IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Extracting Semantic Topics about Development in Africa from Social Media

HARRIET SIBITENDA [1,2] (Graduate Student Member, IEEE), AWA DIATTARA[1], ASSITAN TRAORE[3], RUOFAN HU[2], DONGYU ZHANG[2], ELKE RUNDENSTEINER[2], and CHEIKH BA[1]

[1]University of Gaston Berger, Saint Louis, Senegal: sibitenda.harriet; awa.diattara; cheikh2.ba@ugb.edu.sn
[2]Worcester Polytechnic Institute, Worcester, MA, USA: hsibitenda; rundenst; dzhang5; rhu@wpi.edu
[3]Actroll, France: assitan.traore@free.fr

Corresponding author: Harriet Sibitenda (e-mail: nkarietn@gmail.com)

This work was supported in part by the PASET-RSIF scholars program

ABSTRACT The extraction of knowledge about the prevalent issues discussed on social media in Africa using Artificial Intelligence techniques is vital for informing public governance. The goals of our study are twofold: (a) to develop machine learning-based models to identify common topics of social concern about Africa on social media, and (b) to design a classifier capable of inferring a particular common topic associated with a given social media post. We designed a three-step framework to achieve the former goal, namely, topic identification. The first step uses text-based representation learning methods to generate text embeddings for feature representation. The second step leverages state-of-the-art Natural Language Processing models, commonly called topic modeling, to organize the representations into groups. The third step generates topics from each group, including by means of using large language models to generate meaningful short-sentence labels from the bag-of-tokens associated with each group. Furthermore, we use Llama2 to deduce the words into a single word theme that describes each topic in relation to social concerns about development. To achieve the second goal of classification; we trained classifiers using ensemble voting and stacking learners to infer which among the identified common topics best characterizes the social media post. For our experimental study, we collected a text corpus called Social Media for Africa composed of 22,036 records extracted from social media comments on Twitter (X) and YouTube. The clustering-based model BERTopic yielded 304 topics, at topic coherence 0.81 C-v. On merging the topics into classes, the BERTopic+ created 11 common topic classes at topic coherence 0.76 C-v. For theme extraction, we additionally refined the leading token words with Llama2 to generate concise single-word theme labels, resulting in 98 unique themes by BERTopic_theme with a C-v score of 0.75 and an IRBO score of 0.50. We then utilized the identified topics based on the resulting groupings as labels for training a topic classifier. These labels were created using Llama2 on our SMA corpus. Our comparative study of topic classifiers using stacking and voting schemes shows that the BERTopic model features 0.83 accuracy and 0.82 F1 score with ensemble voting for training on topics. Furthermore, training on topic classes, BERTopic+ with ensemble voting had the highest accuracy of 0.95 and F1 score of 0.95 compared to other alternate methods on our corpus. The BERTopic_theme also achieved higher performance with ensemble voting classifier at 0.93 F1 score and accuracy 0.93. The overall performance of classifiers using the ensemble stacking is slightly better than that of voting methods for short sentence topic labeling. For Africa, policymakers should focus on the most pressing social issues: COVID-19 restrictions affecting public health and economic recovery, promoting entrepreneurial innovation in energy and environmental sustainability to combat climate change, and strategically responding to China's rise in global politics to maintain geopolitical stability and foster international cooperation.

INDEX TERMS social concerns, social media, semantic topic labels, themes, topic modeling, LLMs.

I. INTRODUCTION

A. MOTIVATION AND BACKGROUND

In our daily lives, we face problems that affect many people in society, commonly referred to as social issues [1]. These include terrorism, unemployment, political instability, poverty, human trafficking, and cultism [2]. These social problems often differ from society to society: for example, the growth of Africa has been stagnant since 2019 owing to a

variety of potential factors, such as the coronavirus pandemic and the Russian invasion of Ukraine [3]. Sociologists collect information on social issues to provide informed feedback to policymakers. Sociologists use four major methods to collect social concerns. These include (i) surveys that involve gathering responses from participants in a sample study that are representative of a larger population, (ii) experiments in natural and physical sciences that examine cause-effect relationships, (iii) observations with field-based sessions to watch the participants, and (iv) leveraging existing datasets that have previously been collected for other reasons [4, 5]. They collected data from primary sources such as questionnaires, interviews, online posts, and comments. Alternatively, sociologists may use secondary sources that have already been recorded, such as census data and biographies. For policymakers, gaining a better understanding of the issues affecting people is critical to achieving social protection. Endorsements from international organizations such as the International Labor Organization (ILO) highlight the significance of collecting and utilizing social feedback [6]. They demonstrate the use of feedback about social concerns to (i) invest in data availability to assess action plan policies, (ii) acquire shock-responsive systems to address solutions to unforeseen calamities, (iii) enhance real-time communication between the public and government through social dialogue, and (iv) ensure justice and legal satisfaction with rights-based policies.

A study [7] examining social concerns in East Africa underscores the need for more effective strategies to address these issues. A key strategy proposed was to establish a coherent administration system for tracking, monitoring, and evaluating activities for social protection. The analysis of social issues must be timely and crucial for tracking the impact of changes and strategies for sustainable development on people.

To this end, there is a need to track public complaints and sentiments on social media to monitor developments. The analysis of social concerns may reveal valuable insights into various social issues and the entities actively involved in addressing them.

B. OBJECTIVE

This study aims to introduce and assess artificial intelligence (AI)-driven methodologies tailored to examine the social issues discussed on African social media platforms. These AI models can potentially assist policymakers in promptly gathering valuable insights into societal concerns. The objective of leveraging this public feedback is to achieve more sustainable development goals. To realize this objective, we begin by reviewing the literature to identify methodologies suitable for gathering social concerns from Social Network Platforms (SNPs), and subsequently, customize these approaches as needed.

Thereafter, we collected a dataset of social media suitable for conducting this analysis. Given the assumption that distinct discussion categories about development exist in social comments, we extract common semantic topics from the collected corpus. To refine these topics, we employed LLMs to transform the top k words associated with each topic into concise sentence labels. These initial sentence labels were broad, so we further refined them to specifically highlight social concerns related to development. Thus, we created a general theme description by focusing on words within these top k words that were directly related to social concerns about development. Finally, we evaluate the relative effectiveness of our methods to identify common patterns of interest that indicate public issues about development in social media data sets.

C. RELATED LITERATURE

Pulido et al. [8] examined the impact of social information on health by analyzing posts on Facebook (200 posts with 13,076 comments), Twitter(X) (17,117 tweets), and Reddit (397 comments). They used health-related keywords like 'health,' 'vaccines,' 'nutrition,' and 'Ebola' to collect data. A codebook categorizes the comments into false news, misinformation, opinion, or facts, followed by manual analysis. While thorough, this manual categorization was time-consuming and only practical for relatively small datasets.

Gray [9] analyzed gender bias in comments on political Facebook pages using the Scrapy framework to collect 15,000 posts, bypassing the limitations of the Facebook Graph API. She categorized reactions to these posts in order to assess gender bias. However, the study was limited to public pages and lacked a deeper analysis of public comments on gender issues or common topics in each category.

Blasi, Gobbo, and Sedita [10] used Twitter (X) accounts for Italy's 28 largest municipalities to explore social concerns in near real-time, acting as smart city connectors between citizens and government officials. Over two months, they collected 316,801 comments from these accounts using the Twint scraper. They analyzed social concerns by examining common topics and their most frequent words. However, relying on a single social network limits the views of non-Twitter users. Including data from various social media platforms would offer more comprehensive feedback to the government.

Bihn and Dao [11] analyzed job descriptions from 18,992 online job posts collected using job- and skill-related keywords from the Kaggle dataset. They created topics and labels with unsupervised text clustering using metrics such as Silhouette, Calinski-Harabasz, and Davies-Bouldin scores, as well as C-v and C-UMass coherence, to evaluate the clusters. K-means was the most effective clustering method. However, they did not perform a deeper analysis of the semantic meanings of the generated topic labels.

Vestergaard and Kästel [12]used text clustering to identify topics in datasets, including customer emails (6,644 records), stock exchange posts (13,943 records), and BBC news articles (2,225 records). They preprocessed the data and used BERT sentence embeddings for feature extraction and applied K-means and DBSCAN for clustering. The clusters were manually labeled with single-word titles based on the most frequent

words. The clustering quality was assessed by building and evaluating classifiers. K-means outperformed DBSCAN with 0.88 precision and 0.87 recall, compared to DBSCAN with 0.34 precision and 0.52 recall. Manual labeling relies on the researcher's expertise, highlighting the need for automatic label extraction strategies, which will be part of our proposed research.

In [13], Jaleel et al. explored text clustering from customer reviews using the Kaggle dataset with 65,535 records. They generated features using TF-IDF and created clusters using a hybrid method combining k-means and agglomerative clustering. The CTFIDF method, which uses cosine similarity and TF-IDF, was employed to automatically generate multi-word topic labels by concatenating the top keywords. These topic labels, which categorized and summarized the content of the clusters, were then evaluated using a Naïve Bayes classifier, achieving 0.87 accuracy, 0.73 precision, and 0.79 recall. A notable limitation is the use of only one classifier for evaluation. Providing comparative results with multiple classifiers would be more beneficial.

Hee and Wei [14] and Prakash et al. [15] suggested using Large Language Models (LLMs) to generate more semantically accurate labels. They employed BERTopic, Llama, and ChatGPT to extract topic labels from the top-word tokens. Two methods were proposed for creating topics at the sentence level: Prompt-Based Matching (PBM) and Word Similarity Matching (WSM). PBM prompts ChatGPT to return the top 10 words related to each topic based on c-TF-IDF scores. WSM tracks the similarity of topic pairs using the highest CTFIDF scores and merges similar topics using the CTFIDF algorithm. Our study builds on this by exploring the potential of LLMs to generate concise sentence-level topic labels, a component not covered in their research. Specifically, we employed LMMs to develop semantic short-sentence labels and compared labels gathered from various text clustering and topic modeling methods.

D. OUR APPROACH: SHORT-SENTENCE TOPICS FROM SOCIAL MEDIA DATA

Building on previous research, we followed these steps in this study. Initially, we have data collection and preprocessing. We reviewed past studies to identify the sources of data and methods used by researchers. Based on these comparisons, we selected tools to gather data from Twitter (X) and YouTube social networks using web scrapers like BeautifulSoup, Selenium, and Snscrape. Subsequently, we conducted text preprocessing to clean the data and normalize all texts to a standard format, including translating non-English languages to English where applicable. We then filtered all non-English texts, resulting in a dataset of 22,036 records. We employed feature extraction methods to generate features for machine learning algorithms, including TF-IDF, pretrained Word2Vec, and Word2Vec trained from scratch on the Social Media for Africa (SMA) called Word2Vec-SMA, FastText, and BERT.

For topic identification, we compared different methods of topic modeling (text clustering, conventional, and neural topic models) that generate groups of topics with labels. We deduced the top-token word labels for short sentences with semantic meanings by LLMs. To evaluate topic generation and classification, we used coherence and divergence metrics. The BERTopic has good performance, with 304 topics at 0.81 C-v topic coherence and 0.58 IRBO topic divergence on the SMA corpus. By merging topics into groups, BERTopic+ achieved a topic coherence of 0.76 C-v and a topic divergence of 0.43. The low topic divergence implies topics are closely related because the dataset includes only comments about social concerns to development. Using the BERTopic pipelines, we identified topics that were closely related and not so divergent. For theme title refinement, we further use Llama2 to refine identified topics into concise and meaningful thematic labels that are easy to understand and directly relevant to social development concerns. We design a custom prompt that instructs Llama2 to generate thematic titles for relevant topic sets. If the topics do not align with predetermined social concerns, the output defaults to no theme titles. This ensures precision in thematic categorization. We found that BERTopic_theme identified 98 themes at 0.75 C-v and 0.50 IRBO scores.

To categorize unseen comments, we trained classifiers using ensemble voting and stacking techniques to infer topics from the social media comments. We evaluated the classification models using the accuracy, precision, recall, and F1-score. The BERTopic achieved high results for ensemble voting classification, with an F1 score of 0.82 and an accuracy of 0.83. From the findings, we conclude that using BERTopic for topic identification provided us with meaningful semantic topics and allowed us to train ensemble classifiers with higher accuracy scores than other topic modeling techniques. Further, we deduced the topics by BERTopic to smaller groups, using the cluster-based model BERTopic+ generating 11 topic classes. The ensemble stack-based classifier achieved classification with 0.95 accuracy and 0.95 F1 score. Moreover, the classifier associated with this model achieved a higher accuracy score than BERTopic and other topic modeling algorithms, as indicated by its output. For the BERTopic_theme, the ensemble voting strategy achieved an 0.93 F1 score and accuracy at 0.93.

Contributions.

Our research contributes to the field of social media analysis as follows:

- Comprehensive methodological framework: We design a unified framework integrating a comprehensive collection of alternate topic extraction and modeling techniques on our social media data set relevant to social development in Africa. This holistic framework covers everything from data collection to the evaluation of analysis techniques, setting the stage for studies by governments and other agencies aiming to understand social dynamics through AI.
- Integration of LLMs for topic and theme generation: One contribution of our work is to leverage LLMs to

generate short sentence labels, which represents an innovative utilization of LLM techniques for tackling our topic modeling application goal. We pioneer the use of LLMs to generate single word precise thematic titles from complex social media data based on the topic tokens, enhancing the interpretability and applicability of topic analysis results.

- Automated relevance filtering: Our approach includes an automated mechanism to filter out irrelevant topics, providing outputs only when topics meet the criteria for social development relevance. This increases the utility of our findings for policy-making and social strategy formulation.
- Empirical validation of AI techniques in Social Contexts: By rigorously testing the effectiveness of a diverse set of state-of-the-art AI models in identifying and classifying social media topics, we contribute empirical evidence to the ongoing discourse on the practical applications of AI in social sciences. Our findings also provide guidance on which of the methods interested organizations with social development responsibilities may want to select to do their analysis of social media topics.
- Data Set Resource Contribution:
  In addition, this work curates a relevant social media data set that we will be released to open source to provide a valuable resource to the research community on studying social development issues in Africa.

These contributions not only advance the technical capabilities of social media analytics but also enhance their practical implications, directly impacting how social issues are understood and addressed at the policy level.

The rest of the paper is organized as follows. Section II describes the methods for our KDD pipeline composed of data collection, data preparation, topic extraction, and topic labeling. Section III presents the outcomes of the methods used for topic extraction and model evaluation. This section details the results from different models of topic modeling to generate common topics, as well as the analysis of these topics and the classifiers developed based on them. Section IV discusses our findings, a comparison of results with related studies, and the limitations of the study. Section V presents the conclusions of the study objectives and recommendations for future work.

II. METHODOLOGY

In this study, we adopted the KDD pipeline introduced in [16] to identify valid patterns in data [17]. This approach supports the prediction of future trends and aids in informed decision-making. To extract insights about discussion topics from social network comments, we followed five steps: (i) collecting data from social network platforms (SNPs) and selecting target files; (ii) preprocessing and cleaning the data into a clean format; (iii) transforming the cleaned data into numerical features suitable for machine learning models; (iv) building and training models using machine learning algorithms; and (v) analyzing the results and forming conclusions. Figure 1 illustrates the architecture of this study and the steps adopted.

A. DATA SELECTION FOR TARGET AREA OF INTEREST

The usage of the Internet in Africa was expected to reach 43% by 2021 [18]. By 2022, the number of users surpassed 570 million, with Nigeria accounting for 100 million, Egypt 76 million, and South Africa 41 million [19, 20]. Internet access and usage were the highest in North Africa at 56%, followed by South Africa at 45%, West Africa at 14%, and Central Africa at 8%. A report by the International Telecommunication Union [21] revealed that the most common uses of the Internet involve social networks, with Facebook accounting for 86.55%, YouTube 7.14%, Twitter (X) 2.76%, Instagram 1.7%, and Pinterest 1.56%.

In this study, we focused on the same popular social network platforms (SNPs) that give developers access rights, such as YouTube and Twitter (X). We collected posts on social concerns in Africa from these platforms for further analysis. A public concern is common among many people in the population. According to [22], each public issue has a starting period, and it may take some time for the authorities to gain full awareness. That is, at a certain point in time, people and organizations formally raise concerns and report to the authorities in charge of mobilizing action plans. We observed a need to identify persistent social issues earlier, namely when they were initially triggered, which can be many years before they are large and prevalent.

In our study, we collected social media data over a five-year period from 2018 to 2022, using keywords related to social concerns in Africa, including social problems, challenges, worries, issues, and questions [23].

B. DATASET COLLECTION

Our first goal was to identify appropriate methods for collecting data on social concerns from SNPs. To solve this problem, we reviewed the literature to select open-source tools that support and possibly automate the historical data collection process. We focused on collecting social posts from popular social networks, such as Twitter (X) and YouTube.

1) Methods for Data Collection

To conduct a state-of-the-art review [24] of the methods applied for data collection, we followed the PRISMA protocol guidelines [25]. Using the "Publish and Perish" open software, we selected the Google Scholar index database to collect 2180 articles. We set the time range from 2018 to 2022 and used different search strings. Using "and" the logic operator, we joined words such as "data collection", "social networks", "Twitter (X)", and "YouTube". For the exclusion criteria, we excluded 1612 articles without content related to social network comments and included only 298 articles for manual analysis. The selected articles contained details regarding the methods used for data collection from SNPs. Based on these reviews [24], five methods were identified.

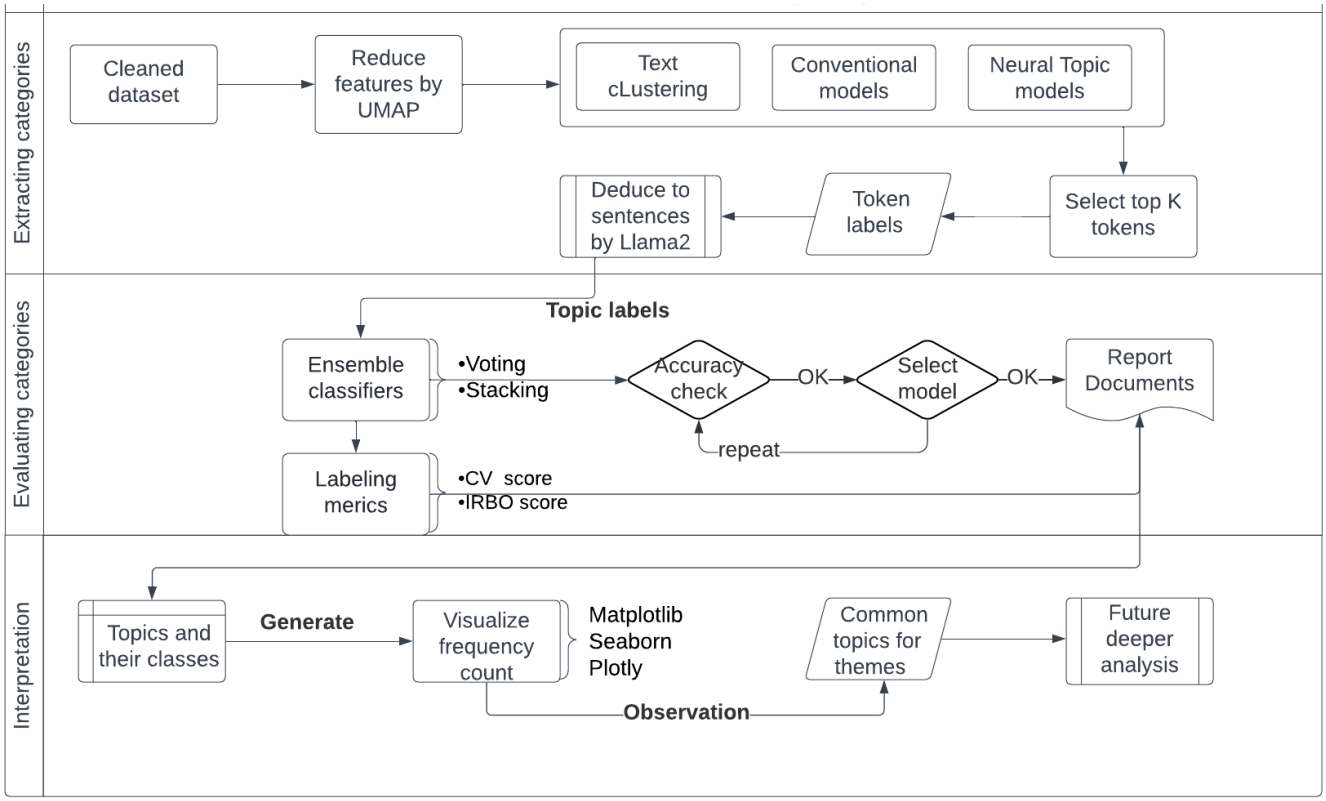


FIGURE 1. **The proposed architecture based on the KDD theory adopted from Fayyad et al., 1996**

These include self-report surveys, public APIs, web crawlers or scrapers, public repositories, and manual observations. Based on this analysis, we selected open software tools in the form of public APIs and web scrapers that support extensive data collection at no cost for tools without restrictions on automatic data access, such as Twitter and YouTube.

Collections from Twitter (X)

First, we used the Twitter (X) social networking tool. According to Stokel-Walker [26], Twitter changed its brand name to X. This study refers to this social media platform as Twitter (X). Twitter (X) is a social networking site where users broadcast short posts known as tweets [27]. It is legal to download content from Twitter (X) by using its API [28]. Because the primary APIs limit data access to seven past days, we used web scrapers or crawlers to retrieve historical social comments. We used a web scraper called Snscrape [29]. A scraping tool released in 2020 to collect data from social networking services such as Twitter (X), Facebook, and Reddit, offers a simpler method for collecting large historical datasets than tools such as Twint, TwitterScraper, and Twarc.

Using Snscrape, we collected a dataset of 9,198 records that included columns such as text, ID, username, likes, replies, retweet counts, date, and language. In adherence to research ethics, we removed the username. The short text posts have an average word count of 30 words.

Collections from YouTube

Second, we used YouTube as a social networking tool. YouTube is "a video sharing service where users can watch, like, share, comment and upload their own videos" [30]. We utilized two methods for data collection: the YouTube Data API and the web scraper tools BeautifulSoup and Selenium. This dual approach allowed us to collect a larger volume of data (18.853 records), circumventing the limitations imposed by the API.

Beautifulsoup is "a Python library used to crawl over the data in HTML and XML documents" [31]. Selenium is "an open-source tool that automates testing on web browsers using JavaScript and iframes." [32], [33], [34]. The dataset columns include text about the video title and description, ID, username, views, and time. We excluded the username column to preserve user privacy. The dataset contained short texts with a mean word count distribution of 20 words.

2) Description of Our Collected Dataset: In a Nutshell

In summary, our dataset was composed of 18,853 records from YouTube and 9,198 records from Twitter (X). In our subsequent analysis, we focused on columns containing the actual text comments on YouTube and Twitter (X). These columns were called "Title," "Description," and "Text," respectively. The word count for each record was less than 200.

C. TECHNIQUES FOR EXTRACTING TOPICS

1) Data preprocessing

In this study, we used data integration methods to combine data from various sources. Automated database systems were used to combine the files and form the structured data. For this, we loaded each csv using PySpark to create table schemas, generating new columns: "id" and "sourcetype." We aggregated the text columns to create a single data frame and dropped empty records and duplicates. Finally, we created a MySQL database to save the data as persistent tables.

For data cleaning, we removed unwanted characters such as punctuation marks, numbers, and hashtags, converted short word forms, and corrected spellings. We also converted or removed the emoticons and emojis. Lastly, we detected language, and when a language other than English was detected, we attempted to convert it to English. We deleted some texts whose local languages were not detected by Google translator. Thus, our cleaned dataset consisted of only English text.

Data normalization is the conversion of text into a single standard form to reduce word redundancy. For this, we performed the standard NLP text processing steps. This included tokenization of the text into single-word chunks called tokens. We also removed stop words and the most common words from the texts, and applied stemming and lemmatization to reduce inflected words to their root forms or lemmas.

Because different NLP tasks may require different forms of text input, we kept records of both cleaned and normalized columns in the dataset. The cleaned Social Media from Africa dataset (SMA) had 22,036 records. In this study, the SMA corpus was used for topic identification and classification. Table 1 presents a sample of cleaned data.

2) Data Transformation and Representation

Next, using feature extraction, we transformed the data into numerical features required or machine learning and subsequent processing [35, 36]. In particular, we used two types of feature extraction techniques: statistical feature representation and word embedding.

Statistical Feature Representation

Statistical feature representation encompasses methods that create feature vectors based on word frequency, including one-hot encoding, bag-of-words, bag-of-n-grams, and term frequency–inverse document frequency (TF-IDF). One-hot encoding transforms a categorical variable into a binary vector, where the value is zero if a word is not in the document and 1 if a word appears in the document. The Bag-of-Words (BoW) model creates a vocabulary of unique words in a document and computes word frequencies. The Bag-of-N-grams (BoN) model creates unique n-grams, which are groups of n adjacent words like unigrams, bigrams, and trigrams. It captures syntactic patterns and critical semantic context by combining adjacent words. The limitation is that the frequency of the common words may not be very informative and important about the document's content.

Term Frequency–Inverse Document Frequency (TF-IDF) extracts features based on the importance of terms across the corpus. Term frequency (TF) measures the frequency with which a term appears in a document. IDF (Inverse document frequency) is the exclusivity or uniqueness of a particular word in a particular document. [37, 38]. TF-IDF is better than the other models because it captures the importance of words to a certain degree. Although TF-IDF captures word similarity and sequencing, it ignores the local semantic context. Moreover, it still suffers from high dimensionality of vector problems [39, 40, 41, 42, 43]. In this study, we used TF-IDF as the baseline word representation method.

Wording Embeddings

Last but not least, word embeddings generate feature vectors based on the vector space distribution of words using neural networks. Word embeddings have several advantages over more traditional representations. They are compact and controllable in size, and words with similar meanings have a similar representation [44, 45]. There are two types of word embedding: static and contextualized. Static word embeddings consider words as fixed dense vectors for grouping similar sentences. Each word is represented as a single prototype vector that does not change with the context meaning. Thus, they ignore the grammatical sense of polysemous target words with many contextual meanings in different documents. [43, 45]. Alternatively, contextualized word embeddings are more semantically meaningful.They are capable of representing words in context across different documents as the context changes.

Static Word Embeddings. In this study, we mainly explored Word2vec and FastText. Word2vec [46] creates vector spaces of words using shallow neural networks (NNs). The model works by either learning from new corpus data or training on existing pretrained vectors. The model analyzes semantics and similarities based on the existence of target words in documents [47]. It has two basic architectures, Continuous Bag-of-Words Model (CBOW) and Skip-gram [48]. The CBOW mopdel predicts the target word based on the context of surrounding words. Skip-gram uses a target word to predict the context of surrounding words. [49, 50]. CBOW operates faster than Skip-gram on large datasets and has a higher prediction accuracy for frequent words. In this study, we trained a Word2Vec model with CBOW from scratch on our SMA dataset called Word2Vec-SMA. Pretrained Word2Vec provides a global word representation of vectors [51, 52]. In this study, we compared the performance of Word2Vec-SMA and a pretrained Word2Vec model trained on the Google News corpus [53].

FastText is based on Bojanowski's SG model, which treats each word as a bag of character n-grams [54]. It has improved the representation of out-of-vocabulary words and works well with large corpora. The limitations of FastText include ignoring the grammatical sense of words. Our study compared the embeddings of Word2Vec-SMA, pretrained Word2Vec, and FastText.

Contextualized Word Embeddings are words based on the meanings derived from the sentences in which they ap-

TABLE 1. Example of changes to text made by NLP tools for text preprocessing

	Activities	Result
Raw Text	Original text	🔴 Pr Macky Sall répond à Sonko "je ne laisserai pas... HSA Social Science Important Questions 2021 | ഈ ചോദ്യങ്ങൾ തീർച്ചയായും അറിഞ്ഞിരിക്കുക
Data cleaning	Convert emoji, emoticon	**red_circle** Pr Macky Sall répond à Sonko "je ne laisserai pas... HSA Social Science Important Questions 2021 | ഈ ചോദ്യങ്ങൾ തീർച്ചയായും അറിഞ്ഞിരിക്കുക
	Translate text to English	red_circle Pr Macky Sall **responds to Sonko "I won't let..." HSA Social Science Important Questions 2021 | Be sure to know these questions**
	Lowercase, remove punctuations	redcircle pr macky sall responds to sonko i wont let hsa social science important questions be sure to know these questions
	Handling short forms and spellings	redcircle pr macky sall responds to sonko i wont let hsa social science important questions be sure to know these questions
Data normalization	Removing stopwords	redcircle pr macky sall responds _sonko _wont let hsa social science important questions _sure _ know _ questions
	Rooting by lemmatizers	redcircle pr macky sall responds sonko wont let hsa social science important question sure know question

pear. Language model-driven embeddings are contextual; they consider the terms surrounding the target words to generate linguistically-based representations [55]. Contextualized embeddings are widely used in various NLP applications, including classification, question-answering, and summarization. The most popular models for generating dynamic word embeddings are from the Bidirectional Encoder Representation Transformer (BERT) family. BERT uses self-supervised learning to learn the contextual connections between words without labeled data. In the pre-training phase, BERT utilizes two main strategies: the Masked Language Model (MLM) and Next Sentence Prediction (NSP) to learn the grammatical semantics of sentences. In our study, we used the pretrained sentence BERT embeddings called "paraphrase-MiniLM-L6-v2," which supports paraphrasing text for NLP tasks with efficiency and high computation power[56].

3) Dimension Reduction

Word embeddings typically have high dimensionality. For an easier analysis, it is necessary to reduce these to lower dimensional features through feature selection or dimension reduction. Commonly used methods for dimension reduction include PCA, t-SNE and UMAP [57]. PCA is a simple technique that requires high computational resources, such as memory, to generate densely clustered features. t-Distributed Stochastic Neighbor Embedding (t-SNE is a nonlinear dimension-reduction technique [58], that uses the perplexity parameter to compute similarity and govern the nearest neighbors for each point. t-SNE maintains the global structure of the data. Uniform Manifold Approximation and Projection (UMAP) is an advanced nonlinear dimension reduction technique that preserves local and global data structures in lower-dimensional space [58]. UMAP determines a local radius using parameters such as the number of neighbors and minimum distance, excelling over t-SNE to maintain the integrity of the nearest local points, while disregarding distant points to preserve the global structure [59]. This is faster than t-SNE and produces denser and, more compact clusters. In our work, we utilize UMAP owing to its speed and efficiency in maintaining the global structure at a relatively low computational cost.

4) Clustering

Text clustering is a text-mining technique used to group similar texts into clusters [60]. Treating each cluster as a topic facilitates the discovery of topics by analyzing common words within each cluster. Text clustering is typically an unsupervised process in which patterns in an unlabeled dataset are identified based on distance functions. The results of this unsupervised learning can then be applied to new unlabeled data to automatically assign documents to clusters, eliminating the need for human experts to categorize the data manually.

Methods of clustering include: (i) partitional clustering, which assigns each data point to one of K clusters; the typical algorithm used is K-means. The main limitation is that it requires determination of the number of clusters and may suffer from outliers. (ii) Hierarchical clustering, which divides documents into a hierarchy using divisive or agglomerative strategies, and may perform poorly with large datasets. (iii) Density-based methods, such as DBSCAN and OPTIC group points by density but struggle with clusters of varying densities. (iv) Hybrid methods, such as HDBSCAN, combine density and hierarchical techniques, making them fast and suitable for large datasets by ignoring sparse regions and forming a cluster tree.

HDBSCAN tackles outliers by ignoring sparse regions and by marking them as noise [61]. A top-down process forms a cluster tree by repeatedly splitting a single large cluster into smaller ones clusters. For this purpose, a hyperparameter denoting the minimum cluster size must be provided. In our study, we worked with HDBSCAN and applied HDBSCAN

to various word representation methods.

5) Conventional Topic Modeling Methods

These techniques aim to model topics from text documents. Conventional topic modeling methods represent documents using the document-term matrix or TF-IDF. Documents are modeled as a mixture of latent topics [62, 63]. Typical methods include Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Process (HDP). The prior versions of the LSA and PLSA have a limitation in handling polysemy problems based on the context of target words. HDP, a method extended from LDA, has the ability to generate latent groups of topics automatically and provides a possibility of new documents triggering the emergence of new topics dynamically, but without altering the dataset shape. This supports Bayesian non-parametric models in adapting many possible topics to create flexibility.

6) Topic Labeling

After generating groups of clusters, our next step was to create a topic representation for each cluster to gain insight. The most intuitive approach is to determine the topic representation by selecting each cluster's most important common tokens. However, simply combining several tokens may result in a representation that lacks semantic meaning and clarity. Therefore, we decided to determine clusters based on the grammatical sense of tokens instead of the frequency of word occurrence.

Topic Representation with Top Tokens

When clustering algorithms are deployed to construct topics, labels can be assigned to those clusters simply by extracting the most common keywords [13]. These keywords were constructed based on the frequent weight occurrences [64]. In particular, we utilize the CTFIDF algorithm, which captures the meaning of texts instead of simply the frequency occurrence of words. As a baseline, we employed the CTFIDF method to identify important words per cluster based on cosine similarity [65]. The CTFIDF method was integrated with HDBSCAN clustering. This involves using the TFIDF procedure to calculate the term frequency within each class c, where c denotes the group of clustered texts. Inverse class frequency measures the weight of the importance of terms in each class. Thus, the CTFIDF generates a list of keywords using cluster-class IDs and documents. We iteratively merged the clusters to reduce their numbers. Traditional models that employ the CTFIDF to generate topics often produce relatively low level tokens, making them somewhat detached from human interpretation. This led us to explore the idea of generating a meaningful short sentence as the topic representation for each cluster, moving beyond traditional methods to enhance the interpretability of the clustering results.

Topic Representation with Sentence

To generate topic representations with more semantic meaning and interpretability, we explored the use of large-language models(LLMs) for topic label generation. Our study used the Llama2-Chat 7B model to generate short sentences that serve as topic labels.

D. METHODS OF TOPIC MODELING

Next, we describe the prevalent techniques for extracting *common semantic topics* inherent in the text with the aim of relating them to key sectors of development. Based on the techniques introduced in Section II, we developed a series of pipelines for extracting and modeling common topics, as detailed in Table 2. These pipelines cover three popular topic modeling methods: text clustering, conventional topic modeling, and neural topic models.

1) Text Clustering based Models

For text clustering-based models, we adopted HDBSCAN as the clustering method, utilizing the following steps. For data representation, we applied and compared the following methods; TF-IDF, Word2Vec-SMA, pretrained Word2Vec, FastText, and BERT (Section II, subsection II-C2). Thereafter, dimension reduction was performed using the UMAP method (Section II, subsection II-C3). Finally, we generated topic labels based on the most frequent top ten tokens of each cluster, and some pipelines used the CTFIDF algorithm to generate these labels.

BERTopic extracts topic representations using a custom class-based variation of TF-IDF [66]. It involves dynamic topic modeling to create topics based on how topics evolve over time. Each document is converted to its embedding representation using a pretrained language model [65]. The BERTopic algorithm reduces the dimensionality of the resulting embeddings to optimize the clustering process.

In our study, while we employed the standard BERTopic pipeline, which inputs the top k tokens determined by CTFIDF to create topic labels, we recognized the need for more interpretable labels. To address this, we designed a custom pipeline that generates semantically meaningful short sentence labels. Specifically, we extended the BERTopic+ pipeline to merge topics to smaller groups by integrating Llama2 for meaningful topic label generation for each new group, an approach conceptually similar to [14], which combines the BERTopic model with Llama2 to generate token word labels. However, our contribution lies in advancing this method by creating a specialized pipeline, termed BERTopic_theme, which utilizes LLMs to generate concise single-word or short phrase labels that better describe the topics. To further refine topic identification related to social development concerns, we designed a specific prompt using BERTopic_theme. This prompt leverages the top k words from CTFIDF and an LLM to generate relevant titles specifically for topics related to social concerns. If the generated topics are unrelated to social development, the prompt returns no theme title, ensuring that only relevant topics are highlighted.

For this study, we set our parameters to BERT embeddings (paraphrase-MiniLM-L6-v2), 15 minimum cluster size, 5 n_components, 15 minimum samples, and 42 random_state. The high-level steps of the BERTopic are presented in Algorithm 1.

Top2Vec is also a text clustering-based method that does not use CTFIDF to create meaningful semantic topics. This model produces jointly embedded topic, document, and word vectors such that the distance between them represents semantic similarity [67]. The number of dense document areas found in the semantic space was assumed to be the number of prominent topics. This model can operate without removing stop words or word normalization by lemmatizing or stemming from learning good topic vectors. Top2Vec is limited by its assumption that topics are based on words near a cluster's centroid, thus ignoring the semantic meaning of the topics. To create topics using Top2Vec, we also set similar parameters to those for BERTopic.

2) Conventional Topic Modeling

Classical methods extract semantic information from a collection of texts based on the conditional probability distribution of words using statistical algorithms [68]. These conventional topic modeling methods represent documents using word representations, where each document is modeled as a mixture of latent topics [62, 63]. Examples include Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Process (HDP). In our study, we use the HDP model due to its advantage in applying to large datasets. Also, we used the Bag-of-Word (BoW) model to create the initial document representations. The topic labels were generated by selecting the ten most common words from documents within the same topic. Although HDP automatically generates a fixed number of optimal topics using the BoW model, these topics have less semantic meanings.

3) Neural Topic Models

To increase semantic meaning of conventional models, neural topic models serve this purpose. These methods incorporate word embeddings into a conventional model to generate latent topics using deep neural networks [63, 69]. Neural topic models incorporating neural components have significantly advanced the modeling results compared to the traditional Latent Dirichlet Allocation (LDA). These include ProdLDA, Scholar, GSMLDA, and NVLDA. NTMs are generally based on a variational autoencoder framework (VAE), which suffers from hyperparameter tuning and computational overheads. Moreover, the integration of pretrained embeddings into the standard VAE framework adds additional model complexity, we thus ignored exploration of these NTMs Recent studies have successfully incorporated contextualized topic embeddings into NTMs for example Contextualized Topic Models. These combine contextualized embeddings to extract topics [70]. Examples of this include CombinedTM and ZeroshotTM. For our study, we used the CombinedTM (CTM), which integrates static and contextualized embedding to discover topics from large datasets [66, 70].

Algorithm 1 BERTopic: Approach for Generating Topics using Llama2

1: **Cluster Initialization:**
$C \leftarrow \text{HDBSCAN}(D)$ ▷ C: Clusters, D: Documents
2: **Document Grouping:**
$G \leftarrow \text{Group}(C)$ ▷ G: Grouped Documents
3: **CTFIDF Calculation:**
$M \leftarrow \text{CTFIDF}(G)$ ▷ M: CTFIDF Matrix
4: **Top Words Extraction:**
$T \leftarrow \text{TopWords}(M)$ ▷ T: Top Words
5: **Prompt Definition:**
$P \leftarrow S + E + Q$ ▷ P: Prompt, S: System, E: Example, Q: Main Prompt
6: **Label Generation:**
$L \leftarrow \text{GenerateLabels}(Q, T)$ ▷ L: Labels

E. EVALUATION OF TOPIC MODELING

1) Unsupervised Evaluation Metrics

For unlabeled datasets, topic coherence and divergence metrics can be used to evaluate the topic labels assigned to each cluster, reflecting their semantic meaning [71]. Topic coherence (TC) measures how understandable a topic is. In our study, we used the Coherence Score(C-v), which uses normalized pointwise mutual information(NPMI) and cosine similarity between the most probable words in a topic to evaluate the quality of the topic [72]. Topic diversity (TD) measures how well topics generated by a topic model are differentiated from each other [73]. We employ the Inverted Rank Based Overlap (IRBO) to evaluate the degree of separation between topics.

F. TOPIC CLASSIFICATION

Our aim was to develop lightweight classifiers using identified topics to infer the semantic labels of unlabeled social media posts. We anticipate a better classifier performance when the quality of the generated topics closely matches the content of the text. We explored multiple classification methods and selected the best models based on standard classification performance metrics.

Dataset split. Our SMA dataset of size 22,036 records was split into 80% (17626) and 20% (4408) training and testing sets. To implement the train and test sets, we fit UMAP on only training data and thereafter apply the transformation learned to the test data to avoid data leakage and inflated performance. We employed stratified sampling to ensure proportional representation based on the number of unique labels in the topics. The labels derived from the top common topics were used as training labels.

Topic Classifiers

We opted for lightweight machine learning methods such as logistic regression, k-nearest neighbor, decision tree, SVM,

TABLE 2. A Summary of model pros and cons and how to extract topics (a) text clustering vs (b) topic modeling

Model Name	Components and Pros	Cons	Topic Labeling
		Classical models	
TF-IDF	(TFIDF-UMAP-HDBSCAN): can track the weight of word importance	Ignores local semantic context	Automatically set to use eom settings of minimum cluster size and samples. Extract most common 10 words
Word2Vec	(Word2Vec-UMAP-HDBSCAN): tracks similarity and word context on any large datasets,	ignores the grammatical sense of polysemous target words	Automatically set to use eom settings of minimum cluster size and samples. Extract most common 10 words
Word2Vec-SMA	(Word2Vec-SMA-UMAP-HDBSCAN): has better tracking of similarity and word context on smaller datasets	Ignores the grammatical sense of polysemous target words	Automatically set to use eom settings of minimum cluster size and samples. Extract most common 10 words
FastText	(FastText-UMAP-HDBSCAN): has improved tracking of similarity, context, and sequence order of words	ignores the grammatical sense of polysemous target words	Automatically set to use eom settings of minimum cluster size and samples. Extract most common 10 words
BERT	(BERT-UMAP-HDBSCAN): captures the grammatical sense of polysemous target words,	works poorly with long length sentences	Automatically set to use eom settings of minimum cluster size and samples. Extract most common 10 words
HDP	(Bow-HDP): automatically generates topics based on a default value	The topics have less semantic meaning	Extract most common 10 words
CTM	(Bow-BERT-CTM): uses neural topic models to generate topics,	Limited number of topics	Extract most common 10 words
		Advanced models	
Top2Vec	(BERT-UMAP-HDBSCAN): a standard model for BERT that produces topics by joint embedding to track semantic similarity,	Assumes words of proximity to have the same group	Use default settings for minimum cluster and size and samples, and extract most common 10 words
BERTopic	(BERT-UMAP-HDBSCAN-CTFIDF): generates token based topics and merges them to topic clusters	It assumes that every document has one topic only	Randomly set a minimum cluster size 15 to create clusters. Use CTFIDF to extract topic token labels, and create short-sentences using Llama2 from these tokens
BERTopic+	(BERT-UMAP-HDBSCAN-CTFIDF+): generates sentence based topics and merges them to topic clusters	It assumes that every document has one topic only	Randomly set a merge range 10 or 20 to deduce topics to classes. Extract token labels for each class using CTFIDF, and use Llama2 to deduce the tokens to short sentences
BERTopic_theme	(BERT-UMAP-HDBSCAN-CTFIDF): generates token based topics and merges them to topic clusters	It assumes that every document has one topic only	Randomly set a minimum cluster size 15 to create clusters. Use CTFIDF to extract topic token labels, and create single word themes based on social concerns using Llama2

Gaussian NB, Random Forest, MLP, and XGB. We chose these basic classifiers for integration with ensemble classifier algorithms, specifically voting and stacking techniques[74].

Ensemble voting involves training multiple models independently on the same dataset. Finally, a final prediction was made by combining the individual predictions. The final prediction was based on the majority vote of the models (hard voting) or by selecting the model with the highest average probability (soft voting).

Ensemble stacking involves training all models first and then feeding the classification decisions of all base classifiers to the stacked classifier on top to predict the final output based on their performance. For ensemble stacking, we used the XGB classifier as the top-stacked model.

Evaluation of the Topic Classifiers.

We used the top ten token words as labels to train the ensemble classifiers and evaluate the performance of models using of the traditional classification metrics, including accuracy, precision, recall, and F1-score [75].

III. RESULTS ON THE SOCIAL MEDIA AFRICA CORPUS

In Table 3, we observe that topics created with TF-IDF embedding have the highest topic coherence of 0.85 C-v. The general topic divergence is fair because the dataset originally contained user comments closely related to African social concerns. The Top2Vec model had the highest divergence in topics at 0.99 IRBO, but this model generates topics with less semantic meaning and hence its poor performance. For topic classification, ensemble stacking occasionally performed slightly better than stacking. We observed a generally poor performance with prior pipelines (TF-IDF, Word2Vec-SMA, Word2Vec, FastText, and BERT) that used the most common words as labels for these topics. This positively differs when the CTFIDF algorithm is used. The BERTopic pipelines outperformed than the previous methods because

the word embeddings have semantic meaning. Although FastText generates topics with a high performance, the embeddings lack grammatical meaning because they are static context embeddings.

In our study, the performance of the BERTopic pipelines is superior to that of the other models in terms of the potential to build classifiers with high performance. Moreover, these pipelines generate many topics that can be reduced to any suitable small number. The BERTopic yielded 304 topics at 0.81 C-v coherence, 0.58 IRBO divergence, 0.83 F1, 0.82 recall, 0.84 precision, and 0.83 accuracy scores. For training on topic classes, BERTopic+ used the CTFIDF algorithm to merge the topics into 11 topic classes, at 0.76 C-v coherence, 0.43 IRBO divergence, 0.97 F1, 0.97 recall, 0.97 precision, and 0.97 accuracy scores. We observe a smaller standard deviation($\pm$0.02) with BERTopic+ which conforms to a 0.95 confidence for the classification results. In addition, the BERTopic pipelines that use CTFIDF algorithm, have competitive topic coherence and the classifiers perform better with these topic labels. We also note that the poor performance of IRBO topic divergence for BERTopic occurs because these topics are closely related to social development issues. The best topic modeling algorithms generated topics that were closely related. Using prompt generation with Llama2, on topics by BERTopic we refined the top ten words to derive a general theme related to social concerns about development. We attained 98 themes, at coherence 0.75 C-V and divergence 0.50 IRBO, and with 0.93 F1 score, 0.93 precision, 0.93 recall and 0.93 accuracy scores.

For the conventional probability based models, we use HDP. As discussed in (Sections II, II-D, II-D2), these models use features extracted by the BoW model. Based on the experiments, HDP has the highest topic divergence score of 1.00 IRBO. However, this model generates tokens with less semantic meaning and thus poor topic coherence and classification of these labels in the text.

For the Neural topic-based models, we use CTM. As discussed in (Sections II, II-D, II-D3) CTM uses a combination of the BoW model and contextualized embeddings (BERT). From our findings, the CTM model exhibits poor performance and yields a default number of token-based topics.

IV. DISCUSSION

A. ANALYSIS OF EXTRACTED TOPICS

1) Topics extracted by Comparison Methods

Our goal was to obtain topic labels with semantic meaning. As discussed earlier, exploring topic modeling approaches is a key strategy for achieving this. Our experiments showed that extracting topics through text clustering-based models performed significantly better than other approaches. The use of contextualized embeddings, such as BERT, significantly outperforms previous methods of static word representations and word embeddings. Topics extracted are in the form of word tokens, and in the study, we demonstrated the extraction of short-sentence topic labels using Llama2. We achieved this using BERTopic pipelines. We extracted short-sentence labels for both topics and their classes using Llama2.

In Figure 2, we display the sample frequency count number of the top 15 common topics extracted using the best text clustering approach, BERTopic. Generally, these topics are closely related to social development issues. They include discussions such as the impact of climate change, corrupt government leadership, unemployment, poverty crisis, Covid Pandemic, youth empowerment, visa applications for migrations, reactions and opinions to policies, economic impact and challenges, political discourse and engagement, spiritual support and studies, business action plans, human rights and abuse, and entertainment. We will further discuss the categories within these topics extracted by BERTopic.

First, the topic "Visa Application Process in Australia for African Nationals" describes comments on the impact of Covid-19 and visa application processes. A sample of comments under this topic cluster includes: (i) 'Social Work Month 2021 Observance, Brooke Army Medical Center.' (ii) 'Prof. Karim. 1. He talked about the trends of COVID-19. Where the epidemic occurred, how the COVID-19 travelled to South Africa.' Furthermore, Vearey et al. [76] explained that during the COVID-19 pandemic, many movements in different countries increased the search for social, health, and political responses to sustain economies during the pandemic.

Second, the "Corruption and Governance in Developing Countries" topic includes both negative comments from individuals with challenges and some proposed strategies for African leadership. For instance, comments such as: (i) "@everest_254 @ian_g3 @NjeriBt Yes. There is no pure capitalist or Socialist or communism. The question is which predominate. In Africa, there's hyena type of capitalism since IMF/WB imposed conditions decades ago. No welfare; cut social spending; govt agencies privatized etc. It's survival of the fittest." (ii) "Corruption undermines political; social; and economic development; which is a growing concern in fragile economies in parts of Africa." - @Lola_Kanye https://t.co/VR9tZMsCv5 via @BizAfricaDaily." This category of issues was repeatedly reported by the United Nations in 2020 [77]. This highlights that the common problems in Africa included largely depend on political instabilities and crime.

Third, the topic "Entrepreneurial Innovation Challenge" highlights the suggested economic strategies to solve problems in Africa. The comments in this topic cluster include: (i) 'We believe in building strong partnerships so that we may collaboratively drive social change for an alcohol-harm free South Africa. Why? Because the problem is huge; complex and multifaceted, and; the truth is; we just cannot do it on our own.' (ii) '9/17/2021 Are Corporate 'Win-Win' Strategies an Effective Way of Alleviating Social...Problems? by Rutgers Institute for Corporate Social Innovation.'

2) Analysis between Topics and their Classes

After collecting topics with the BERTopic pipelines, we merged them into smaller groups using the CTFIDF algo-

TABLE 3. Comparative study of models of Topic Modeling

Model Name	Topics	Metrics		Ensemble voting				Ensemble Stack			
		C-v↑	IRBO ↑	Acc ↑	F1 ↑	Prec ↑	Rec ↑	Acc ↑	F1 ↑	Prec ↑	Rec ↑
Classical models											
TF-IDF	31	**0.85**	0.46	0.56	0.40	0.31	0.56	0.56	0.40	0.33	0.56
std ↓				±.007	±.008	±.008	±.007	±.007	±.009	±.015	±.007
Word2Vec-SMA	18	0.80	0.53	0.38	0.31	0.27	0.38	0.39	0.30	0.27	0.39
std ↓				±.009	±.010	±.013	±.010	±.011	±.012	±.017	±.012
Word2Vec	12	0.83	0.34	0.72	0.62	0.61	0.72	0.73	0.64	0.62	0.73
std ↓				±.005	±.007	±.006	±.005	±.006	±.007	±.012	±.006
FastText	20	0.83	0.47	**0.94**	**0.94**	**0.94**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**
std ↓				±.006	±.006	±.006	±.006	±.005	±.005	±.004	±.005
BERT	7	0.82	0.37	0.41	0.33	0.31	0.41	0.45	0.32	0.30	0.45
std ↓				±.008	±.009	±.011	±.008	±.007	±.008	±.008	±.007
HDP	10	0.33	**1.00**	0.64	0.61	0.63	0.64	0.63	0.60	0.61	0.63
std ↓				±.005	±.006	±.011	±.005	±.002	±.002	±.008	±.002
CTM	10	0.32	0.98	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10
std ↓				±.006	±.006	±.006	±.006	±.005	±.006	±.006	±.005
Advanced models											
Top2Vec	288	0.30	0.99	0.27	0.27	0.28	0.27	0.25	0.26	0.31	0.25
std ↓				±.009	±.009	±.008	±.009	±.008	±.008	±.010	±.008
BERTopic	**304**	0.81	0.58	0.83	0.82	0.84	0.83	0.78	0.78	0.80	0.78
std ↓				±.004	±.005	±.005	±.004	±.007	±.007	±.005	±.007
BERTopic+	11	0.76	0.43	**0.94**	**0.94**	**0.94**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**
std ↓				**±.002**	**±.003**	**±.002**	**±.002**	**±.002**	**±.002**	**±.002**	**±.002**
BERTopic_theme	98	0.75	0.50	0.93	0.93	0.93	0.93	0.91	0.91	0.92	0.91
std ↓				±.005	±.005	±.005	±.005	±.005	±.006	±.006	±.006

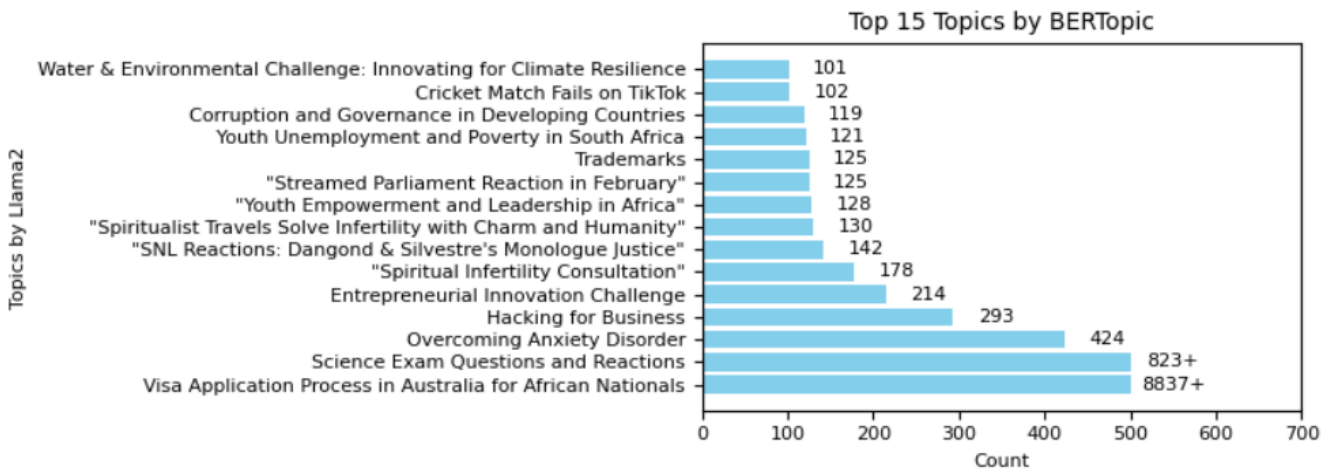


FIGURE 2. Percentage frequency count of topics extracted by BERTopic

rithm. In this study, we displayed the topic classes gathered using the BERTopic+ model. Labeling of topic classes was performed using Lllama2. These consisted of the top common token words with the highest weight importance for each group of topics. Figure 3 shows the results of the merged topic classes or clusters. These include: "African countriesĆOVID-19 travel restrictions and requirements," "Nicki Minaj Reactions at Davos and Bidenś Daily Documentary," "Entrepreneurial Innovation in Energy & Environmental Sustainability," "Spiritual Infertility Consultation," "Reactions to Godsentś Anxiety-Inducing Video Confusion," "Spiritual Journey Through Infertility," "Tillerson-Trump controversy involving Balvin, NBC, Tillersons, TragediaTrademark, Mill-Trademarks, Meek, FaceFace, and Safari," "Pitbulls in Washington: Shouting Match between Schumer, Pelosi, Alicia Había, and Comptes," "Chinaś Rise in Global Politics: Prosperity, Influence, and Reactions," "Robotics and Technology."

For example, we display the top 10 common topics in one class labeled "African countriesĆOVID-19 travel restrictions and requirements" in Figure 3. (b). This topic is highlighted by the red overlay box in Figure 3.(a). We further explore different topics inside this category that highlight challenges and strategies proposed to address challenges in Africa. This cluster encompasses several dominant topics, including visa application processing cases and immigration, social class

studies and exams, impact of technology under business sector, movements for youth empowerment, public engagement to political issues, trends of market products, political instabilities and elections.

3) Analysis between Topics and General Theme Titles

Our analysis of topics and their general themes reveals that certain themes dominate social media discussions in Africa. "Poverty" leads significantly with 15,391 instances, highlighting widespread concern over economic hardships. Other key themes include "empowerment" (1,230) and "health" (541), emphasizing social and economic uplift and public health issues, "Security" (389) and "stewardship" (293) reflect concerns about safety and responsible governance. Environmental themes like "sustainability" (168) also appear relatively frequently. Cultural and social issues, such as "youth empowerment" (131) and "social justice" (105), point to ongoing discussions about equity and the role of youth in societal change.

In summary, poverty, empowerment, and health are the most prevalent themes, reflecting a complex landscape of social concerns that require diverse approaches for effective intervention. Figure 5 displays key discussions across three themes: poverty, empowerment, and health. In the poverty theme, topics such as the visa application process for African nationals, youth unemployment in South Africa, and corruption dominate, reflecting a mix of regional and global socioeconomic concerns. The empowerment theme centers on entrepreneurial innovation and spiritual journeys, with cultural aspects like Afro-beats Mix showing how empowerment is expressed through both personal and cultural growth. The health theme focuses heavily on mental health, with discussions on anxiety disorders, bipolar disorder, and public health topics like vaccine confusion during pregnancy. This highlights the importance of mental and public health awareness. Together, these themes capture a wide range of issues that resonate globally, affecting both individual and collective well-being.

B. FINDINGS ON STUDY OBJECTIVE

In this study, we achieved the following objectives. First, we discuss the collection of social comments on concerns in Africa from social media sites (SNPs). Second, we compared different machine learning (ML) algorithms to extract common topics of discussion related to African development from these corpora of social media posts. Finally, we evaluated the generated topics and trained classifiers to automatically predict topic labels for social media posts. These classifiers can be employed in the future to predict topics in new datasets. We also display a sample of common topics retrieved by our methods related to key sectors under development.

For a recap, we initially collected historical social comments from YouTube using web scrapers and the YouTube API, for Twitter (X), we used Snscrape. We integrated the datasets into one database and cleaned them to 22,036 records. We compared the different feature extraction methods of TF-IDF, Word2Vec-SMA, Word2Vec, FastText and BERT. To create common topics for discussion, we compared different models for topic modeling (text clustering, conventional probability, and neural topic models).

C. FINDINGS FROM RELATED STUDIES

We aggregated findings from different related studies to compare our findings. A study by [78] used a much smaller dataset of 2472 records to extract topics. Their study was restricted to the text clustering method. They compared vectorizers such as TFIDF, Word2Vec, FastText, and BERT, to extract clusters using HDBSCAN and UMAP. They obtained 89 clusters using FastText, HDBSCAN, and UMAP, with the highest accuracy of 0.92. Although their study suggested using BERT embeddings, they ultimately recommended using FastText to generate the topic clusters. In our study, we observed that BER Topic-based solutions achieved better performance and, more importantly, that the extraction of topics could be aided by assigning meaningful short sentences as topic labels. This would aid their future use by stakeholders who may want to analyze the results and utilize them for specific purposes. Although Fast Text yields a high topic identification and classification performance, it ignores the contextual meaning of embeddings.

The study by [79] compared Word2vec and contextual word embeddings using BERT. Their study focused only on topic modeling methods. They created clusters and topic representations using the CTFIDF. The ensemble stacking classifier had the highest prediction performance, with 0.98 accuracy and 0.97 precision and recall. From their study, we borrowed the idea of using BERT as contextualized embedding instead of traditional feature extraction methods. In addition to their work, we explored not only topic modeling but also clustering, and our comparative best models can attain semantically meaningful tokens at higher topic coherence by using BERT embeddings. By introducing LLMs (in particular, Llama2), we deduce the tokens into short-sentence topic labels in the BERTopic model.

A study by [14] proposed combining the BERTopic algorithm and LLMs (Llama2) to extract topics and topic classes. Their study was restricted to topic modeling. They used a smaller dataset of 2,472 records to extract topics using Prompt WSM_Llama with a divergence score of 0.97 IRBO and 0.04 NPMI topic coherence. Their study extracted token words as labels, which are typically not as easily interpretable by humans as our proposed short sentences generated based on these top meaningful tokens. In our study, we compared the performance of BERTopic+ in generating short-sentence topics using Llama2 for topic classes. With a bigger dataset (22,036), the BERTopic+ approach achieved 11 topics at topic coherence 0.76 C-v, 0.27 NPMI, topic divergence 0.43, and classifier accuracy 0.95 and 0.95 F1 score. Thus, our experiments revealed better performance with short-sentence labels.

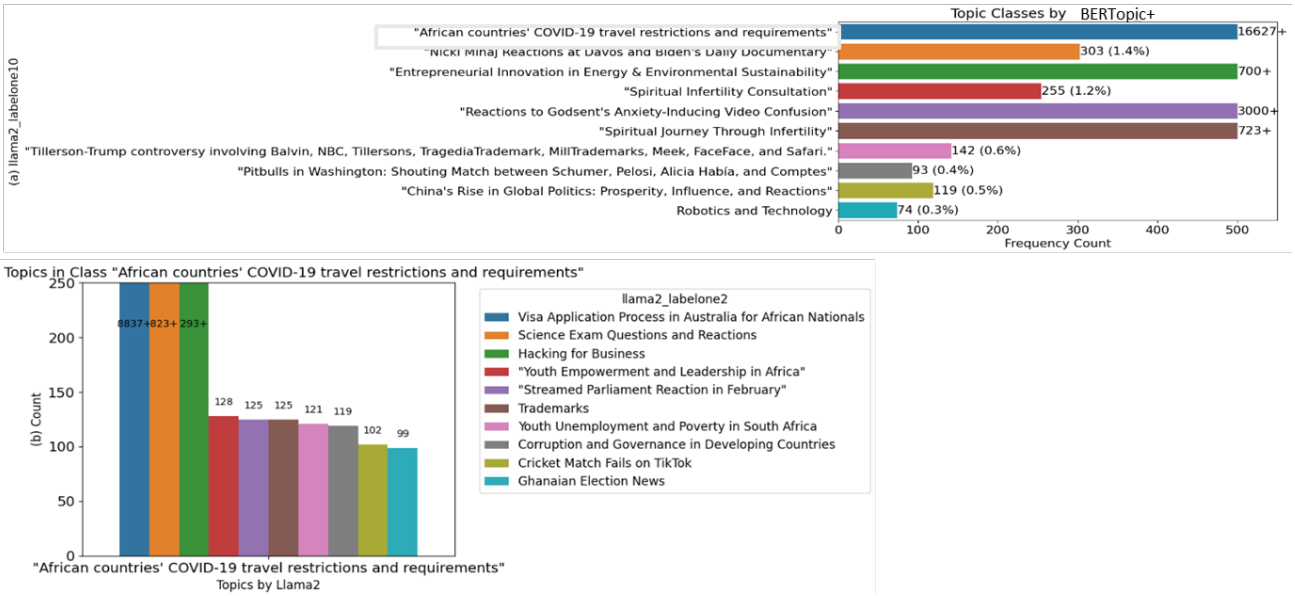


FIGURE 3. **Display for the frequent count for topic classes by BERTopic+ in 3.(a) and frequency count of topics within a specific class in 3.(b)**

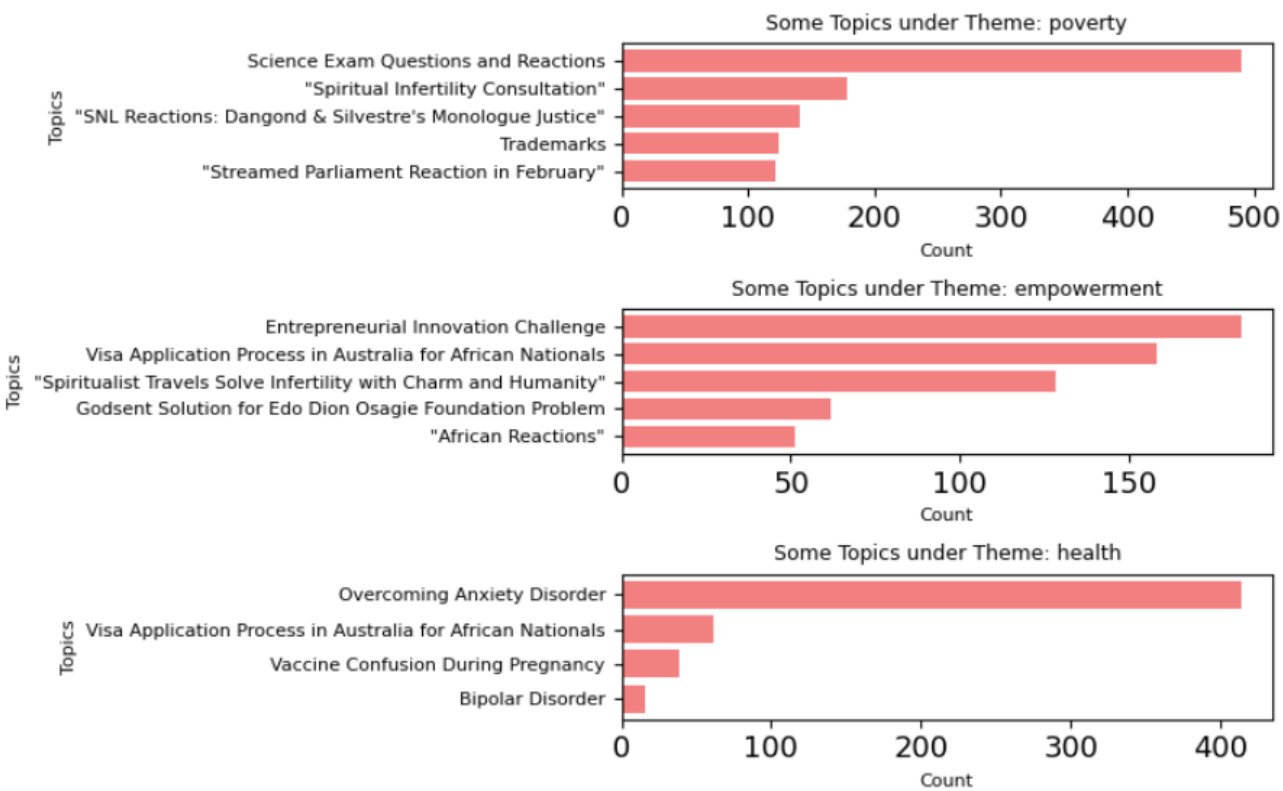


FIGURE 4. **Display for top three themes about social concerns for development and some topics**

D. LIMITATIONS OF OUR STUDY

We followed the literature in that we leveraged metrics such as topic coherence (C-v) and divergence (IRBO) to evaluate the quality of the extracted topics, along with our case study. This is done because the data sets, both ours and those used in related studies, are unlabeled data sources. In the future, it may be interesting to conduct a user study of either human label subsets of the extracted dataset and/or to discern the deeper semantic meaning of labels by humans. This may help us determine the degree to which the identified topics are

considered important and relevant. Furthermore, a case study may be an interesting next step to explore the utilization of the extracted topics in practical objectives.

Second, LLMs, such as the Llama2 model, require high computational resources and powerful GPUs. To support this advancement, we conducted LLM-related experiments using an A100 GPU. However, we acknowledge that governments and non-profit agencies may not have access to these computing resources when applying our proposed methods to their organizational goals at this pivotal moment of the disruptive AI revolution.

V. CONCLUSION

In conclusion, our study aims to achieve two main objectives: first, to develop models to extract common social concerns about Africa from social media sites, and second, to create a classifier that infers the particular topics of concern for a given social media text. We collected and analyzed a text corpus Social Media in Africa (SMA) consisting of 22,036 records extracted from Twitter (X) and YouTube comments in Africa. We followed three steps to achieve the first goal: we initially utilized feature extraction techniques, such as BERT and Word2Vec models, to generate text embeddings for representation. Next, we compared a wide variety of models that fall into one of two core strategies, text clustering, and topic modeling, to organize features into meaningful topic groups. Finally, we conduct a comparative study of alternate methods to label these groups, including generated short sentences as topic labels using the CTFIDF algorithm and LLMs (Llama2).

To meet the second goal, we trained classifiers using ensemble voting and stacking techniques to efficiently infer the existence of identified topics in particular individual social media posts. The text clustering-based pipelines of BERTopic are the best pipelines for topic identification. BERTopic generated 304 topics at 0.81 C-v topic coherence and topic divergence at 0.58 IRBO. We consolidated the topics into classes using the BERTopic+ approach and achieved a topic coherence of 0.76 C-v and a topic divergence of 0.43 IRBO. For single theme extraction from topics, the BERTopic_theme attained 98 general theme titles at 0.75 C-v in coherence and at 0.50 IRBO in divergence.

Thereafter, we assigned topic labels to social media posts for supervised training. For topic classification, BERTopic achieved a 0.82 F1 score and 0.83 accuracy with ensemble voting. Generally, using BERTopic pipelines is ideal for collecting many topics automatically because we had more topics at higher performance, and these were merged into smaller topic classes. The BERTopic+ is an ideal method for attaining smaller clusters within text because we attained 11 topics with higher 0.95 F1 score and 0.95 accuracy for ensemble stacking compared to other models of topic modeling. The BERTopic_theme with ensemble voting also achieved 0.93 F1 score and 0.93 accuracy.

Hence, for Africa, policymakers must prioritize tackling the most urgent social issues, such as the effects of COVID-19 restrictions on public health, economic recovery, and international travel. Encouraging entrepreneurial innovation in energy and environmental sustainability is essential for addressing climate change and promoting sustainable development with green technologies. Lastly, China's increasing influence in global politics necessitates strategic actions to ensure geopolitical stability and enhance international cooperation. Our analysis of topic correlations with general themes reveals that "poverty" is the most dominant concern on African social media, followed by "empowerment," "health," "security," and "sustainability." These themes reflect the key socio-economic and environmental issues discussed online. This highlights the need for targeted policies to address these interconnected challenges for sustainable development in Africa. It is of particular interest to note that our trained classifier model can now be deployed to infer insights in the form of common topics for other social media platforms collected in the future to observe trends over time.

Looking ahead, our study reveals the potential for extracting common topics of discussion as sentences with semantic meanings using modern LLM models. In the future, it would be interesting to analyze correlations within social media between positive and negative emotions related to these topics, as well as the key entities involved. This study may have the potential to identify and track the evolution of common social issues that the community believes hinder development.

ACKNOWLEDGMENT

We express our gratitude to all those who contributed to the completion of this study. We are also thankful to PASET-RSIF for their financial support, without which this research would not have been possible.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest. The funders had no role in the study design; collection, analyses, interpretation of data, writing of the manuscript or decision to publish the results. We provided public access to the codes for data collection and a sample dataset from the public GitHub repository (https://github.com/SibitendaHarriet/Data_collection_common_topics)

REFERENCES

[1] Harold R Kerbo and James W Coleman. Social Problems. *The Social Science Encyclopedia*, pages 949–951, July 2018.

[2] Abolaji Adewale Obileye and Richard Aborisade. Chapter Twenty Social Problem. *Introduction to Sociology: African Culture, Context and Complexity*, page 12, June 2020.

[3] Albert G. Zeufack, Cesar Calderon, Alaine Kabundi, Dhushyanth Raju, Megumi Kubota, Vijdan Korman, Kaleb Girma Abreha, Woubet Kassa, and Solomon Owusu. *Africa's Pulse, An Analysis of Issues Shaping Africa's Economic Future*, volume 25. April 2022.

[4] Shah Fahad Khyber, Shah Faisal Khyber, Naushad Khan, and Shah Fahad. Review of Social Problems in the World. *Researchgate.Net*, January 2021.
[5] Buhner Martin. Sociological Research Methods. *Sociological Research Methods*, pages 36–67, 2017.
[6] ILO. *Africa Regional Social Protection Strategy, 2021-2025*. pages 1–38, 2021.
[7] *OECD (2017), Social Protection in East Africa: Harnessing the Future, OECD Publishing, Paris*. pages 1–98, 2017.
[8] Cristina M. Pulido, Laura Ruiz-Eugenio, Gisela Redondo-Sama, and Beatriz Villarejo-Carballido. A new application of social impact in social media for overcoming fake news in health. *International Journal of Environmental Research and Public Health*, 17(7):pages 1–15, 2020.
[9] Lillian Gray. Gender Bias Detection Using Facebook Reactions. pages 1–6, 2020.
[10] S. Blasi, E. Gobbo, and S. R. Sedita. Smart cities and citizen engagement: Evidence from Twitter data analysis on Italian municipalities. *Journal of Urban Management*, 11(2):pages 153–165, 2022.
[11] Thien Binh and Hoang Dao. Job Clustering : An unsupervised approach for a recommender system of skill requirements. pages 1–28, June, 2021.
[12] Christian Vestergaard and Arne Morten Kästel. Comparing performance of K- Means and DBSCAN on customer support queries (Bsc Thesis). page 12, 2019.
[13] Hanan Qassim Jaleel, Jane Jaleel Stephan, and Sinan A. Naji. Products dataset analysis using data mining techniques. *Journal of Engineering Science and Technology*, 16(5):pages 3880–3906, 2021.
[14] Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka Wei Lee. Prompting Large Language Models for Topic Modeling. *Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023*, pages 1236–1241, 2023.
[15] Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka Wei Lee. PromptMTopic: Unsupervised Multimodal Topic Modeling of Memes using Large Language Models. *MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia*, pages 621–631, 2023.
[16] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):pages 37–53, 1996.
[17] Gregory Piatetsky-Shapiro, Ron Brachman, Tom Khabaza, Willi Kloesgen, and Evangelos Simoudis. An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 89–95, 1996.
[18] Statcounter. Social Media Stats Africa | StatCounter Global Stats, https://gs.statcounter.com/social-media-stats/all/africa, 2021, accessed on 28th/June/2022.
[19] The Internet Society Pledges to Expand Internet Access in Africa, 2022, accessed on Febraury/9/2023, https://www.internetsociety.org/news/press-releases/2022/the-internet-society-pledges-to-expand-internet-access-in-africa/.
[20] Mariam Saleh. Internet usage in Africa - statistics facts, 2022, https://www.statista.com/topics/9813/internet-usage-in-africa/topicOverview.
[21] International Telecommunication Union. World Telecommunication/ ICT Development Report and database. Individuals using the Internet (% of population) | Data, 2018, accessed on Febraury/9/2023, https://data.worldbank.org/indicator/IT.NET.USER.ZS.
[22] Frank W. Geels and Caetano C.R. Penna. Societal problems and industry reorientation: Elaborating the Dialectic Issue LifeCycle (DILC) model and a case study of car safety in the USA (1900-1995). *Research Policy*, 44(1):pages 67–82, 2015.
[23] Social Concern synonyms - 21 Words and Phrases for Social Concern, accessed on March/6/2023, https://www.powerthesaurus.org/social_concern/synonyms.
[24] Harriet Sibitenda. Retrieving Data from Social Network Platforms : A State-of-art Review. (i):pages 1–24, 2023.
[25] David Moher, Alessandro Liberatl, Jennifer Tetzlaff, Douglas G. Alttman, and PARISMA Group. So schaffst du deine Ausbildung. Ausbildungsbegleitende Hilfen (abH). 151(4):pages 264–269, 2009.
[26] Chris Stokel-Walker. Why is Twitter becoming X? *New Scientist*, 259(3449):Page 9, 2023.
[27] Pamela E. Walck. BOOK REVIEW: Social Communication in the Twitter Age. *International Journal of Interactive Communication Systems and Technologies*, 3(2):pages 66–69, 2013.
[28] Axel Bruns and Stefan Stieglitz. Twitter data: What do they represent? *IT - Information Technology*, 56(5):pages 240–245, 2014.
[29] SNScrape: A social networking service scraper in Python, 2021, accessed on March/6/2023, https://github.com/JustAnotherArchivist/snscrape.
[30] GCFLearnFree. What is YouTube?, 2020, accessed on March/6/2023,https://edu.gcfglobal.org/en/youtube/what-is-youtube/1/.
[31] Vineeth Nair. *Getting Started with Beautiful Soup*. pages 1-113, 2014.
[32] Nisha Gogna. Study of Browser Based Automated Test Tools WATIR and Selenium. *International Journal of Information and Education Technology*, 4(4):pages 336–339, 2014.
[33] Antawan Holmes and Marc Kellogg. Automating functional tests using selenium. *Proceedings - AGILE Conference, 2006*, 2006:pages 270–275, 2006.
[34] Renu Patil and Rohini Temkar. Intelligent Testing Tool : Selenium Web Driver. *International Research Journal of Engineering and Technology(IRJET)*, 4(6):pages 1920–1923, 2017.

[35] Fredrik Hallström and David Adolfsson. Data Cleaning Extension on IoT Gateway: An Extended ThingsBoard Gateway. pages 1–44, 2021.

[36] Kashshaf Labeeb, Kuraish Bin Quader Chowdhury, Rabea Basri Riha, Mohammad Zoynul Abedin, Sarmila Yesmin, and Mohammad Nasfikur Rahman Khan. Pre-Processing Data in Weather Monitoring Application by Using Big Data Quality Framework. *Proceedings of 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, WIECON-ECE 2020*, (April):pages 284–287, 2020.

[37] Wie Hsing Li and Wie Hsing. Detecting Non-Credible News Using Machine Learning. pages 1–27, 2018.

[38] Arvind Kumar Gautam and Abhishek Bansal. Performance Analysis of Supervised Machine Learning Techniques for Cyberstalking Detection in Social Media. *Journal of Theoretical and Applied Information Technology*, 100(2):pages 449–461, 2022.

[39] Pingchuan Wen and Xianbiao Li. Research on User Demand Based on E-commerce Consumer Negative Reviews. 1(04):pages 45–51, 2019.

[40] Peya Mowar, Mini Jain, Ruchika Goel, and Dinesh Kumar Vishwakarma. Clickbait in YouTube Prevention, Detection and Analysis of the Bait using Ensemble Learning. pages 1–26, 2021.

[41] Lifeng Li, Wenxing Li, and Daqing Gong. Naive bayesian automatic classification of railway service complaint text based on eigenvalue extraction. *Tehnicki Vjesnik*, 26(3):pages 778–785, 2019.

[42] Eivind Strøm. Multi-label style change detection by solving a binary classification problem. *CEUR Workshop Proceedings*, 2936:pages 2146–2157, 2021.

[43] Varun Dogra, Sahil Verma, Kavita, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience*, 2022:pages 1–21, 2022.

[44] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 55–65, 2019.

[45] Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. From static to dynamic word representations : a survey. *International Journal of Machine Learning and Cybernetics*, 11(7):pages 1611–1630, 2020.

[46] Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. Afaan Oromo Hate Speech Detection and Classification on Social Media. *2022 Language Resources and Evaluation Conference, LREC 2022*, (June):pages 6612–6619, 2022.

[47] Chunyu Jiao. Review of Artificial Intelligence Classification of Short Texts in Weibo Based on BERT Model. 2:pages 1–12, 2022.

[48] Amgad Muneer and Suliman Mohamed Fati. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11):pages 1–21, 2020.

[49] Oguzhan Gencoglu. Deep Representation Learning for Clustering of Health Tweets. pages 1–8, 2018.

[50] B Wang, M Liang, and A Li. Cross-media Scientific Research Achievements Query based on Ranking Learning. *arXiv preprint arXiv:2204.12121*, pages 1–7, 2022.

[51] Vishal Anand, Ravi Shukla, Ashwani Gupta, and Abhishek Kumar. Customized video filtering on YouTube. pages 1–13, 2019.

[52] Lingfei Wu, Ian En Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. Word mover's embedding: From word2vec to document embedding. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4524–4534, 2018.

[53] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:pages 139–147, 2019.

[54] Erjon Skenderi, Jukka Huhtamäki, and Kostas Stefanidis. Multi-Keyword Classification: A Case Study in Finnish Social Sciences Data Archive. *Information (Switzerland)*, 12(12):pages 1–19, 2021.

[55] Valentin Hofmann and Janet B Pierrehumbert. arXiv : 2010 . 12684v3 [cs . CL] 8 Jun 2021 d. 2018.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019*, (Mlm):pages 4171–4186, 2018.

[57] Leland McInnes. Performance Comparison of Dimension Reduction Implementations, 2018, accessed on June/21/2023, https://umap-learn.readthedocs.io/en/latest/benchmarking.html.

[58] Tina Smets, Nico Verbeeck, Marc Claesen, Arndt Asperger, Gerard Griffioen, Thomas Tousseyn, Wim Waelput, Etienne Waelkens, and Bart De Moor. Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data. *Analytical Chemistry*, 91(9):pages 5706–5714, 2019.

[59] Deepak Upreti and Hyunil Kim. Defending against Label-Flipping Attacks in Federated Learning Systems with UMAP Defending against Label-Flipping Attacks in Federated Learning Systems with UMAP. pages 0–13, 2022.

[60] Meghana Santoshi Janapareddy, Nirmala Paul Nirujogi, Nandini Prasada, and R Sree Meghana. Article Clustering Using A Normalised Semantic Data Representation. 11(4):pages 147–154, 2020.

[61] Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-

Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung Sup Kwak. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, 174(March):pages 27–42, 2019.

[62] D Yamunathangam. An Overview of Topic Representation and Topic Modelling Methods for Short Texts and Long Corpus. *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pages 1–6, 2021.

[63] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):1–30, 2024.

[64] Mohamed Mahyoub, Jade Hind, David Woods, Carl Wong, Abir Hussain, and Dhiya Aljumeily. Hierarchical text clustering and categorisation using a semi-supervised framework. *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, October-2019:pages 153–159, 2019.

[65] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. pages 1–10, 2022.

[66] Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi-Rad. Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 3886–3893, 2022.

[67] Dimo Angelov. Top2Vec: Distributed Representations of Topics. pages 1–25, 2020.

[68] Xieling Chen, Ziqing Liu, Li Wei, Jun Yan, Tianyong Hao, and Ruoyao Ding. A comparative quantitative study of utilizing artificial intelligence on electronic health records in the USA and China during 2008 – 2017. 18(Suppl 5):pages 1–15, 2018.

[69] Olga Kellert and Md Mahmud Uz Zaman. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. *LChange 2022 - 3rd International Workshop on Computational Approaches to Historical Language Change 2022, Proceedings of the Workshop*, (April 2020):131–139, 2022.

[70] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pretraining is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2:pages 759–766, 2021.

[71] Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1728–1736, 2020.

[72] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445, 2022.

[73] Silvia Terragni, Elisabetta Fersini, Bruno Galuzzi, Pietro Tropeano, and Antonio Candelieri. OCTIS: Comparing and optimizing topic models is simple! *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*, pages 263–270, 2021.

[74] Stamatios-aggelos N Alexandropoulos, Christos Aridas, Sotiris Kotsiantis, Michael N Vrahatis, Stamatios-aggelos N Alexandropoulos, Christos Aridas, Sotiris Kotsiantis, Michael N Vrahatis, Stamatios-aggelos N Alexandropoulos, Christos K Aridas, and B Sotiris. Stacking Strong Ensembles of Classifiers To cite this version : HAL Id : hal-02331304 Stacking strong ensembles of classi ers. pages 0–12, 2019.

[75] Alaa Tharwat. Classification assessment methods. 17(1):168–192, 2021.

[76] Jo Vearey, Thea de Gruchy, and Nicholas Maple. Global health (security), immigration governance and Covid-19 in South(ern) Africa: An evolving research agenda. *Journal of Migration and Health*, 3(March):page 100040, 2021.

[77] United Nations. Africa | United Nations, 2020, accessed by April/27/2024, https://www.un.org/en/sections/issues-depth/africa/index.html.

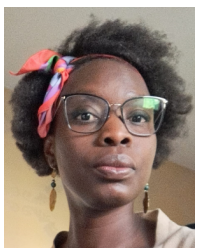
[78] Muhammad Sidik Asyaky and Rila Mandala. Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP. *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, pages 1–6, 2021.

[79] Fatima Alhaj, Ali Al-Haj, Ahmad Sharieh, and Riad Jabri. Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic. *International Journal of Advanced Computer Science and Applications*, 13(1):pages 854–860, 2022.

HARRIET SIBITENDA received a B.S. in Education (Mathematics, Computer Science) from Busitema University, Uganda. She attained an M.S. degree in Information Systems from Uganda Martyrs University. Currently, she is a doctorate pursuing her PhD in Computer Science at the University of Gaston Berger, Senegal. She is a visiting scholar with an exchange program in Data Science at Worcester Polytechnic Institute, Worcester, MA, USA.

AWA DIATTARA , an Associate Professor at Computer Science, University of Gaston Berger, St. Louis, Senegal. She works with the Computer Science Department.

CHEIKH BA , an Associate Professor at Computer Science, University of Gaston Berger, St. Louis, Senegal, since 2012. He is the head of the Computer Science Department.



ASSITAN TRAORE , She works at Actroll, France, in collaboration with the University of Gaston Berger. She affiliates to the Computer Science Department

RUOFAN HU , a PhD student at Data Science, Worcester Polytechnic Institute. Worcester, MA, USA. Education: Southwestern University of Finance and Economics, Bachelor of Science in Statistics, 2018 Worcester Polytechnic Institute, Master of Science in Data Science, 2020. Her advisor is Professor Rundensteiner and she is a member of Daisy lab. Her research interests include self-supervised learning, weakly supervised learning, and natural language processing.

DONGYU ZHANG , a PhD student at Data Science, Worcester Polytechnic Institute. Worcester, MA, USA. Education: He received his Master's degree in Business Analytics from the University of Texas at Dallas. He is now a Ph.D. student majoring in Data Science at WPI, working with Prof. Elke Rundensteiner. His research interests include natural language processing, model interpretability, multi-task learning, and weakly supervised learning.

ELKE RUNDENSTEINER , a Professor of Computer Science, the William Smith Professor in Computer Science, Worcester Polytechnic Institute. Worcester, MA, USA. She serves as the Founding Head of the Data Science Program at WPI. Education: BS Computer Science J.W. Goethe University, Frankfurt, Germany MS Computer Science Computer Science, Florida State University, Tallahassee, Florida. PhD Computer Science University of California, Irvine, California