and deep dense neural network (DDNW) approaches were incorporated. The procedure also incorporated machine learning and deep dense neural network (DDNW) approaches. Finally, Naive Bayes, XGBoost, Bagging, and DDNW were utilized; Bagging and DDNW fared better than other algorithms, with Bagging having a 95% f1 score and DDNW having a 92% score.

This proposal paper [19] focused on the test that was carried out to identify user behavior on Twitter and assist decision-makers. Word2Vec, Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA) are some of the methods used to identify this. Performance evaluation methods included K-Means. Tweets from the Turkey earthquake were used in this case study. Tweets were grouped under fifteen different hashtags, and the aforementioned procedures were used. While Word2Vec performed well with small data sets, LDA performed better with medium and large data sets than Word2Vec and LSA.

In [20], the author proposed that in order to make a cluster human-readable, it needed to be given a meaningful term. The suggested method was contrasted with two graph-based techniques—TextRank and PositionRank—as well as Z-Order, M-Order, T-Order, and YAKE, as well as four other statistically-based methods. The technique made use of the Headlinesnk and PositionRank Urdu dataset. Human assessors and the Jaccard index were used to measure the outcome. In contrast to other methods, the labels created through this technique were found to be more relevant and semantically rich.

In [21], the paper aimed to help users find the information they were seeking in Urdu newspapers. To achieve similar or better results, techniques such as NLP (pre-processing), TFIDF, and BERT were applied. However, BERT performed better than TF-IDF in terms of results. The user was advised when similarity was greater than 60%.

In [22], careful planning was required to anticipate accident circumstances. The Pegasus layer model was employed in this study to explain traffic incidents in Korea. Three different types of accident situations were identified according to frequency: typical traffic, critical traffic, and edge case. The edge case was used to apply Topic Modeling (the ones least likely to occur and harder to predict).

In [23], GDP was described as a macroeconomic measure that required data for approaches that make short-term predictions. Although most of the data comes from private sources, in the event of an occurrence like COVID, this data may not be complete. The economic policy uncertainty (EPU) index, which was based on the Topic Modeling of newspapers, was used to quantify uncertainty. By calculating the EPU index, which was a quick and effective method for Topic Modeling of digital news based on semantic clustering using word embedding, the index could be updated in real-time, thus reducing the time needed to document assignments into subjects.

In [24], the author proposed that understanding and analyzing trustworthy latent subjects in online discussion texts that predominated with little word co-occurrence had always been difficult. Such works for Urdu text were lacking in the extant literature. The study presented experiments on 0.8 million Urdu tweets using Latent Dirichlet Allocation (LDA), Nonnegative Matrix Factorization (NMF), Probabilistic Latent Semantic Analysis (PLSA), and Latent Semantic Analysis (LSA). The authors produced the three types of the collected dataset, preprocessed the text of the tweets, and extracted several features to represent documents on various n-grams. The observed results showed that LDA performed best with merging, whereas NMF outperformed the approaches with TF-IDF feature vectors in Urdu tweets text.

In [25], the authors presented various coherent topics extracted from Urdu text. Urdu being a low-resourced language, unsupervised models that work well with English documents are not very effective with Urdu. There is limited work available on Urdu language. The authors introduced a semi-supervised topic model named "seeded-Urdu LDA" specifically for Urdu language that produces coherent topics, taking into account the morphological structure of the Urdu language. The study indicated that word embedding was not sufficient for extracting coherent topics in Urdu language. Therefore, the proposed Seeded-ULDA model was compared to the existing ULDA model based on coherent measures, and was found to be 39% more effective.

The assessment of self-attention mechanisms in [17] and [27] deviated from traditional recurrent architectures, which involve token prediction followed by Masked Language Modeling (MLM). In a MLM experiment, Goldberg [28] demonstrated that BERT consistently assigned higher scores to the correct verb forms than to the incorrect ones. Additionally, BERT's ability to capture relationships between sentences was attributed to another underlying mechanism called NSP (Next Sentence Prediction) [29].

Overall, these studies demonstrate various techniques for Natural Language Processing in Urdu, including Topic Modeling, semantic clustering, and self-attention mechanisms. They also highlight the challenges associated with low-resource languages such as Urdu and the importance of developing tailored models and techniques to address them.

The below table define the summary of literature review.

**Table 2. 1** Summary table

| Reference | Title | Year | Dataset | Description |
|---|---|---|---|---|
| [12] | The twenty-first century of structural engineering | 2022 | 51,346 article abstracts from 2000 to 2020 year. | The author proposed a LDA, topic modeling approach, to |

| | | | | |
|---|---|---|---|---|
| | research: A topic modeling approach | | | analysis articles abstract from 23 prestigious journal in structural engineering. Whereas the results shows tha LDA successfully identified 50 research topics that the current state of research in the community. |
| [13] | Topic Modeling Technique on Covid19 Tweets in Serbian | 2022 | Covid19 vaccine 1,768 tweets | The author main objective was to determine which topic appear in tweets related vaccination. As when they compare results they came up that LDA method provides a very |

| | | | | |
|---|---|---|---|---|
| | | | | good interpretation of the topics. |
| [14] | Machine learning in finance: A topic modeling approach | 2022 | Elsevier Scopus database | The researcher identifies the core topics applying machine learning to finance. Through LDA topic modeling approach they extract coherent topic. |
| [15] | A framework of Urdu topic modeling using latent Dirichlet allocation (LDA) | 2022 | Urdu headlines news (such as BBC, Nawa-i-Waqt and Jang) | They proposed a topic model for Urdu languages as they used the standard LDA model therefore they named it Urdu Latent Dirichlet Allocation (ULDA). The result shows the efficacy of this model |

| | | | | |
|---|---|---|---|---|
| [16] | Urdu Documents Clustering with Unsupervised and Semi-Supervised Probabilistic Topic Modeling | 2022 | Urdu news datasets | The author proposed that Urdu is a less resourced language. The semi-supervised model, Seeded-ULDA, provides significant results as compared to unsupervised algorithms |
| [17] | A Suggestion on the LDA-Based Topic Modeling Technique Based on Elastic Search for Indexing Academic Research Results | 2022 | the ElasticSearch classification method and topic-based LDA model were applied to extract the characteristics of academic papers | They proposed best method for doing the use of restructured classification system and topic modelling that is based on LDA. Additionally, by examining the distribution of Document weight in topics, system efficiency was |

| | | | | |
|---|---|---|---|---|
| | | | | proven excellent. |
| [18] | Benchmarking Performance of Document Level Classification and Topic Modeling | 2022 | Pakistani news sources | The text classification is a low-resource language like Urdu is always difficult. The procedure also incorporates machine learning and deep dense neural network (DDNW) approaches, finally they came up with Bagging and DDNW fared better than other algorithm. |
| [19] | Comparative analysis with topic modeling and word embedding methods after the Aegean Sea earthquake on Twitter | 2022 | Tweets were gathered after 6.6 magnitude earthquake in October 30, 2020 | This study was carried out to identify user behavior on Twitter and assist decision-makers. Word2Vec, (LDA), and |

| | | | | |
|---|---|---|---|---|
| | | | | latent semantic analysis are some of the methods used to identify this, While Word2Vec performs well with small data sets, LDA often performs better with medium and large data sets than Word2Vec and LSA |
| [20] | Overview of the Transformer-based Models for NLP Tasks | 2020 | large number of raw texts are available online (e.g. Wikipedia, Web blogs, Reddit) | This paper reviews the architecture of Transformers and the common transformers used in the industry today. |
| [21] | Attention is All You Need | 2017 | WMT 2014 English-German dataset consisting of | A groundbreaking research work that proposed the mechanisms |

| | | | | |
|---|---|---|---|---|
| | | | about 4.5 million sentence pairs and for English-French, they used the significantly larger WMT 2014 dataset consisting of 36M sentences | behind Transformers. |

# CHAPTER 3

# RESEARCH METHODOLOGY

## 3.1 Research Methodology

We have used transformer-based models such as BERT as they have shown amazing results in various NLP tasks over the last few years. BERT provides a wide range of pre-trained models to help getting started with different NLP tasks. Pre-trained models are especially helpful as they contain accurate representations of words and sentences from a large corpus.

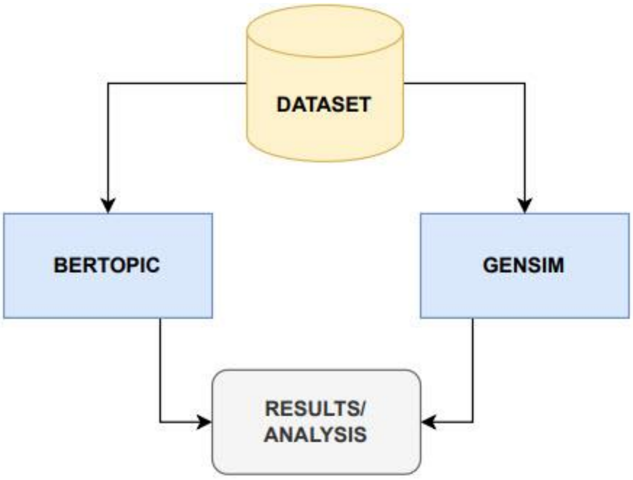

**Figure 3. 1** Proposed Design

However, to implement model we have used Urdu new dataset. After that We utilized a pre-trained transformer model along with an open-source Topic Modeling library; Bertopic, to automatically generate topic models for our dataset [30]. We also utilized Gensim (LDA); another open-source library, to perform unsupervised statistical analysis-based topic modeling [31]. After implementing both models then we Compared and analyzed the results and performance of each technique. We also focused on human judgments analysis because at the end it about human that how good topics are interpretable by humans.

### 3.1.1 BERT steps

Here are the detailed steps for performing topic modeling using BERTopic and analyzing the results [4]:

Preprocessing: Before running the BERTopic algorithm, the text data needs to be preprocessed to remove stop words, punctuation, and other noise, and lemmatized to normalize the words. BERTopic provides built-in preprocessing functions.

Vectorization: BERTopic uses the pre-trained BERT model to generate embeddings for each document in the corpus. This is done using the BERTopic.embed() function, which takes the preprocessed text data as input and returns a matrix of embeddings.

Topic modeling: The BERTopic algorithm is then applied to the embeddings matrix to generate the topic clusters. This is done using the BERTopic() function, which takes the embeddings matrix as input and returns a BERTopic object containing the topic clusters and their associated documents.

Topic visualization: The resulting topics can be visualized using various techniques, such as word clouds or scatter plots. BERTopic provides a built-in function for generating a scatter plot of the topics, which can be customized using various parameters such as the number of topics to display, the color scheme, and the distance metric.

Analysis: At the end we analysis the results of bert model with the existing model and made a visualization of each of results so that it would be easily interpreted by human.

However, the below are the detailed steps for performing topic modeling using LDA and analyzing the results:

### 3.1.2 LDA steps

Preprocessing: Before running the LDA algorithm, the text data needs to be preprocessed to remove stop words, punctuation, and other noise, and lemmatized to normalize the words. This involves steps such as tokenization, lowercasing, and stemming/lemmatization. There are many libraries available for this purpose in various programming languages, such as NLTK for Python.

Vectorization: The preprocessed text data is then transformed into a numerical format that can be used for topic modeling. This is done using techniques such as bag-of-words or term frequency-inverse document frequency (TF-IDF) vectorization, which represent each document as a vector of word counts or weights. Many libraries are available for this purpose, such as scikit-learn for Python.

Topic modeling: The LDA algorithm is then applied to the vectorized text data to generate the topic clusters. This is done using libraries such as Gensim for Python. The LDA algorithm requires specifying the number of topics to generate, as well as other hyperparameters such as the number of iterations and the alpha and beta value.

Topic visualization: The resulting topics can be visualized using various techniques, such as word clouds or bar charts. This involves examining the most frequently occurring words associated with each topic and identifying any notable patterns or trends. Libraries such as pyLDAvis provide interactive visualizations of LDA topics.

Analysis: Once the topics have been generated and visualized, they can be analyzed to gain understandings into the underlying patterns and themes in the text data. We can Start by reviewing the topics that were generated by the LDA algorithm. Look at the top keywords associated with each topic and consider whether they form coherent themes or not. Compare the topic, interpret the topic and evaluate the quality of LDA topic use a matric coherence score, through all these we had done comprehensive analysis of the LDA and Bert model.

## 3.2 Dataset

For our study, we used the Urdu News Dataset 1M by Hussain et al. [32] [33]. The documents in the dataset are divided into four news categories, namely: business and economics, science and technology, entertainment, and sports. Although the dataset contains 1 million news items, the dataset was sampled for computational resource reasons. The text contained only alphabetic, numeric, and symbolic words. In the context of using pre-trained language representation models like BERT, preprocessing is unnecessary. This is because BERT incorporates a multi-head self-attention mechanism that enables it to leverage all the information in a sentence, including punctuation and stop-words, from various perspectives. Therefore, BERT can effectively process raw text without the need for any additional preprocessing steps.

## 3.3 Transformer-Based Model

The Transformer model operates as follows:

1. Each word of an input sequence is transformed into a dmodel-dimensional embedding vector.
2. Positional information is incorporated into the embedding vectors by adding a positional encoding vector of the same length.
3. The Transformer encoder, consisting of two sub-layers, takes in the augmented embedding vectors. The encoder is bidirectional, attending to all terms in the input sequence, regardless of their position.
4. At time step t-1, the decoder is provided with its own previous output as input.
5. Similar to the encoder, positional encoding is applied to the decoder input.
6. The decoder unit consists of three sub-layers. In the first sub-layer, masking is applied to prevent the decoder from attending to future positions. The second sub-layer takes in the output of the encoder, allowing the decoder to attend to all positions in the input sequence.
7. Finally, a convolutional neural network followed by a Softmax layer is used to generate an output vector estimating the validity of the output sequence [34].

### 3.3.1 Transformer Architecture and Its Adaptation

Transformers were introduced in 2017 and have proven to be a valuable tool for natural language processing (NLP) tasks. Unlike sequential models, Transformer allow for more parallelism and faster training as data does not need to be fed into it sequentially. Transformers employ a semi-supervised learning approach in training, combining a large amount of unsupervised, unlabeled data with a smaller amount of labeled data.

Transformers have found extensive use in a variety of NLP tasks, including machine translation, time series prediction, document generation, named entity recognition, and biological sequence analysis [34, 35, 38]. Figure 3,1 illustrates the building block of the Transformer model, which comprises three primary working units: embedding, encoder decoder, and output generation.

In the embedding layer, each word in a sentence is converted into a word embedding vector, which is then augmented with a positional encoding vector of the same dimension ranging from -1 to 1 [35]. This results in a vector that contains all the necessary information about the word sequence and distance between different words in the input sentence. The resulting vector is then fed into the encoder unit for further processing.

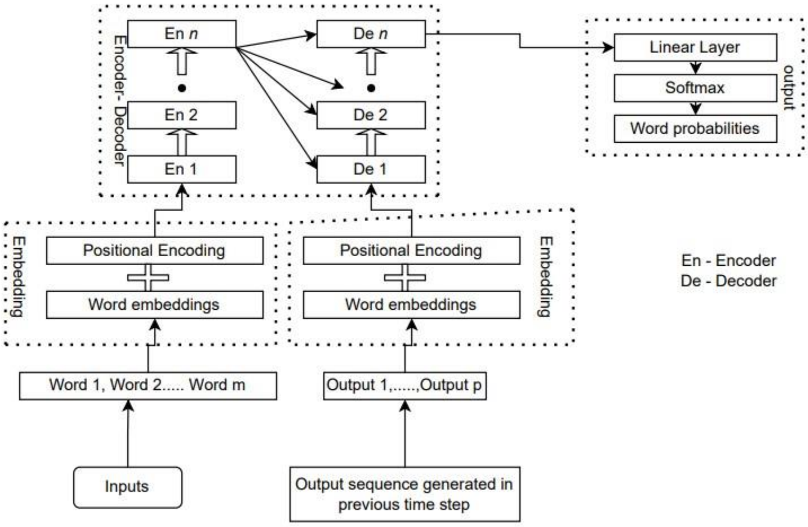

**Figure 3. 2** A fundamental component of a deep learning transformer model

The transformer model's core, the encoder-decoder module, manages the attention mechanism and enhances the performance of NMT. It is necessary for the number of encoders and decoders to be equal, even though they are independently stacked. Fig. 3.2 depicts the layers of an encoder and decoder [35]. The encoder consists of two sub-layers: Self-Attention and Feed Forward Neural Network (FFNN). Similarly, the decoder is divided into three sub-layers: SelfAttention, Encoder-Decoder Attention, and FFNN. The encoder processes a fixed-size vector list as input through the self-attention and FNN layers. The output of a lower-level encoder is passed as input to its adjacent higher-level encoder in a cascade fashion, continuing until the topmost level of the encoder. The output of the top encoder is transformed into attention vectors K & V. Each decoder receives these vectors, which are then utilized by "Encoder-Decoder Atdone in timestep 1." For the decoder to predict the following word in the subsequent time step, this output serves as an additional input. The output sentence is formed when word generation continues up until the encounter of a special character [36]. The decoder output vector is simply converted into some probabilistic values by the Linear and Softmax layers, and these values assist the model in generating the subsequent token.

The transformer model's hyperparameters need to be adjusted for optimal performance. The transformer model includes a number of hyperparameters, including batch size, dropout, learning rate, the number of encoder-decoder layers, the number of heads, and others [36]. The capacity of the model to guess what other words are referring to a specific word that is presently being processed is improved by having a larger number of heads in the attention layer.

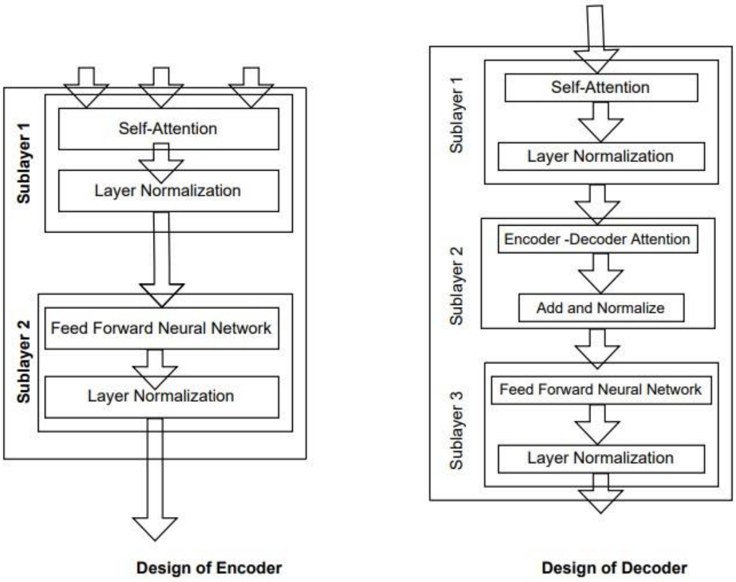

**Figure 3. 3** The architecture of the encoder and decoder in the transformer model

.

## 3.4 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a transformers model that has been pretrained on a large corpus of multilingual data in an unsupervised manner [36]. This indicates that it was pretrained using simply the raw texts, with no human labelling of any kind (thus, it can use a lot of material that is readily accessible to the public), and with an automatic procedure to generate inputs and labels from those texts.

It was specifically pretrained with two objectives.:

The following are two techniques used in language modeling:

1. Masked Language Modeling (MLM): The model randomly selects 15% of the words in a sentence and hides them, and then predicts the missing words [37]. Unlike conventional RNNs that process words sequentially, and autoregressive models like GPT that internally mask the next tokens, MLM allows the model to learn a two-way representation of the sentence.
2. Next Sentence Prediction (NSP): During training, the model combines two masked words as inputs. Sometimes, the two words belong to consecutive sentences in the original text, and other times they don't. The model must determine whether or not the two sentences were next to each other [37].

### 3.4.1 BERTopic library

With the use of transformers and the c-TF-IDF (class-based TF-IDF), the topic modelling library BERTopic uses complex groups to produce topics that are simple to understand while preserving key terms from the subject specifications [38]. Directed, (semi-)supervised, multilevel, dynamic, and interactive simulations are supported by BERTopic.

Three phases are used by BERTopic to develop subject descriptions. Each text is first transformed using a developed language model into its embedded version. The complexity of the generated embeddings is then decreased prior to grouping these embeddings in order to improve the grouping procedure. Finally, concept descriptions are retrieved from the file sets using a customized class-based form of TF-IDF [38].

#### 3.4.1.1 Document embedding

By embedding documents, we can produce interpretations in vector space in BERTopic that can be contrasted contextually. It is presumed that texts dealing with the same subject are highly interrelated. Thus, rather than directly producing the subjects, such embeddings are mainly employed to group documents that share comparable semantic properties [35]. If the machine learning model that created the document embeddings was adjusted for semantic relatedness, then any other embedding method was able for this reason. As newer and better word modeling are created, the efficiency of grouping in BERTopic will therefore enhance. This enables BERTopic to advance along with the state-of-the-art embed methodologies.

#### 3.4.1.2 Document clustering

It has been demonstrated that the space between the closest and farthest sets of data approaches as dataset dimensions rise (Aggarwal et al., 2001; Beyer et al., 1999). As an outcome, in highdimensional area, the idea of spatial proximity is ill-defined and there are few significant differences in distance measurements [39].

#### 3.4.1.3 Topic Representation

The documents in every group are used to describe the subject interpretations, and every group will be given a specific subject. We are interested in every title's distinctive characteristics depending on the distribution of cluster words per subject. For this reason, we can alter TFIDF, a metric for quantifying the significance of a message within a text, to indicate the significance of a title within a concept. This process is generalized to groups of documents. First, by just appending the texts, we consider each item in a group as a separate file. Then, by converting docs to groups, TF-IDF is modified to carefully consider this interpretation. We can lower the total quantity of subjects to a user-specified amount by repeatedly combining the cTF-IDF models of the lowest prevalent subject with its most comparable counterpart [39].

#### 3.4.1.4 Dynamic Topic Modeling

Traditional topic modelling methods are static in character and do not support modelling of papers that are structured chronologically. By simulating how concepts may have changed over time and the degree to which text visualizations indicate that, dynamic topic modeling approaches, first presented by (Blei and Lafferty, 2006) as an application of LDA, circumvent this. By utilising the c-TF-IDF descriptions of subject in BERTopic, we are able to simulate this behavior [39].

1. Mainly, our experiment is based on the open-source BERTopic model.

2. As BERT is based on the transformer mechanism, it requires a substantial corpus for effective training. Its initial training was performed using 3.3 billion words collected from the vast English Wikipedia and the Book Corpus.
3. One of the advantages of BERTopic over LDA is that it does not require predefining the number of topics, as it can extract the number of topics mentioned in the documents.

## 3.5 Gensim library

Gensim ("Generate Similar") is a python-based open-source framework for unsupervised topic modeling and NLP [40]. Gensim transforms documents from one vector representation into another. It is designed to handle large text collections. This process serves two goals: To bring out hidden structure in the corpus, discover relationships between words and use them to describe the documents in a new and more semantic way, also to make the document representation more compact and this will lead to improving efficiency and efficacy.

Advanced statistical machine learning and best academic algorithms are used to carry out a variety of difficult projects, including

- Creating word or text sequences
- Corpora
- Recognition of the subject
- Correlation of files (retrieving semantically similar documents)
- Evaluating the core concept of simple texts

Gensim, a Python and Cython implementation that is based on Numpy and scipy packages, is built to deal with big textual sets employing progressive interactive techniques as well as information broadcasting, in addition to carrying out the aforementioned complicated duties. This sets it apart from software for machine learning programs that only focus on in-memory computation [40].

Following are some of the features and capabilities offered by Gensim

### 3.5.1 Scalability

Gensim's incremental online training methods enable it to process huge and web-scale corpora with ease. It is scalable since the entire input corpus does not have to be kept in Random Access Memory (RAM) at once. In other words, regardless of the size of the corpus, all of its methods are memory-independent [41].

### 3.5.2 Robust

Gensim is a powerful program that has been used for more than 4 years in a variety of systems by a wide range of people and organizations. We can quickly insert in our own data stream or input corpus. Additionally, adding other Vector Space Algorithms to it is quite simple [41].

### 3.5.3 Platform Agnostic

Since Gensim is written entirely in Python, it may be used on any platform that supports Python and Numpy, including Windows, Mac OS, and Linux.

### 3.5.4 Efficient Multicore Implementations

Gensim offers multicore effective versions of several well-known techniques, such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Random Projections (RP), and Hierarchical Dirichlet Process, to speed up processing and retrieval on computer clusters (HDP) [42].

### 3.5.5 Advantages of Gensim

The following are some of Gensim's key benefits:

- Although other programming languages like "scikit-learn" and "R" may offer topic modeling and transfer learning capabilities, Gensim's capabilities are unmatched in these areas. Additionally, it offers textual data centers that are more practical.

- Gensim’s ability to process enormous documents even without putting the entire document into storage is one of its biggest advantages.
- Gensim employs unsupervised modeling instead of expensive notes or traditional manual labeling [43].

# CHAPTER 4

# RESULTS AND DISCUSSION

In this section, we describe the experimental outcomes of the proposed study on the benchmark dataset. We run a comparative analysis of statistical analysis-based model LDA and transformer-based models BERT for Topic Modeling.

In this research work, we have analyzed the performance of each of these techniques based on different evaluation criteria such as coherence score and the number of topics generated. Coherence measures how semantically coherent the topics are by comparing the similarity of the top words in the topic with each other and across different topics. The transformer-based modeling techniques have seen wide adoption because of the availability of a large number of pre-trained multi-lingual models. Pre-trained models are especially helpful as they contain accurate representations of words and sentences from a large corpus.

For analyzing and comparing the results using BERTopic and LDA, the Urdu news dataset was used. This data has been divided into four different categories such as science and technology, business and economics, entertainment, and sports. Although we did not use Guided or SemiSupervised Topic Modeling, these categories were used to evaluate the results from an understanding point of view. The dataset was sampled due to the lack of computational resource reasons.

## 4.1 Models Evaluation and Interpretation

Transformer-based model was compared with LDA. For a fair comparison, we compared these models on the bases of the performance of the coherence score and the number of topics generated. Since LDA requires the number of topics to be initialized beforehand, we chose the same number as the number of topics generated by the Transformer-based model.

## 4.2 Results and Comparison

For the Urdu News dataset, two of the significant topic modeling algorithms were implemented, namely, LDA and c-TF-IDF with Transformer-based model. The both models were trained on 60k dataset. All the preprocessing approaches, such as tokenization, raw text, stop word removal as well as lemmatization were applied only for the data used for LDA. As mentioned previously, Transformer-based models like BERT do not require explicit preprocessing to be performed on the data.

## 4.3 Comparison of BERTopic and Gensim with respect to coherence and topic score

**Table 4. 1** Results on Urdu dataset

| S.no | Evaluation | BERTopic | LDA |
|---|---|---|---|
| 1. | Number of topics | 526 | 526* |
| 2. | Coherence score | 0.52 | 0.45 |

* We manually assigned this number of topics for comparison purposes

The first topic model was generated for using the Transformer-based architecture using the BERTopic library. We used the BERT-multilingual pre-trained model as it has been trained over 100+ language including Urdu. BERTopic generated 526 topics in accordance with the automated model, with a coherence score of *0.52*. While in the case of LDA, 526 topics were allocated, and the coherence score for the model was *0.45*.

In assessing the performance of BERTopic and LDA, it was found that BERTopic outperformed LDA in terms of coherence score, indicating that the automated matrices were able to accurately cluster topic-related words based on their semantic similarity. BERT models are known to be more powerful and accurate than LDA models for various NLP tasks, including topic modeling, due to their ability to capture complex patterns and relationships between words and sentences by leveraging a large amount of text data. However, BERT models typically require more computational resources than LDA models, which can affect their training and inference speed. The performance, coherence score, heatmap, and similarity matrix of both models are presented below.

### 4.4.1 Topic word score results BERT vs LDA

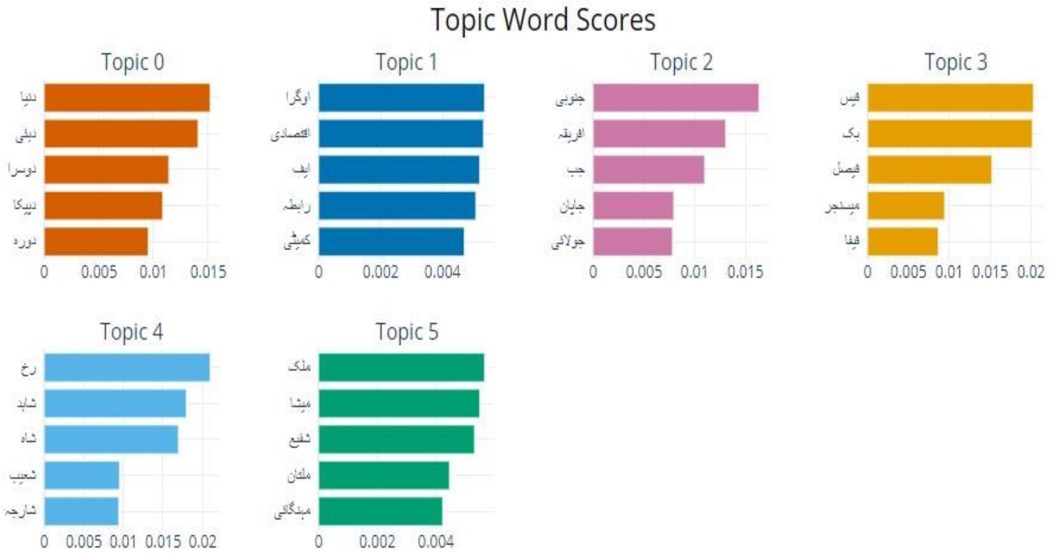

**Figure 4. 1** BERTopic word score

BERTopic calculates a coherence score for each topic, which measures how closely related the words in the topic are to each other based on their embeddings. However, in this way, we interpret the words present in the topic easily. Gensim uses a statistical measure called topic coherence, which evaluates the degree of semantic similarity between words in a topic based on their co-occurrence within documents.

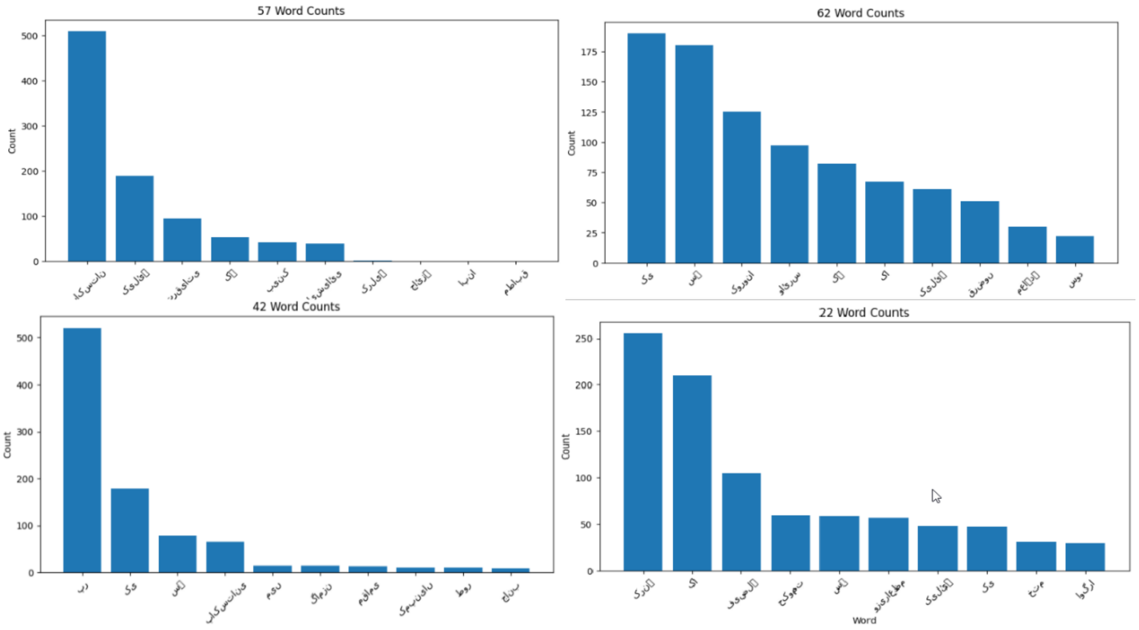

**Figure 4. 2** LDA word score

Whereas in BERTopic's word scores are based on the semantic meaning of words, which can make the resulting topics more interpretable and meaningful. LDA word scores are based on word frequency and co-occurrence, which are less intuitive for some users.

### 4.4.2 Distance map results BERT vs LDA

Both BERTopic and LDA provide distance maps as a way to visualize the similarity between topics in a topic model. In Bertopic each point represents a topic and its position is based on the embedding of its top words. Similar topics are closer together in the plot, and dissimilar topics are farther apart. The visualization can be helpful in identifying clusters of related topics and the overall structure of the topic model.

Whereas The distance map for LDA shows the hierarchical clustering of topics based on their similarity. The visualization allows for exploration of the relationships between topics at different levels of granularity, and can help identify subtopics or related themes within a larger topic.

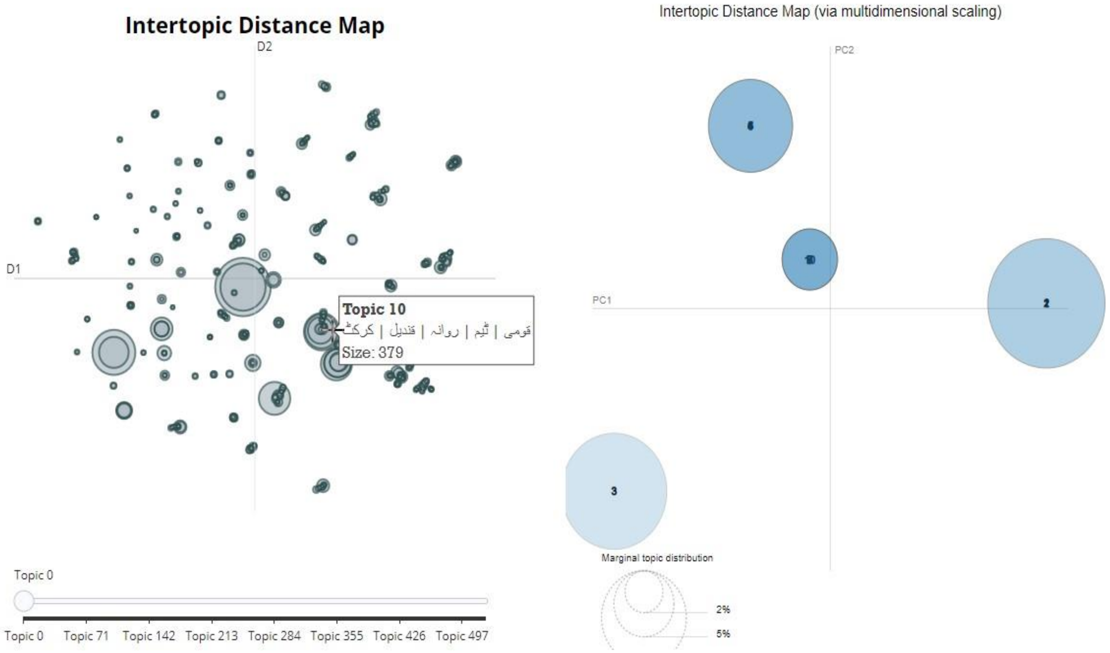

**Figure 4. 3** BERTopic and LDA Distance Map

When we clicked on any particular topic, we can see what are the important words over there, for example, topic 10 displays that this cluster belongs to the sport category furthermore the overlapping is done within the cluster which means these topics consist of the same document and that makes easier to identify which topic consist of which document. So here we have visualized, what are the important words/ topics and how these topics are disturbed. We had also seen that there is a lot of overlap between these topics which helps us to identify that these clusters are talking about the same document. In the BERT model, the topics were very specific whereas in LDA topics are generative.

However, we compared the Topic reduction and BERTopic experiment, as it is an important step in the BERTopic algorithm, we had done reduced topic because in BERT model number of topics are generated by -1 by default and basically BERTopic accept outliers so that to remove outliers we had done reduced BERT topic which number of topic start with 0. Similarity, by reducing the number of topics, we had focus on the most relevant and meaningful topics, which can improve the interpretability of the results. We had improve the clustering quality and reduce redundancy and we had also reduce the amount of computation required and made the algorithm more efficient.

### 4.4.3 Similarity matrix BERT vs LDA:

In terms of visualization, the BERTopic heatmap provides a clearer and more direct view of topic similarity scores, while the LDA circle grid provides a more abstract and visually appealing representation of the topics and their relationships.

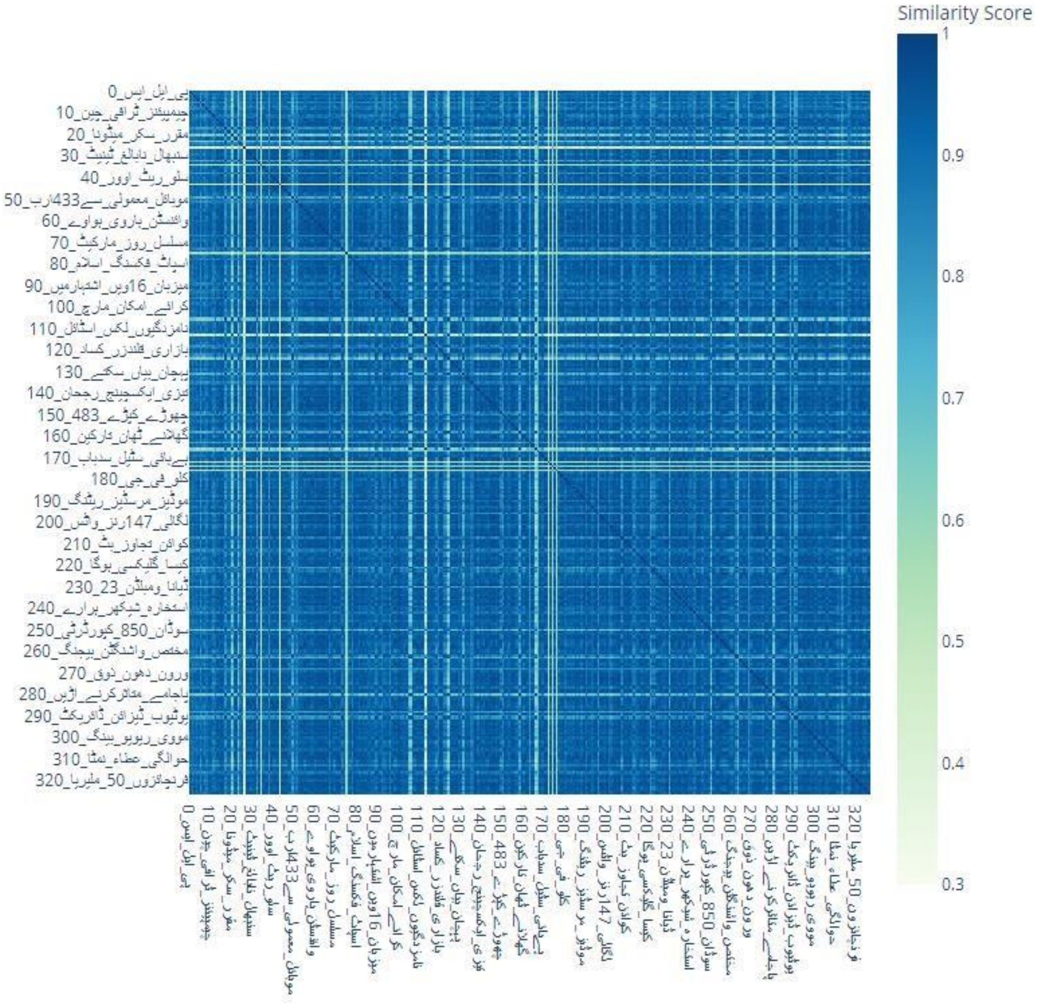

**Figure 4. 4** BERT similarity matrix

Here, the similarity matrix for BERTopic is often visualized as a heatmap, where each cell represents the similarity score between two topics. The heatmap is colored according to the similarity score, with higher similarity scores being assigned darker colors. This visualization can be useful for identifying clusters of similar topics, and can also help in identifying outlier topics.

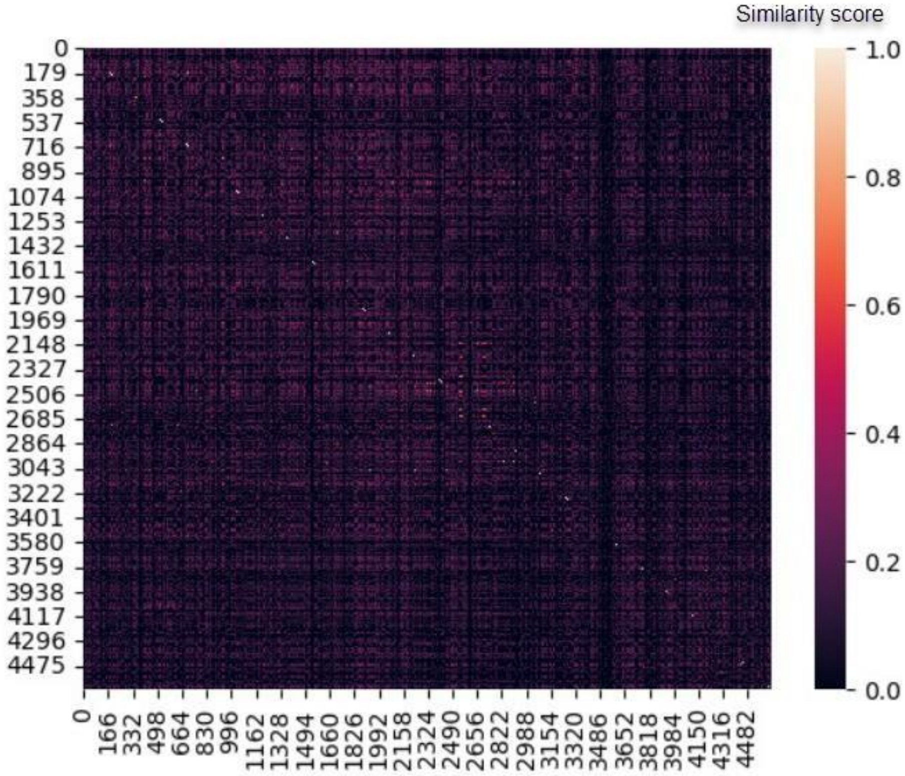

**Figure 4. 5** LDA similarity matrix

Whereas on the other side LDA, it is visualized as a grid of circles, where each circle represents a topic and its size corresponds to its prevalence in the corpus. The proximity of the circles indicates the similarity between the corresponding topics. This visualization allows for exploration of the relationships between topics at different levels of granularity, and can help identify subtopics or related themes within a larger topic.

### 4.4.4 Hierarchical clustering results BERT vs LDA

It's a kind of dendrogram visualization. BERTopic uses a graph-based clustering algorithm that can group similar topics into clusters at different levels of granularity. In the below fig, clusters are organized in a tree-like structure, with the top-level clusters representing broad themes and lower-level clusters representing more specific subtopics. BERTopic's clustering is based on the cosine similarity between topic embeddings, and the resulting tree structure can be visualized as a dendrogram; for example, in the picture below, topics 27 and 13 show similar topics whereas topic 27 and 21 cluster same topics.

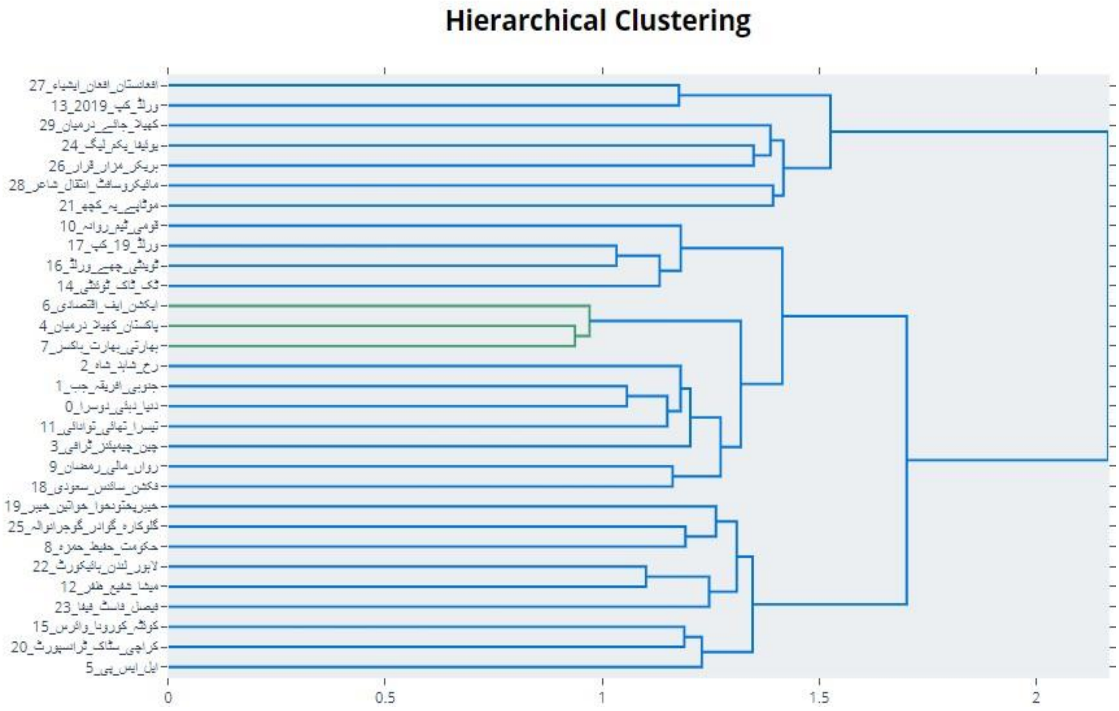

**Figure 4. 6** BERT Hierarchical clustering

Similar to this, LDA also uses hierarchical clustering to group similar topics into clusters, but it uses a different method than BERTopic. In below fig, LDA clustering is based on the topic probability distributions, and the resulting clusters can be visualized as a dendrogram. The topic distribution is not clearly clustered as BERTopic

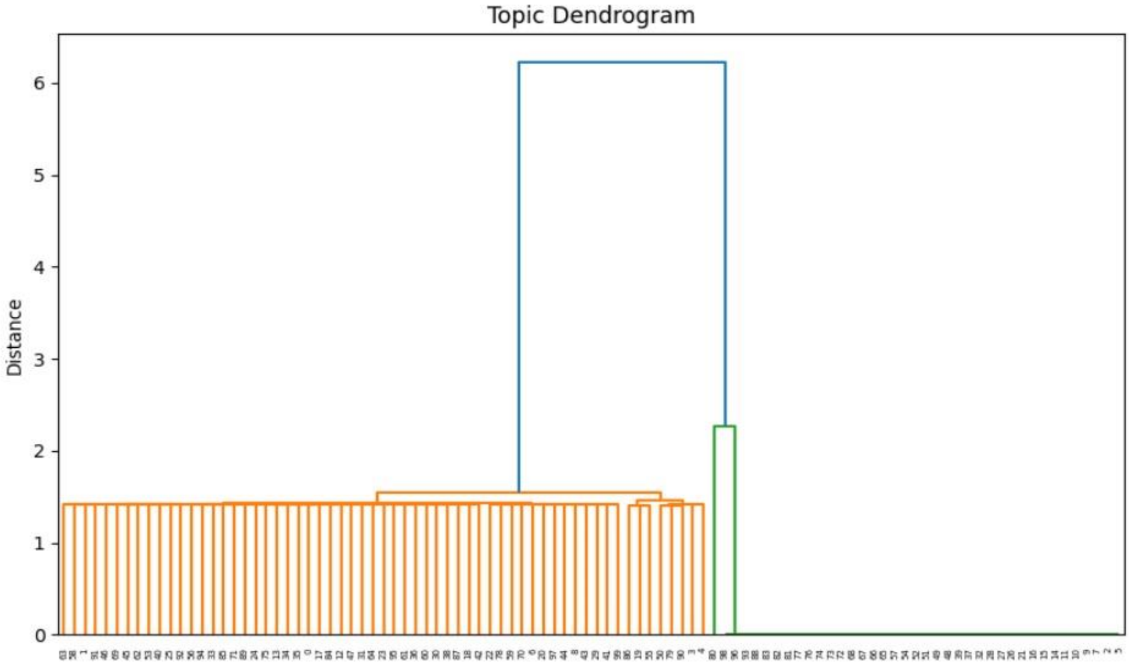

**Figure 4. 7** LDA Hierarchical clustering

However, Transformer based model, BERT is very efficient to generate topics from the document.

### 4.4.5 Word Cloud of BERT model vs LDA

Creating a word cloud was a successful strategy for presenting the model's output in a visually appealing and easily understandable way. The word cloud was created using categories of topics that were simple for people to recognize, and it helped to highlight the most important and relevant terms associated with each topic.

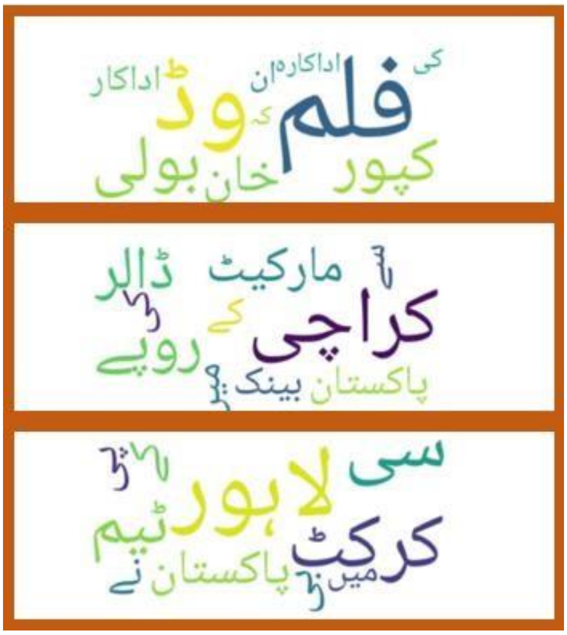

**Figure 4. 8** Main keywords of BERT Wordcloud

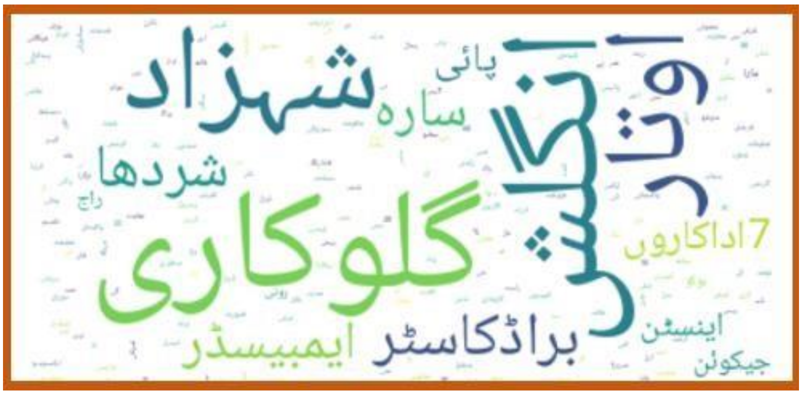

**Figure 4. 9** Main keywords of LDA Wordcloud

By presenting the results in this way, we were able to provide a clear and concise summary of the main themes and topics represented in the BERT model's output. This was an effective way to communicate the results of the study to a wider audience, including those who may not have a deep understanding of natural language processing or topic modeling techniques.

Transformer-based model, BERT is very efficient to generate topics from the document and can capture complex linguistic patterns but may require a large amount of training data and computational resources. LDA, on the other hand, can be useful for identifying topics in a corpus of text, but it may not capture more subtle semantic relationships between words.

## 4.5 Human analysis:

In order to draw a meaningful conclusion from human analysis. However, [44] research demonstrates that many NLP experiments, particularly those that rely on human evaluation, lack the necessary power to identify model differences at reported levels. Human evaluation plays an important role because in the real-world human perspective matters more rather than automated models’ results. We Automated evaluation measures and human annotation to assess the quality of the models' output.

One of the human raters who was a data analyst by profession conducted the results for us. The human rater they employed found that the BERT model's topic representation was more specific and easier to comprehend than the LDA model's, which had more ambiguous topics. Also show them different visualizations, such as intertopic distance maps and word clouds, to better understand the output of the models. They found that the BERT model's topics exhibited clustering and similarity, indicating that the model was performing well, while the LDA model's topics were less clustered and less related to each other.

According to their interpretation of the Intertopic distance map, the BERT topic exhibits a clustered image, it overlaps the topics, and this indicates that there is a similarity between topics, indicating that BERT was a good sign to consider further. In contrast, in the LDA model, the overlap of the topic is not scattered, indicating that the majority of the topics are not clustering in a similar corpus. In terms of the similarity matrix, they identify that the x and y axes show a similarity and correlation between topics. In LDA intertopic distance is low, and the topics are not cluttered, indicating a distinction between them as the majority of the topics are unrelated.

However, we also created a word cloud for the Bertopic model using categories of themes that are simple for people to recognize. They found this strategy to be quite successful because the topics are generated in a cloud format and are simple and easily to understand.

After the automated results and human judgment, we concluded that the transformer-based model BERT model performed better than the LDA model in terms of accuracy and understandability.

# CHAPTER 5

# CONCLUSION

In this Chapter, we describe the conclusion of our research.

## 5.1 Conclusion

Even though Urdu is a widely spoken language, the amount of computer linguistic research available is quite small. Since language is important to millions of people around the world, developing its computation linguistics assets from infancy to maturity is of vital importance. Although some topic modeling techniques have been used, such as the popular LDA, there is an identified research gap concerning the use of transformers in topic modeling Urdu texts. Statistical analysis methods have been predominantly used for Urdu NLP. This was due to the scarcity of resources for Urdu. In recent years, Transformer-based models have become state-of-the-art for many NLP tasks including Topic Modeling. As they are pre-trained model are helpful as they accurately represent the words and sentences from large corpus. Transformer-based models, such as BERTopic, are typically more computationally intensive than LDA and require more powerful hardware, such as GPUs, to train efficiently. However, once trained, they can be very fast and scalable in inference, particularly when used for document classification or clustering.

In this research, we intended to perform a comparative analysis of Bertopic and LDA methods to analyze their performance. As we come up with the results that Bertopic has maintained a good semantic connection among the topic words and they have higher coherence score which mean they were able to accurately cluster topic-related words based on their semantic similarity. On the other hand, LDA topics were very generative and they were not interpretable by human. In order to when we did human analysis, the rate observed that results of BERTopic are effective and understandable as compare to LDA. After the automated results and human judgment, we concluded that the transformer-based model BERT model performed better than the LDA model in terms of accuracy and understandability.

## 5.2 Future Work

Additionally, Urdu language, like many other languages, has its own unique linguistic features and challenges, which present opportunities for future work on Transformer-based models. Some of the potential direction for the future work are:

Pre-training: While pre-trained Transformer-based models have achieved remarkable success in NLP tasks, they are often trained on large corpora of English language data. The Future work could focus on developing pre-training models specifically for Urdu language, using large corpora of Urdu text data

Corpus development: Currently, there is a lack of large and diverse Urdu language corpora for training and evaluating NLP models. Future work could focus on developing and curating such corpora, covering different domains, genres, and styles of Urdu text.

Part-of-speech tagging: Accurate part-of-speech (POS) tagging is a crucial component of many NLP tasks, such as text classification and named entity recognition. Future work could focus on developing more accurate and efficient POS tagging models for Urdu language, using Transformer-based models.

Named entity recognition: Named entity recognition (NER) is an important NLP task that involves identifying and extracting entities such as person names, organizations, and locations from text. Future work could focus on developing more effective and robust NER models for Urdu language, using Transformer-based models.

Machine translation: Machine translation is an important application of NLP that involves automatically translating text from one language to another. Future work could focus on developing more accurate and efficient machine translation models for Urdu language, using Transformer-based models, to support applications such as cross-lingual information retrieval and communication.

In future we can also compare transformer-based model with other topic modeling technique to get better comparison and effective results.

# REFERENCES

1. Syed Zain Abbas, Dr Rahman, Abdul Basit Mughal, Syed Mujtaba Haider, et al. Urdu news article recommendation model using natural language processing techniques. arXiv preprint arXiv:2206.11862, 2022.

2. Saqib Aziz, Michael Dowling, Helmi Hammami, and Anke Piepenbrink. Machine learning in finance: A topic modeling approach. European Finan- cial Management, 28(3):744–770, 2022.

3. Md Kowsher, M Ashraful Alam, Md Jashim Uddin, Faisal Ahmed, Md Wali Ullah, and Md Rafiqul Islam. Detecting third umpire decisions & automated scoring system of cricket. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), pages 1–8. IEEE, 2019.

4. Md Kowsher, Anik Tahabilder, Md Zahidul Islam Sanjid, Nusrat Jahan Prottasha, Md Shihab Uddin, Md Arman Hossain, and Md Abdul Kader Jilani. Lstm-ann & bilstm-ann: Hybrid deep learning models for enhanced classification accuracy. Procedia Computer Science, 193:131–140, 2021.

5. Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. Parsbert: Transformer-based model for persian language understanding. Neural Processing Letters, pages 1–17, 2021.

6. Mihai Masala, Stefan Ruseti, and Mihai Dascalu. Robert–a romanian bert model. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6626– 6637, 2020 .

7. Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226, 2018.

8. Md Kowsher, Md Shohanur Islam Sobuj, Md Fahim Shahriar, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba.

An enhanced neural word embedding model for transfer learning. Applied Sciences, 12(6):2848, 2022.

9. Md Jamiur Rahman Rifat, Sheikh Abujar, Sheak Rashed Haider Noori, and Syed Akhter Hossain. Bengali named entity recognition: A survey with deep learning benchmark. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pages 1–5. IEEE, 2019.
10. Salim Sazzed. Cross-lingual sentiment classification in low-resource bengali language. In Proceedings of the Sixth Workshop on Noisy Usergenerated Text (W-NUT 2020), pages 50– 60, 2020.
11. Roman Egger. Topic modelling. In Applied Data Science in Tourism, pages 375–403. Springer, 2022.
12. Nazmiye Eligu¨zel, Cihan C¸ etinkaya, and Tu¨rkay Dereli. Comparative anal- ysis with topic modeling and word embedding methods after the aegean sea earthquake on twitter. Evolving Systems, pages 1–17, 2022.
13. Maarten Grootendorst. Bertopic: Neural topic modeling with a classbased tf-idf procedure, 2022.
14. Mi Kim and Dosung Kim. A suggestion on the lda-based topic modeling technique based on elasticsearch for indexing academic research results. Applied Sciences, 12(6):3118, 2022.
15. Adela Ljaji´c, Nikola Prodanovi´c, Darija Medvecki, Bojana Baˇsaragin, and Jelena Mitrovi´c. Topic modeling technique on covid19 tweets in serbian.
16. Mubashar Mustafa, Feng Zeng, Hussain Ghulam, and Wenjia Li. Discov- ering coherent topics from urdu text. 08 2021.
17. Mubashar Mustafa, Feng Zeng, Hussain Ghulam, and Hafiz Muham- mad Arslan. Urdu documents clustering with unsupervised and semi- supervised probabilistic topic modeling. Information, 11(11):518, 2020.
18. Zarmeen Nasim. On building an interpretable topic modeling approach for the urdu language. In IJCAI, 2020.

19. Zarmeen Nasim and Sajjad Haider. Automatic labeling of clusters for a low-resource urdu language. Transactions on Asian and Low-Resource Lan- guage Information Processing, 21(5):1–22, 2022.

20. Radim Rˇeh˚uˇrek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta, May 2010. ELRA.

21. Khadija Shakeel, Ghulam Rasool Tahir, Irsha Tehseen, and Mubashir Ali. A framework of urdu topic modeling using latent dirichlet allocation (lda). pages 117–123, 01 2018.

22. Venkatesh Shankar and Sohil Parsana. An overview and empirical compar- ison of natural language processing (nlp) models and an introduction to and empirical application of autoencoder models in marketing. Journal of the Academy of Marketing Science, pages 1–27, 2022.

23. Yuan PY, Du AM, Wang C (2020) Using Word2vec to match knowledge points and test questions: a case study. In: Proc. 2nd Int. Conf. Comput. Sci. Educ. Informatiz. CSEI 2020, pp 272–276.

24. Yazhou Xie, Chunxiao Ning, and Lijun Sun. The twenty-first century of structural engineering research: A topic modeling approach. In Structures, volume 35, pages 577–590. Elsevier, 2022.

25. Zoya, Seemab Latif, Faisal Shafait, and Rabia Latif. Analyzing lda and nmf topic models for urdu tweets via automatic labeling. IEEE Access, PP:1–1, 09 2021.

26. Nusrat Jahan Prottasha, Abdullah As Sami, Md Kowsher, Saydul Akbar Murad, Anupam Kumar Bairagi, Mehedi Masud, and Mohammed Baz. Transfer learning for sentiment analysis using bert based supervised finetuning. Sensors, 22(11):4157, 2022

27. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

28. Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: Getting inside bert's linguistic knowledge. arXiv preprint arXiv:1906.01698, 2019.

29. Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582, 2019 .

30. Deerwester S, Dumais ST, Furnas GW, Landauer TK (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

31. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

32. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

33. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

34. Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

35. Jeremy Howard and Sebastian Ruder. Universal language model finetuning for text classification. arXiv preprint arXiv:1801.06146, 2018.

36. Jose Manuel Gomez-Perez, Ronald Denaux, and Andres Garcia-Silva. Understanding word embeddings and language models. In A Practical Guide to Hybrid Natural Language Processing, pages 17–31. Springer, 2020.

37. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

38. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

39. Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

40. Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising crosslingual effectiveness of bert. arXiv preprint arXiv:1904.09077, 2019. [23] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502, 2019.
41. Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291, 2019.
42. Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. Is multilingual bert fluent in language generation? arXiv preprint arXiv:1910.03806, 2019.
43. Jelodar H et al (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211.
44. Mikolov T, Corrado G, Chen K, Dean J (2013) Efficient estimation of word representations in vector space. arXiv, pp 1–12