Adapting Outperformer from Topic Modeling Methods for Topic Extraction and Analysis: The Case of Afaan Oromo, Amharic, and Tigrigna Facebook Text Comments

Naol Bakala Defersha[1], Kula Kekeba Tune[2], Solomon Teferra Abate[3]
Ph.D. student, Core Member of Center of Excellence for HPC and Big Data Analytics, Software Engineering,
College of Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia[1]
Software Engineering, College of Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia[2]
Information Science, College of Natural and Computational Science, Addis Ababa University, Addis Ababa, Ethiopia[3]

Abstract*—Facebook users generate a vast amount of data, including posts, comments, and replies, in various formats such as short text, long text, structured, unstructured, and semi structured. Consequently, obtaining import information from social media data becomes a significant challenge for low-resource languages such as Afaan Oromo, Amharic, and Tigrigna. Topic modeling algorithms are designed to identify and categorize topics within a set of documents based on their semantic similarity which helps obtain insight from documents. This study proposes latent Dirichlet allocation, matrix factorization, probabilistic latent semantic analysis, and BERTopic to extract topics from Facebook text comments in Afaan Oromo, Amharic, and Tigrigna. The study utilized text comments from the Facebook pages of various individuals, including activists, politicians, athletes, media companies, and government offices. BERTopic was found to be the most effective for identifying major topics and providing valuable insights into user conversations and social media trends with coherence scores of 82.74%, 87.85%, and 81.79% respectively.

Keywords*—*Afaan oromo; amharic; tigrigna; BERTopic; topic extraction; social media data

I. Introduction

Social media platforms now offer various features and tools for online resource sharing, rapid conversation, and improved communication through digital technologies such as comments such as shares, and replies. The vast amount of data generated on these platforms is valuable for comprehending user behavior, predicting trends, and analyzing sentiments [1]. Analyzing social media data is vital for gaining insight into user discussions and identifying emerging trends [2].

Topic modeling is used in natural language processing and text mining to automatically identify and extract meaningful topics from textual data and to provide insights into user discussions. Common topic modeling techniques include Latent Dirichlet Allocation (LDA), non-negative matrix factorization (NMF), Probabilistic Latent Semantic Analysis (PLSA), Latent Semantic Analysis (LSA), and BER Topic. Topic modeling algorithms are categorized into traditional and recent topic modeling algorithms. Topic modeling algorithms like LDA, NMF, PLSA, and LSA are used for social media data analysis, but traditional methods may be less effective due to deployment costs and limited metadata sources [2].

Despite their popularity, these techniques have limitations, requiring multiple topics, stop-word lists, stemming, and lemmatization. They use a bag-of-words representation that disregards word order and meaning [3].

Recent studies have explored a topic modeling approach for text classification, using sentence embedding, semantic similarity grouping, and c-TF-IDF to determine topic distribution among texts [4].

BERTopic is a recent method that uses UMAP and HDBSCAN to generate a comprehensive list of document topics for lexicon, text classification, information retrieval, and abuse detection [5].

The study analyzed over 240,000 Spanish immigrants' hate speech tweets using unsupervised machine learning and latent Dirichlet allocation techniques between November 2018 and April 2019 [6].

A study analyzed a dataset of 1,424 NATO-supporting videos, revealing 8,276 comments, while 3,461 NATO-opposing videos had 46,464 comments [7]. The researchers utilized LDA topic modeling to extract semantic information from documents and analyzed the impact of toxicity levels on narratives, including pro- or anti-NATO videos and their linked comments [7].

Obadimu et al. [8] used structural topic modeling to analyze racial prejudice in social media posts, identify hate speech, and determine abusive language frequency using Gibbs sampling and human assessment.

The study used an unsupervised topic model to cluster Afaan Oromo documents, learning a combination of latent topics in a probability distribution representation over vocabularies [9]. Researchers utilized word embedding techniques, semantic correlation with LDA, and Gibbs sampling to improve topic quality. Evaluations include perplexity, topic coherence, and human assessment [9].

The authors conducted a study on STTM algorithms for short-text topic modeling using Real-World Pandemic Twitter and Real-World Cyberbullying Twitter datasets, to evaluate topic coherence, purity, NMI, and accuracy [2].

Topic modeling is a method that evaluates social media texts using tools such as latent semantic analysis, latent Dirichlet allocation, non-negative matrix factorization, random projection, and principal component analysis [10]. The authors use a Comparative Analysis (LDA) method to identify hate speech types on social media, focusing on religion, race, disability, and sexual orientation, requiring text cleaning [11]. BERTopic is a recent method for extracting topics from documents by clustering lower-dimension approximations, thereby reducing the computational difficulty in determining related word embedding closeness [12]. BERTopic offers UMAP to reduce dimensionality by eliminating noisy data, while HDBSCAN clusters samples and minimizes outliers. The study utilized LDA for topic modeling in web content classification, concluding that sentiment analysis and topic modeling are crucial for effective classification [13]. T. Davidson and D. Bhattacharya [14] used a structural topic modeling method to examine racial bias within an online abuse dataset.

The second types of topic modeling algorithms are BERTopic and To2Vec used to represent topics from the document. BERTopic and Top2Vec frameworks enhance topic representation accuracy, prompting further research on social media text data topics, a class-based variant of TF-IDF. Angelov [15], employed joint document and word semantic embedding to identify topic vectors without relying on stop-word lists, stemming, or lemmatization. The experiment reveals that top2vec outperforms probabilistic generative models in identifying more informative and representative themes in the trained corpus [15].

Topic modeling, an unsupervised technique using UMAP and HDBSCAN, is used to create a comprehensive document topic collection and interpretable c-TF-IDF for social media text data topics [10].

Silva et al. [16] employ BERTopic topic models for Portuguese political comments on Brazil's Chamber of Deputies bills, adjusting parameters to improve accuracy and align with research direction.

Topic modeling techniques for Afaan Oromo, Amharic, and Tigrigna, face unique challenges due to their rich morphology, complex syntax, lack of pre-trained models, and informal expressions, necessitating adaptation of existing algorithms.

Previous studies primarily focus on resource-rich languages such as Arabic, English, Portuguese, and Spain, which have their own compiled list of stop words, spelling error checkers, and word disambiguating tools. Limited studies have explored the recent BERTopic and Top2vec topic representation algorithms to extract topics from social media data that involve short text, and nonstandard language.

Research Objectives: the aim of this study is to 1) apply LDA, LSA, NMF, PLSA, and BERTopic models to extract topics from Afaan Oromo, Amharic, and Tigrigna, particularly to a) compare the performance of proposed topic modeling methods, b) adapt the outperformer from LDA, LSA, NMF, PLSA, and BERTopic for extracting topics from Afaan Oromo, Amharic, and Tigrigna Facebook Text comments, 2) investigate organic discussion in various languages on a Facebook platform such as cross-cultural communication, language analysis, hateful content analysis, and social Trends/Topics.

This study contributes to 1) the development of a large-scale multilingual dataset for Afaan Oromo, Amharic, and Tigrigna from Facebook pages, 2) the construction of language embedding for document transformation, 3) comparing the performance of LDA, LSA, NMF, and PLSA and BERTopic's effectiveness in extracting quality topics, 4) adapt BERTopic and evaluates its hyperparameter tuning, and 5) Investigation of organic discussion across languages in Facebook.

The findings of this study will enable researchers and practitioners to effectively analyze and understand user discussions, trends, and sentiments in Afaan Oromo, Amharic, and Tigrigna social media platforms, facilitating a deeper understanding of user behavior and the development of targeted strategies and interventions.

II. Related Work

To extract topics from Afaan Oromo for getting insight information text comments, a few studies were attempted on Afaan Oromo and Amharic whereas no study on the Tigrigna text document.

The study uses an unsupervised topic model for document clustering in Afaan Oromo documents, learning latent topics in probability distribution representation over vocabulary [9]. Word embedding and LDA algorithm improve theme quality. Performance is validated through Perplexity, Topic Coherence, and human assessment [9].

The study evaluated the effectiveness of pre-trained word embedding techniques, deep learning algorithms, and BERTopic in extracting topics and classifying hateful speech in Afaan Oromo Facebook comments [17].

There are also topic modeling approaches applied for Amharic [18] and [19]. The paper proposes a concept-based single-document Amharic text summarization system using topic modeling, specifically probabilistic latent semantic analysis (PLSA)[19]. The algorithms are language and domain-independent, allowing for use in other local languages [19]. The authors propose six algorithms, each with two common steps: selecting keywords and selecting sentences with the best keywords[19]. They experimented with news articles and found encouraging results after varying extraction rates [19].

This study develops a supervised topic model using LDA for an Amharic corpus, examining the impact of stemming on topic detection using four supervised machine learning tools: Support Vector Machine (SVM), Naive Bayesian (NB), Logistic Regression (LR), and Neural Nets (NN) [18]. The approach outperforms state-of-the-art TF-IDF word features with an 88% accuracy rate, suggesting that stemming slightly improves the topic classifier's performance [18]. On the other hand, no study applied topic modeling for the Tigrigna text document.

III. Methodology

This paper proposes a topic modeling approach for extracting topics from Afaan Oromo, Amharic, and Tigrigna text comments using LDA, LSA, PLSA, NMF, and BERTopic, and evaluates it using topic coherence score and then adapts the outperformer algorithm. Fig. 1 shows the details of the proposed methodology. In this study, we applied the following methodology to achieve the proposed objectives.

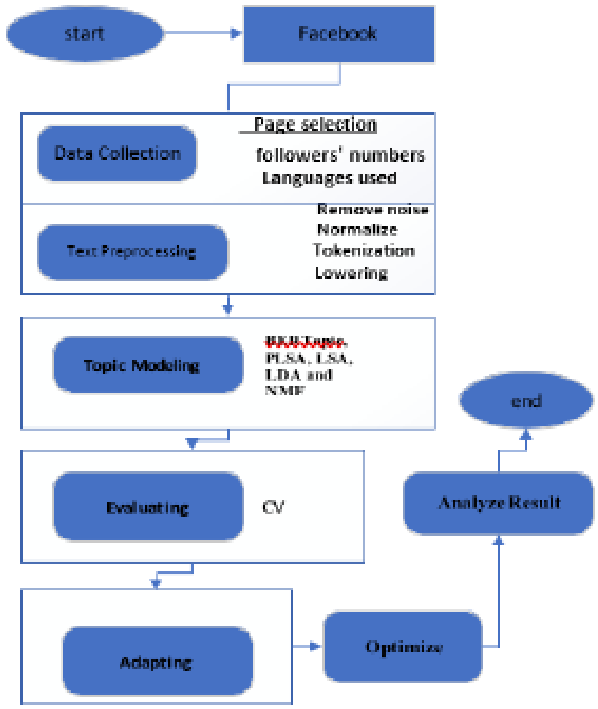

Fig. 1. Topic modeling methods topics extraction proposed framework.

A. Data Collection

Because Facebook is the most dominantly used social media network in Ethiopia, it was used as a source of data to collect data from Afaan Oromo, Amharic, and Tigrigna. In addition to collecting new data, we used an existing dataset to construct a large dataset. A large Afaan Oromo social media dataset was prepared by collecting data from various Facebook pages, including those of activists, politicians, athletes, media companies, and government offices[17]. We selected Fakebook pages with at least 21000 followers of Afaan Oromo. We recruited five master's students from computer science and information technology, two from information technology, and three from computer science, to gather data. The experts who collected Afaan Oromo's text comments from September 2019 to October 2022 completed the data collection within two months. Amharic dataset available at https://data.mendeley.com/datasets/fhvsvsbvtg/3 and also available at https://data.mendeley.com/datasets/ymtmxx385m/1.

The 10827 and 35000text comments were gathered from specified those sources respectively. We also gathered 13882 Tigrigna text comments from Facebook pages of media companies, politicians, activists, websites, public services, interests, people, and political parties with over 15,000 followers. From June 2023 to August 2023, text comments from 2018 to 2023 were gathered by experts from five master's students of computer science and information technology. The Amharic hate speech detection dataset, which includes 35000 text comments, was translated into Tigrigna using Google Translator. Finally, to compile 48882 Tigrigna text comments.

B. Text Preprocessing

This study focuses on text preprocessing and removing unnecessary content such as HTML links, tags, numbers, punctuation, and stop words to create clean comments. It also tokenizes plain text, eliminates non-Afaan Oromo, non-Amharic, and non-Tigrigna words, and converts all comments to lowercase.

C. Applying LDA, LSA, PLSA, NMF, and BERTopic to Develop the Proposed Model

The proposed model was developed using LDA, LSA, PLSA, NMF, and BERTopic to extract quality topics, and its performance was compared. LDA automatically extracts topics from documents based on the relevance of connected topics within texts and documents [3]. Studies have primarily focused on long-text topic modeling approaches, including classic methods, such as latent Dirichlet allocation[3], LSA[9], PLSA[19], and NMF[20] which are commonly used to extract latent semantic structures in long texts. Short-text generation is becoming more prevalent, but long-text TMs are less promising owing to their limited content and difficulty in finding topic co- occurrence [20]. Despite their popularity, these techniques have drawbacks, such as requiring numerous topics, stop-word lists, stemming, lemmatization, and relying on a bag-of-words representation that disregards word order. The study utilized coherence measurement CV to evaluate the efficacy of topic modeling [15]. In 2019, Angelov tests demonstrated that top2vec outperformed probabilistic generative models in identifying more informative and representative themes in a trained corpus [2]. Researchers have employed joint document and word semantic embedding to identify topic vectors without the need for stop-word lists, stemming, or lemmatization [15]. BERTopic is a recent method that uses lower-dimension approximations to extract topics from documents, thereby reducing the computational complexity in determining related word embedding closeness[12]. Unlike traditional topic modeling methods, BERTopic eliminates human judgment or intervention in developing suitable models, focusing solely on selecting parameters for model training.

D. Comparing the Performance of LDA, LSA, PLSA, NMF, and BERTopic Models

The study utilized LDA, LSA, NMF, PLSA, and BERTopic to extract topics from a dataset with topic coherence scores to assess their performance.

E. Topic Extraction

The proposed topic extraction techniques, including LDA, LSA, NMF, PLSA, and BERtopic, were applied to develop

Topic Modeling Topic extraction and Hate Speech Analysis (TMBTEHSA), evaluating topic quality.

F. Evaluation

Evaluating the effectiveness of LDA, LSA, PLSA, NMF, and BERTopic in extracting topics from Afaan Oromo, Amharic, and Tigrigna social media data is crucial for assessing semantic coherence and topic diversity. Topic coherence is a linguistic concept that automatically evaluates the interpretability of latent topics based on the distributional theory, which suggests similar meanings often appear in similar contexts [16]. The study confirms that the coherence measurement CV is a reliable tool for evaluating the effectiveness of topic modeling, showing a positive correlation with human interpretability [21]. In this study, we utilized coherence measurement CV to assess the efficacy of topic modeling, as detailed in Table II.

G. Optimization

Hyperparameter tuning is a technique used to optimize the parameters and configurations of an outperformer model based on the evaluation results to enhance the quality of the extracted topics. During this phase, the hyperparameters of the selected techniques were adjusted and tuned to improve the quality of the extracted topics (details of BERTopic optimization and adaption techniques are indicated in Fig. 2).

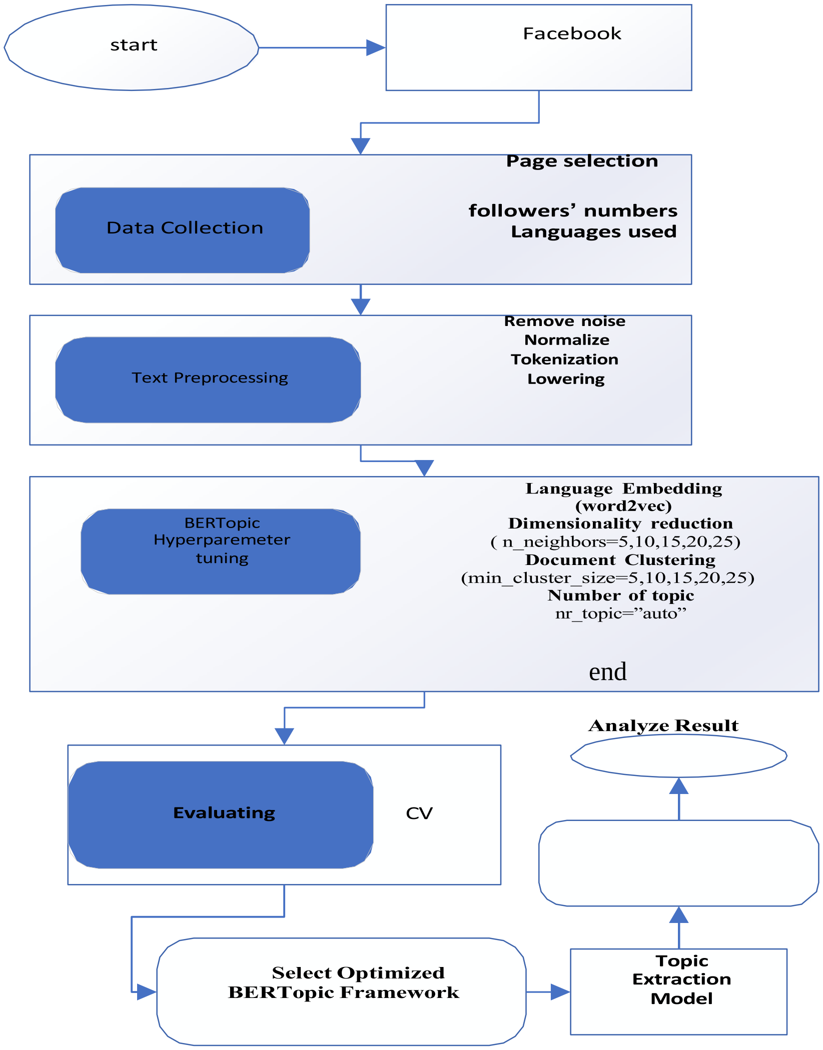

Fig. 2. Proposed framework for BERTopic parameters tuning.

H. Implementation Environment and Programming Language

Python was used for the experiment, with Google Collaboratory scripts written using GPUs for faster data analysis and Excel for dataset preparation and CSV file saving.

IV. Experiment

Two experiments were conducted to achieve the proposed objectives. First, the performance of the topic modeling algorithms was evaluated based on text comments prepared by Afaan Oromo, Amharic, and Tigrigna. Second, outperforming algorithms were adapted and optimized to develop the proposed model.

A. Dataset

As indicated in the previous section, we used Facebook as a source of data to prepare a dataset consisting of Afaan Oromo, Amharic, and Tigrigna, with sizes of 59529, 45522, and 48882 text comments, respectively.

B. Experiment One

The study compared topic modeling algorithms such as LDA, PLSA, LSA, NMF, and BERTopic in extracting topics from Afaan Oromo, Amharic, and Tigrigna social media data, revealing that BERTopic performed better in terms of topic quality, computational efficiency, and ease of implementation (details of the effectiveness of the algorithm in Table I). The framework of experiment two is indicated in Fig. 1.

C. Experiment Two

As indicated in Table VI, the BERTopic scored a higher coherence score than LDA, PLSA, LSA, and NMF. Therefore, BERTopic was selected and optimized in experiment two to develop the topic extraction model from Afaan Oromo, Amharic, and Tigrigna social media data. In this study, we adapted the BERTopic algorithm to accommodate the characteristics of Afaan Oromo, Amharic, and Tigrigna's social media text. In the BERtopic adaptation phase, the framework of the BERTopic was modified based on the input Afaan Oromo, Amharic, and Tigrigna to handle the input text. Accordingly, the BERTopic framework was modified based on parameters from language embedding, dimensionality reduction, clustering document, and the size of topics (the detail approach described in Fig. 2).

D. BERTopic Hyperparameters Tuning Models

Fine-tuning helps the models learn the patterns and semantics specific to Afaan Oromo, Amharic, and Tigrigna's social media text (see Table I). NLP faces the challenge of organizing and summarizing large text corpora, often utilizing topic modeling when intelligent reading and sorting are impossible.

Topic modeling is a popular research area in natural language processing that aims to extract topics from documents and words with minimal computer resources. When a person cannot read and sort through an enormous text corpus, topic modeling is employed [12]. P. Ghasiya and K. Okamura [12] used topic modeling to address the challenges of processing and understanding short text documents. BERTopic is a recent method for extracting topics from documents by clustering lower-dimension approximations, reducing the computational difficulty in determining related word embedding closeness [12].

Unlike traditional topic modeling approaches, BERTopic will not need to be upfront with current topics to develop a topic extraction model. In this sense, human judgment or intervention is absent from BERTopic, except in selecting the parameters for model training. The most popular embeddings in BERTopic are Sentence Transformers, Hugging Face Transformers, Flair, Spacy, Universal Sentence Encoder, Gensim, Scikit-Learn Embeddings, OpenAI, TF-IDF, Custom Embeddings, Custom Backend, and Multimodal.

In the first steps, since sentence transformers are pretty good at capturing the semantic similarity of documents, BERTopic begins by converting our input documents into numerical representations. In the second step, dimensionality reduction of the input embeddings is a critical component of BERTopic.

TABLE I. BERTOPIC HYPERPARAMETER TUNING SETUP

Lang Emb	Dim Red	Cluste ring	No of Topics	lang
word2vec	5	5	auto	Afaan Oromo
word2vec	10	10	auto	Afaan Oromo
word2vec	15	15	auto	Afaan Oromo
word2vec	20	20	auto	Afaan Oromo
word2vec	25	25	auto	Afaan Oromo
word2vec	5	5	auto	Amharic
word2vec	10	10	auto	Amharic
word2vec	15	15	auto	Amharic
word2vec	20	20	auto	Amharic
word2vec	25	25	auto	Amharic
word2vec	5	5	auto	Tigrigna
word2vec	10	10	auto	Tigrigna
word2vec	15	15	auto	Tigrigna
word2vec	20	20	auto	Tigrigna
word2vec	25	25	auto	Tigrigna

The calamity of dimensionality makes clustering challenging since embeddings are frequently highly dimensional. Because UMAP is the default value in BERTopic that can capture the local and global high-dimensional space in lower dimensions, researchers may find it worthwhile to experiment with alternative solutions, such as PCA. We can apply any other dimensionality reduction approach such reduction because BERTopic requires some degree of independence between phases.

The accuracy of our topic representations increases with the performance of our clustering technique, which makes the clustering process crucial. HDBSCAN is one component of BERTopic that can capture structures with varying densities. In addition, HDBSCAN and BERTopic use cuML HDBSCAN, agglomerative clustering, and k-means as examples of clustering mechanisms.

To accurately depict the topics from our bag-of- words matrix, TF-IDF was modified in BERTopic to work at the cluster, topic, and topic levels rather than the document level. "c-TF-IDF" refers to this modified TF-IDF representation, which accounts for variations across documents inside a cluster. BERTopic allows for directly adjusting several hyperparameters to improve the model's performance such as PCA, Truncated SVD, call MAP, and skip dimensionality.

1) Language embedding: BERTopic uses the sentence transformer's English version for document embeddings but requires an additional sentence-transformer model for low-resource languages. In this study, we build a word2vec pre-trained embedding model for Afaan Oromo, Amharic, and Tigrigna to transform text documents for dimensionality reduction in the BERTopic framework.

2) Dimensionality reduction: The UMAP algorithm is a clustering model that reduces dimensionality while maintaining local and global data structure. It can be customized with hyperparameters such as n_neighbors, and its default value is 15, resulting in larger cluster sizes. The BERTopic model uses UMAP's stochasticity for distinct outcomes.In this study, we focused on the n_neighbors parameters and tuned them to 25,15,10 and 5, as indicated in Table II, and then its performance was evaluated.

3) Clustering: The study uses HDBSCAN, a density-based clustering algorithm, to reduce dimensionality in embedded documents. It automatically determines cluster size using hyperparameters such as min_cluster_size, min_samples, metric, and prediction_data. To avoid new document prediction, set it to False, feed the model into the BERTopic technique, and include umap_model for comparability. In this study, we tuned the number min_cluster_size to 25,15,10 and 5, as indicated in Table II.

4) Number of topics: BERTopic uses the HDBSCAN model's clusters as topic count but can be adjusted by changing the nr_topics parameter. The default value for nr_topics parameters is 15, and the topic reduction procedure reduces related topics based on the c-TF-IDF feature vector, starting with low- frequency topics. As indicated in Table II, in this study, the BERTopic hyperparameter tuning nr_topics to auto throughout all experiments.

E. BERTopic *Adaptation and Optimization*

Adapt the BERTopic algorithm to accommodate the characteristics of Afaan Oromo, Amharic, and Tigrigna social media text. In the BERtopic adaptation phase, the framework of the BERTopic was modified based on the input Afaan Oromo, Amharic, and Tigrigna to handle the input text. Accordingly, the BERTopic framework was modified based on parameters from language embedding, dimensionality reduction, clustering documents, and several topics extracted. After experimenting with evaluating the adapted BERTopic algorithm for topic extraction in Afaan Oromo, Amharic, and Tigrigna social media data, the hereunder results and discussions presented: - In this paper, the coherence score used to evaluate the performance of the adapted BERTopic. The coherence score calculates coherence scores for the extracted topics to assess their semantic coherence. Higher coherence scores indicate more coherent and meaningful topics. The coherence score for the adapted BERTopic is shown in Table II. As described in the Table III, the coherence score of the adapted BERTopic for Afaan Oromo is 82.74%. In contrast, the coherence score for Amharic is 87.85%. Similarly, the adjusted BERTopic coherence score for Tigrigna is 81.79%.

V. Result and Discussion

This section describes the results of the topic extracted from Afaan Oromo, Amharic, and Tigrigna Facebook text comments. As we observed from experiment one above, the BERTopic scored highest accuracy than others applied topic modeling methods such LDA, LSA, PLSA, and NMF.

A. *Afaan Oromo Social Media Data Topics and Description*

BERTopic was applied to Afaan Oromo text comments, revealing 1562 topics, with 14409 and 351 representing ethnically based hate and media hate, respectively. The discussion covered various topics, such as identity-based attacks, ethnic group-based attacks, and information that undermines others' ideas. Table IV revealed that both normal and hateful information is also delivered online, as indicated by topics extracted from text documents.

TABLE II. The Comparison of LDA, LSA, PLSA, NMF, and Bertopic in Developing Topic Extraction

Methods	Performance per Languages		
	Afaan Oromo	Amharic	Tigrigna
BERTopic	82.74	87.85	81.79
LDA	-16.83	-14.77	-14.52
PLSA	41.48	43.52	41.24
LSA	58.23	59.49	58.38
NMF	48.71	32.89	49.78
BERtopic	73.33	77.00	69.54

TABLE III. BERTopic Hyperparameter Tuning Setup

Lang Emb	Dim Red	Clustering	No of Topics	Acc	lang
word2vec	5	5	auto	82.74	Afaan Oromo
word2vec	10	10	auto	76.22	Afaan Oromo
word2vec	15	15	auto	73.33	Afaan Oromo
word2vec	20	20	auto	72.3	Afaan Oromo
word2vec	25	25	auto	70.02	Afaan Oromo
word2vec	5	5	auto	87.85	Amharic
word2vec	10	10	auto	82.81	Amharic
word2vec	15	15	auto	77.00	Amharic
word2vec	20	20	auto	74.60	Amharic
word2vec	25	25	auto	72.02	Amharic
word2vec	5	5	auto	81.79	Tigrigna
word2vec	10	10	auto	73.65	Tigrigna
word2vec	15	15	auto	69.54	Tigrigna
word2vec	20	20	auto	64.97	Tigrigna
word2vec	25	25	auto	61.73	Tigrigna

TABLE IV. Top 20 Topics Represented from Afaan Oromo Text Comments Topics

Topic	Count	Name	Topic description
0	14409	0_barnootaa_oromiyaa_godina_gaallaa	Describe about "Ethnically based hate."
1	351	1_bbc_tv_televijiinii_channel	Describe the Media:
2	320	2_sodaa_sodaata_sodaatin_sodaatu	Describe fear
3	285	3_milkii_milkaa_milkiin_minilk	Describe success
4	249	4_amara_amaraan_amaran_kehil	Describe the Amhara ethnic group
5	229	5_galatooma_galatoomaa_galatoomi_galatoomii	Describe about thanks
6	212	6_galatomii_galatomi_galatoma_galatomaa	Describe about thanks
7	211	7_minilik_miniliki_minilikii_diqalaa	Describe hate speech target person

8	188	8_ahmed_ahmad_abiy_abi	Describe about person
9	180	9_poolisiin_poolisii_federaalaa_pool	Describe police commission

B. Amharic Social Media Data Topics and Description

The application of BERTopic on the Amharic social media dataset has resulted in topics containing text comments across the entire dataset. This study selected and analyzed the top 21 topics despite the top 8 listed in Table V. Amharic social media dataset generated 1044 topics from Amharic text comments.

TABLE V. TOPICS EXTRACTED FROM AMHARIC SOCIAL MEDIA DATA

Topic	Count	Name	Descriptio n of Topics Extracted
0	18868	0_ሀይል_መንግስት_ሰራዊት_ከተማ	Describe military
1	560	1_ወሎ_እውነትነው_አይዴላም_እንዳናስ ተውል	Describe appreciation
2	142	2_ደስ_ይላል_አለሽ_አለህ	Describe appreciation
3	142	3_አለ_ፈላሻው_ብትሆንልን_ዛዛታ	Describe hate speech
4	132	4_የኢትዮጵያ_ጠላት_አምላክ_ደራርቱ	Describe hate speech
5	106	5_እውነት_ብለሻል_ይገላልና_ቁጣ	Describe hate speech
6	67	6_ሆይ_ጌታ_ይቅርም_ድረስልን	Describe about religion
7	66	7_ነፍሳቸውን_ያኑርልን_በገነት_በአፀደ	Describe condolenc es
8	65	8_ይማር_ነፍስ_ነብስ_ያማን	Describe condolenc es

Topics 1 to 4 provide normal information. The Amharic social media dataset was used to extract topics with sizes of 18868, 560, 142, and 142, including normal content. Topic 5, with 132 text comments, outlines the Process of insulting and identifying a target to an individual. Topic 10, with 65 text comments, discusses hate crimes against specific individuals, including insult and identity hate. Topic 21 indicates that the normal content is attributed to individuals with a comment size of 45.

C. Tigrigna Social Media Data Topics and Description

We analyzed 21 Tigrigna text comments and revealed that topics 1 and 2 provide normal information, while others contain hateful content about nationalist targets. The description of the top 10 topics is indicated in the Table VI below.

Table VII indicates the common topics extracted by applying BERTopic from the Ethiopian language social media dataset to data by experiment. Those topics extracted from Ethiopian social media data are normal or hate classified as antagonistic, identity hate, insult, and threats. As indicated in Table VI, alongside providing normal content, Afaan Oromo users utilize social media to spread hate against identity. Afaan Oromo users' people are also posting hate speech on Facebook in the form of insults. The experiment shows that insult is hating speech posted on social media platforms from Amharic text comments in addition to identity hate.

The top two topics extracted from Tigrigna indicate normal information, whereas the third describes individual nationality and antagonistic information targeted to individuals. Information emerging on Facebook pages in Afaan Oromo, Amharic, and Tigrigna generally relates normal information and hate speech. As Table VI illustrates, normal, insult, threat, antagonistic, and identity hate are the common types of information disseminated by Ethiopian social media data.

TABLE VI. TOPICS EXTRACTED FROM TIGRIGNA SOCIAL MEDIA DATA

Count	Name	Description
23723	0_እዩ_ናይ_ኣምሓራ_ህዝቢ	Describe the Amhara ethnic group
396	1_ገለቴ_ፖፒ_ኣማርድ_ሓልዎን	Describe individual
147	2_ነፍሲ_ወከፍ_ትመሃር_መድረኽ	Describe the stage of learning
49	3_በዓል_ርሑስ_ልደት_ኢሬቻ	Describe the Irrecha celebration
36	4_ፍትሒ_ደለይቲ_ኣቱም_ብደለይቲ	Describe about justice
30	5_ሰናይ_ለይቲ_ይግበረልኩም_ፍሽኽታ	Describe a good evening
23	6_ጅግና_ኣያና_ኣያኒ_ደብሪፅ	Describe the appreciation of Debretsion
23	7_ዓቃቢ_ሕጊ_ተቓውሞ_ተቓዋሙ	Describe against somebody
20	8_ኣበይ_ኔርካ_ሓሲብካኒ_ራብከን	Describe individual- based hate
19	9_ሰዓት_መስከረም_ንግሆ_ከይሰማዕኩም	Describe about ethnic group- based hate

TABLE VII. COMMON TOPICS FROM ETHIOPIAN LANGUAGES SOCIAL MEDIA DATA

S/N o	Languages	norm al	identit y hate	insult	threat	antag onisti c
1	Afaan Oromo	yes	yes	yes	yes	yes
2	Amharic	yes	yes	yes	yes	yes
3	Tigrigna	yes	yes	yes	yes	yes

VI. CONCLUSION

Topic modeling is used in resource rich languages, such as English, for analyzing massive amounts of data. This study examines various modeling techniques, such as LDA, LSA, PLSA, NMF, and BERtopic, for extracting topics from low-resource Ethiopian-language social media data to analyze hateful content. In this study, we collected 59529 text comments from the Facebook pages of Afaan Oromo, 45522 from Amharic, and 48882 from Tigrigna. The text preprocessing technique was applied to text comments, and a topic modeling approach was used to extract topics from preprocessed texts. The experiment findings show that BERTopic has a better coherence score than the others. This work employed BERTopic techniques to identify topics from Facebook comments in Afaan Oromo, Amharic, and Tigrigna, using pre-trained word2vec as language embedding for the translation of texts. We evaluated BERTopic's performance by setting parameters for topic extraction in low-resource datasets, training Word2Vec for text document transformation, and setting nr-topics to auto for optimal tuning. The second experiment demonstrates that Afaan Oromo, Amharic, and Tigrigna have topic coherence scores of 82.74, 87.85, and 81.79 percent at 5 n_neighbors, minimum_cluster size, and number of topics set to auto. The study revealed that Facebook users in Afaan Oromo, Amharic, and Tigrigna languages are posting both normal and hate content, including identity hatred, insults, threats, and combative language. The BERTopic model was assessed for its efficacy in addressing the challenges of short text, spelling errors, and context words in low-resource languages such as Afaan Oromo, Amharic, and Tigrigna. From the extracted topics we also concluded that both normal and hateful content are posted as comments. The experiment dataset in Ethiopia uses Afaan Oromo, Amharic, and Tigrigna languages, primarily from Facebook, with plans to expand to include other Ethiopian languages on other topic platforms.

ACKNOWLEDGMENT

The support of Addis Ababa Science and Technology is greatly acknowledged. The author also wants to appreciate the support of Dr Kula Kekaba and Dr Solomon Teferra for their support scholarly for the completion of this paper.

REFERENCES

[1] D. Ediger et al., “Massive social network analysis: Mining twitter for social good,” Proc. Int. Conf. Parallel Process., no. May 2014, pp. 583–593, 2010, doi: 10.1109/ICPP.2010.66.

[2] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab, Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis, vol. 56, no. 6. Springer Netherlands, 2023. doi: 10.1007/s10462-022-10254-w.

[3] J. C. Campbell, A. Hindle, and E. Stroulia, “Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data,” Art Sci. Anal. Softw. Data, vol. 3, pp. 139–159, 2015, doi: 10.1016/B978-0-12-411519-4.00006-9.

[4] I. Vayansky and S. A. P. Kumar, “A review of topic modeling methods,” Inf. Syst., vol. 94, no. June, p. 101582, 2020, doi: 10.1016/j.is.2020.101582.

[5] Z. Zhang, M. Fang, L. Chen, and M. R. Namazi-Rad, “Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics,” NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., pp. 3886–3893, 2022, doi: 10.18653/v1/2022.naacl-main.285.

[6] C. A. Calderón, G. de la Vega, and D. B. Herrero, “Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in Spain,” Soc. Sci., vol. 9, no. 11, pp. 1–19, 2020, doi: 10.3390/socsci9110188.

[7] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, “Detection of hate speech in COVID-19-related tweets in the Arab Region: Deep learning and topic modeling approach,” J. Med. Internet Res., vol. 22, no. 12, 2020, doi: 10.2196/22609.

[8] A. Obadimu, E. Mead, T. Khaund, M. Morris, and N. Agarwal, “Utilizing Topic Modeling and Social Network Analysis to Identify and Regulate Toxic COVID-19 Behaviors on YouTube,” Sbp-Brims.Org, pp. 1–9, 2020, [Online]. Available: https://developers.google.com/youtube/v3/docs/search/.

[9] S. Deerwester, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis Scott,” Kehidupan, vol. 3, no. 12, p. 34, 2015.

[10] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” 2022, [Online]. Available: http://arxiv.org/abs/2203.05794.

[11] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” Inf., vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[12] P. Ghasiya and K. Okamura, “Investigating COVID-19 News across Four Nations: A Topic Modeling and Sentiment Analysis Approach,” IEEE Access, vol. 9, pp. 36645–36656, 2021, doi: 10.1109/ACCESS.2021.3062875.

[13] S. Liu and T. Forss, “New Classification Models for Detecting Hate and Violence Web Content,” vol. 1, no. Ic3k, pp. 487–495, 2015.

[14] T. Davidson and D. Bhattacharya, “Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling,” pp. 2–5, 2019.

[15] D. Angelov, “Top2Vec: Distributed Representations of Topics,” pp. 1–25, 2020, [Online]. Available: http://arxiv.org/abs/2008.09470.

[16] N. F. F. d. Silva et al., “Evaluating Topic Models in Portuguese Political Comments About Bills from Brazil’s Chamber of Deputies,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 13074 LNAI, no. September, pp. 104–120, 2021, doi: 10.1007/978-3-030-91699-2_8.

[17] N. B. Defersha, J. Abawajy, and K. Kekeba, “Deep Learning based Multilabel Hateful Speech Text Comments Recognition and Classification Model for Resource Scarce Ethiopian Language: The case of Afaan Oromo,” Proc. 2022 IEEE Int. Conf. Curr. Dev. Eng. Technol. CCET 2022, 2022, doi: 10.1109/CCET56606.2022.10080837.

[18] G. Neshir, A. Rauber, and S. Atnafu, “Topic modeling for amharic user generated texts,” Inf., vol. 12, no. 10, pp. 1–21, 2021, doi: 10.3390/info12100401.

[19] K. Assefa and W. Bank, “Short Amharic Text Clustering Using Topic Modeling,” no. November, 2020, doi: 10.13140/RG.2.2.17462.32326.

[20] A. Abdel-Hafez and Y. Xu, “A Survey of User Modelling in Social Media Websites,” Comput. Inf. Sci., vol. 6, no. 4, 2013, doi: 10.5539/cis.v6n4p59.

[21] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min., pp. 399–408, 2015, doi: 10.1145/2684822.2685324