IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 May 2024, accepted 29 June 2024, date of publication 10 July 2024, date of current version 18 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3426329

RESEARCH ARTICLE

Uncovering Concerns of Citizens Through Machine Learning and Social Network Sentiment Analysis

SANDRA KUMI[1], (Graduate Student Member, IEEE), CHARLES SNOW[2], RICHARD K. LOMOTEY[2], (Member, IEEE), AND RALPH DETERS[1], (Member, IEEE)

[1]Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5A2, Canada
[2]Information Sciences and Technology, The Pennsylvania State University, Monaca, PA 15074, USA

Corresponding author: Richard K. Lomotey (rkl5137@psu.edu)

This work was supported in part by the Penn State Beaver Academic Affairs, and in part by the University of Saskatchewan through the Natural Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT Artificial Intelligence and Machine Learning (AI/ML) as analytical tools can be applied across multiple social domains. Thus, these tools are being deployed in several ways to address societal issues and concerns for "social good". For instance, AI/ML has applicable use cases for crisis response, economic empowerment, educational demands, environmental challenges, equality and inclusion, health and hunger, and security and justice. In this work, we seek to explore the power and capability of AI/ML in understanding citizens' engagement, which can improve governance and smart city deployment. Specifically, we studied the views expressed by online users about the city of Saskatoon in Canada. The analyzed views have become a value chain that community leaders can use to improve the governance structure of the city. In the study, we extracted 114,390 comments from Reddit (i.e., Saskatoon subreddit posts) between January 1, 2019, and September 20, 2023, to discover topics to highlight citizens' concerns. We compare the performance of three major topic models, namely, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic with a K-means clustering algorithm in the discovery of topics from the collected Reddit comments. The BERTopic with the K-means clustering algorithm achieved the highest coherence score of approximately 0.64 in the extraction of 25 topics from the dataset. Our findings showed that BERTopic can discover coherent and diverse topics compared to LDA and NMF. We found 12 underlying themes by merging related topics. Also, we leveraged SiEBERT (a pre-trained transformer model), 4 supervised ML models, and VADER (a lexical sentiment analysis classifier) to identify the sentiments expressed in each theme. The SiEBERT model outperformed the other sentiment classifiers with an accuracy of 89% in the prediction of sentiments. The research discovered factors for smart city engagement such as Housing and Facilities, Education, Downtown Development, Tourism and Entertainment, Policing, Healthcare, Online Community, and Cost.

INDEX TERMS Machine learning, textual mining, social media, citizens engagement, smart city.

I. INTRODUCTION

Social media has become the de facto platform for citizens' engagement. Users find these platforms as avenues to express themselves on social issues, economies, and governance of their daily affairs [1], [2], [3], [4], [5]. The views expressed can have negative, neutral, or positive polarities depending on the user's feelings, views, understanding, and so on. Users' views can be individualized or collective reflections of how a group feels. When these user views are properly analyzed, stakeholders and leaders can better steer the affairs of the citizens since solutions can be customized. Understanding the

The associate editor coordinating the review of this manuscript and approving it for publication was Mehedi Masud.

citizens can also help with proper resource allocations and community project prioritization. The challenge however is that there is no straightforward way to analyze societal issues from data. As a result, we have witnessed the application of Artificial Intelligence and Machine Learning (AI/ML) to social media data/posts to better analyze users' sentiments, views, needs, and demands.

For instance, researchers in [1] showed that ML algorithms such as the logistic regression binary classifier could be used to analyze public social media posts related to pro-vaccine and anti-vaccine discourse. Similarly, [2] studied citizens of the Philippines' sentiments on social media when the country's house proposed a bill for decreasing the minimum age of criminal liability. Their work employed ML techniques such as clustering, natural language processing, and content analysis. The same techniques were used by [3] to study the communicative behavior, conversation themes, and network structures of "*Lockdown*" protest supporters and non-supporters based on their tweets. The work in [4] also showed that the Latent Dirichlet Allocation (LDA) and social network analysis can be used to study public discourse on international trade between countries based on users' expressions.

In this work, our goal is to analyze the concerns of the citizens of Saskatoon, Canada based on their social media engagements. The outcome of the work will enable city leaders to manage the resources of the city better as well as facilitate proper planning and prioritization of developmental projects. This will align with the city's vision for the implementation of a smart city using a social approach. Unlike existing works that focus on the Twitter community, we are using data from Reddit. In the study, we extracted 114,390 comments from Saskatoon subreddit posts between January 1, 2019, to September 20, 2023, to discover topics that emphasize citizens' concerns. Traditional topic models like LDA and NMF leverage text vectorization methods such as bag-of-words (BoW) to represent documents and do not take into consideration the context and semantic similarity of words in a sentence [47]. To resolve this issue, BERTopic uses Bidirectional Encoder Representations from Transformers (BERT) embeddings to capture the contextual word and sentence vector representations of documents [47]. This preprocessing step allows BERTopic to preserve the semantic relationships between words to generate coherent and accurate topic representations [47].

Our study aims to compare the performance of three major topic models, namely, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic with a K-means clustering algorithm in the discovery of topics from the collected Reddit comments. The BERTopic with the K-means clustering algorithm achieved the highest coherence score of approximately 0.64 in the extraction of 25 topics from the dataset. Our findings showed that BERTopic can discover coherent and diverse topics compared to LDA and NMF. We discovered 12 underlying themes by merging related topics. Also, we leveraged SiEBERT (a pre-trained transformer model), 4 supervised ML models, and VADER (a lexical sentiment analysis classifier) to identify the sentiments expressed in each theme. The SiEBERT model obtained an accuracy of 89% compared to the VADER and Random Forest models which achieved an accuracy of 57% and 86% respectively in the prediction of sentiments. The research discovered factors for smart city engagement such as Housing and Facilities, Education, Downtown Development, Tourism and Entertainment, Policing, Healthcare, Online Community, Cost, and Animal Control.

In summary, the paper made the following contributions to social science, social computing, and governance.

- Employed AI/ML, and social media data from the Reddit community to highlight the societal issues affecting the citizens of Saskatoon, Canada. Hence, pivot the vision of the city towards smart city design.
- Explored multiple AI/ML topic models such as LDA, NMF, and BERTopic to determine the best-performing model as applied to the Reddit data.
- The work further evaluated the performance of the BERTopic model (a) when no preprocessing is done, and (b) when data is pre-processed.

The remaining sections of the paper are as follows. Section II describes the background works. Sections III and IV detail the methodologies and the generation of topics respectively. Also, Sections V and VI explain our thematic analysis and sentiments analysis respectively. We discussed our findings in Section VII and the paper concludes in Section VIII.

II. THE APPLICATION OF AI/ML TO SOCIAL MEDIA DATA ANALYSIS

Citizens' engagement aids government officials in understanding the needs of the public and can lead to frictionless decision-making. In this regard, social media platforms have become the go-to place for citizens to communicate and express their concerns about their expectations from community leaders. While community leaders also engage in social media, it is difficult to understand the generality and collective views of social media users due to the vast amount of data available on these platforms. Hence, techniques such as AI/ML algorithms and Natural Language Processing (NLP) have been employed to better analyze these data which can lead to an informed decision-making process [2].

Social media analysis can be used to analyze sentiments and emotional states with public disclosure [5], [6], [7], [8], [9], [10]. To understand the emotional and perceptual dimensions of citizens on innovations within smart cities, Adikari and Alahakoon [5] turned to Twitter to collect related data for opinion analysis. Specifically, they focused on the impact of self-driving cars on city development. They employed NLP and Markov models for the analysis while the negativity (toxicity) in conversations was evaluated using a deep learning-based classifier developed with layers of word embedding, bidirectional Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). Similarly,

Jain et al. [11] studied the emotions of citizens on social media about smart cities using the Bidirectional Encoder Representations from Transformers (BERT). The work also utilized the Dilated Convolutional Neural Network (DCNN) and SenticNet to improve the classification ability of BERT. Other researchers who classified public emotions using machine learning algorithms include Adamu et al. [12].

Additionally, Melton et al. [13] and Liu et al. [14] used NLP, Latent Dirichlet Allocation (LDA) modeling, and BERT sentence clustering to identify semantics within Reddit text as applied to public sentiments on COVID-19 vaccination. Likewise, Singh et al. [15] turned to Twitter data and applied sentiment and clustering analysis to understand the adverse effect that the COVID-19 pandemic has on agriculture stakeholders. Correspondingly, Chau et al. [16] used support vector machines (SMV) and rule-based classification to sift through numerous blog articles to detect emotionally distressed individuals during the COVID-19 pandemic. Basiri et al. [17] also applied a fusion sentiment analysis model that combines deep learning models to understand people's reactions to the coronavirus. Praveen et al. [18] used NLP and LDA to comprehend the sentiments of citizens towards the COVID-19 pandemic.

Further, Yigitcanlar et al. [19] investigated the use of AI and NLP in urban planning and development in Australia. This is better to perceive public perception and implement practical interventions. Besides, Kovacs-Györi et al. [20] employed Artificial Neural Networks (ANN), Support Vector Machine (SVM), geospatial data, and big data analytics in the realm of urban planning and the assessment and improvement of livability in cities. Similarly, Tran et al. [21] and Milusheva et al. [22] used NLP to identify the sentiments of online users describing their transit experience and road crashes in urban areas.

Also, Hodorog et al. [23] and Elabora et al. [24] focused their works on event detection in urban and smart cities with the aid of ML and Social Media Analysis (SMA). In [23], they employed Multiple Regression Analysis (MRA), semantic-based risk classification, and a selected combination of supervised NLP techniques for event detection in smart cities. They focused on analyzing social media data primarily from Twitter with an accuracy rate of 88.5% in risk event detection.

Hassan et al. [25] used SMA and graph convolutional networks to understand statistical trends in data sources such as police reports, call data records, and citizen profile data to discover criminal networks. Also, Chen et al. [26] used Long Short Term Memory (LSTM), which is an extension of RNN architecture, and an online forum to make a military sentiment dictionary.

Also, Fan et al. [27] and Modha et al. [28] study hate and aggressive speech online. They used tools like SVM, Logistic Regression, CNN, and BERT to figure out if a comment was aggressive or not. Also, Zarouali et al. [29] researched the significant relationship that exists between the use of political microtargeting (PMT) on social media and changes in citizens' attitudes and voting patterns. This is like Guess et al. [30] who used SMA to study the attitudinal and behavioral changes in users during the electioneering season. Moreover, Nistor and Zadobrischi [31] and Bojjireddy et al. [32] explored the problems that are associated with the spread of fake news on social media. They used machine learning algorithms like K-Nearest Neighbors, Linear Regression, Multinomial Naïve Bayes, Decision Trees, SVM, Random Forest, Gradient Boosting, and Multilayer Perception for the classification of texts. In [31], they achieved 90% accuracy when identifying fake news.

Kaur et al. [33] discuss the problem of false or misleading COVID-19 information posted through social media sites. The researchers used ML algorithms such as RNN, SVM, and Hybrid Heterogeneous SVM to determine the devastating effects that misleading information could have on the public. Reisach [34] also found that ML algorithms can influence misinformation on social media as pertains to critical issues such as public health.

Moreover, Alipour and Harris [35] explored the complex task of cost-effectively monitoring the conditions of urban infrastructure using computer vision and big data computing. The paper introduced a semi-supervised learning that leverages web images and Google Street View imagery to detect potholes and cracks in buildings. Also, Alahakoon et al. [36] utilized a self-building AI framework to study the challenges faced by the modern urban environment due to the massive figure of urban migrations and also the importance of smart cities as a solution to those challenges.

Furthermore, Bono et al. [37] studied how relevant information related to emergencies could be distilled from online platforms using CNNs and crowdsourcing. The work was tested on the Albania earthquake of 2019, the Covid-19 pandemic, and the Thailand floods of 2021. Similarly, Kankanamge et al. [38] after the use of AI posit that social media in a disaster context carries real-time crisis information such as the status of communication channels, status of roads, needs for the evacuation camps, and other important knowledge. Likewise, Biggers et al. [39] used an implementation of the Word2Vec method of neural network training to present the changing semantics of Twitter within the context of a crisis event, specifically tweets during Hurricane Irma.

Finally, McClure et al. [40] employed ML models to investigate the convergence of AI and citizen science within the field of ecological monitoring.

III. METHODOLOGY

In Fig. 1, we illustrate the overview of the proposed approach of using social media sensing and machine learning to extract the opinions of citizens. Using the city of Saskatoon as a case study, we extracted comments from Saskatoon Subreddit posts to discover topics discussed on Reddit. We used three topic modeling algorithms namely, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and

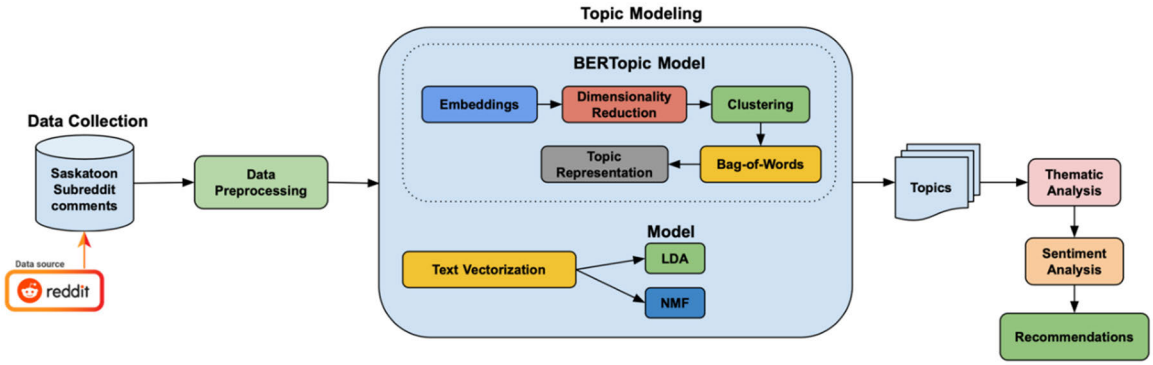

FIGURE 1. Overview of the proposed approach.

BERTopic to discover the hidden topics in the collected comments. We categorized similar topics into themes, and then leveraged sentiment analysis tools to explore the sentiments of each text in the discovered themes. The workflow of the proposed approach is discussed below.

A. DATA COLLECTION

Posts on Reddit are categorized based on the level of engagement and popularity. We leveraged the Python Reddit API Wrapper (PRAW)[1] to collect the hot, top, controversial, and rising posts from the Saskatoon subreddits. We retrieved 2,562 posts between the period of January 1, 2019, to September 20, 2023. Fig. 2 shows the proportion of posts collected for each year. As shown in Fig. 2, the number of posts at the beginning of the period was low.

However, the level of engagement of Saskatoon citizens on Reddit increased in the subsequent years. We extracted 114,390 comments from the collected posts for further analysis. The graphs in Fig. 3 and Fig. 4 show the number of comments in the original dataset and the cleaned dataset respectively.

B. DATA PREPROCESSING

Data preprocessing is an essential step in topic modeling and sentiment analysis. Performing data preprocessing before applying topic models on collected texts improves the interpretability of topics generated by the models. We employed the following Natural Language Processing (NLP) techniques to preprocess the collected comments.

1) DATA CLEANING

a. *Removal of bots and AutoModerator accounts.* AutoModerator is a built-in system in Reddit that allows moderators of a Reddit community (subreddit) to define rules. Comments from bots and AutoModerator accounts often add noise to a dataset and may affect the quality of topics generated by topic models. Hence, in this study, our goal is to analyze comments from only human users. We detected bot accounts from the collected data by analyzing the account names and comment contents. Most bots often use very short and common phrases. To detect bots by comment content, we removed accounts with comments of less than 30 characters. To identify bots by names, account names that consist of more than two words separated by an underscore or hyphen, and followed by the keyword bot were discarded.

b. *Removal of [deleted] or [removed].* In a Reddit comment section, "[deleted]" means a comment was deleted by the user who posted the comment, and "[removed]" indicates that the comment was deleted by the moderator of the subreddit. We eliminated comment rows with such contents as they are irrelevant.

c. *Removal of Links.* We removed HyperText Markup Language (HTML) entities and Uniform Resource Locators (URLs) from the text as they can introduce noise and increase the dimension of the dataset.

d. *Removal of non-English comments*. Our work focuses on analyzing comments in English as most Saskatoon citizens use English. Additionally, ML translation models may not accurately translate non-English comments to English, which may affect the efficiency of topic modeling algorithms and sentiment classifiers [41]. We used the *FastText* Python library to eliminate non-English comments.

e. *Expand contractions*. We expand shortened versions of words such as *can't* into *cannot*, I'll becomes *I will*. The expansion of contractions in a text improves tokenization and enhances the performance of models for sentiment analysis and topic modeling.

f. *Check duplicate comments*. Topic models may assign higher importance to duplicate documents in a corpus, which may impact the inference of the topic model negatively [42]. We discarded duplicate comments to obtain a unique dataset to ensure that topics are accurately distributed across the dataset. After data cleaning, we had a total of 85,113 comments for further analysis. All comments in the cleaned dataset are converted to lowercase before performing additional data preprocessing.

[1] https://praw.readthedocs.io/en/stable/index.html

TABLE 1. Bag of words numerical vector representation of preprocessed comments.

T	child	cost	dog	gay	gender	life	man	ownership	parent	require	school	student	transition
1	0	1	1	0	0	1	1	1	0	0	0	0	0
2	1	0	0	1	1	1	0	0	2	1	1	1	1

TABLE 2. TF-IDF numerical vector representation of preprocessed comments.

T	cost	dog	man	ownership	life	child	gay	gender	parent	require	school	student	transition
TF-IDF	0.471078	0.471078	0.471078	0.471078	0.335176	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

2) TOKENIZATION

Tokenization is the process of splitting the comments into individual words or tokens. It transforms raw comments into a format that can be easily processed by ML models.

3) LEMMATIZATION

We lemmatized the words (terms) in the corpus to normalize them into their root form. We transformed words to only their noun and verbs to obtain meaningful words to enhance the accuracy of the topic model and sentiment analysis models.

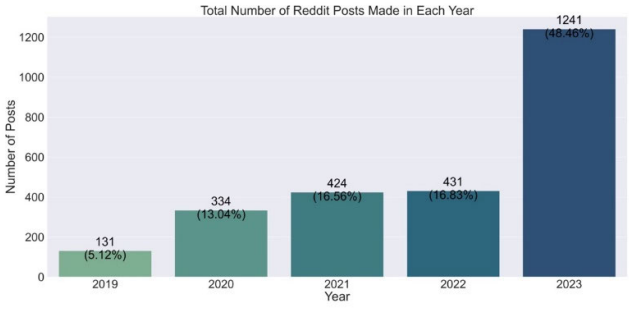

FIGURE 2. Total number of posts collected for each year from Saskatoon subreddit.

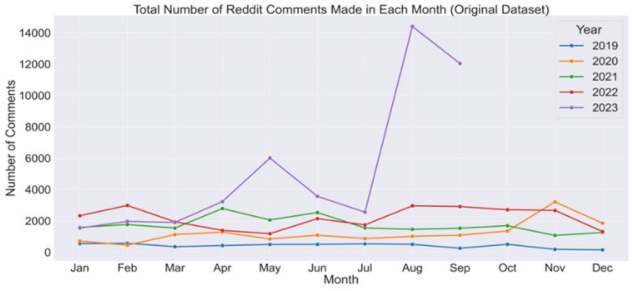

FIGURE 3. The number of comments in the original dataset.

4) STOPWORD REMOVAL

We eliminated stop words from our dataset that carry little meaning to obtain more meaningful terms in our corpus.

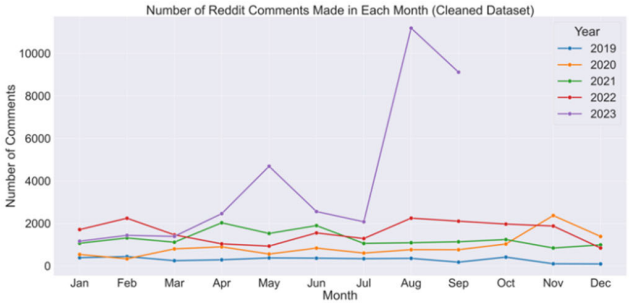

FIGURE 4. The number of comments in the cleaned dataset.

We extended the NLTK's stopword Python library with custom stopwords such as *"sask", "Saskatoon", "SK", and "Saskatchewan"*. These words might be common in our dataset, which may affect topic modeling results.

5) HANDLING N-GRAM

We identified the captured sequence of words in the comments to reduce ambiguity and improve the coherence of topics. We applied unigram (1-gram) i.e. single words and bigram (2-gram), which is the combination of two consecutive words.

C. TEXT VECTORIZATION

After the preprocessing of comments, we leveraged text vectorization techniques to transform the texts into document term matrices (DTM) for topic models. DTM is the transformation of a text corpus into numerical representations for ML models. We applied three text vectorization techniques namely, Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and document embeddings. The BoW technique is based on the occurrence of words in a collection of documents. It ignores the order of words in a document. The BoW algorithm converts each word in a document to a numerical vector based on the word counts. Table 1 shows the BoW numerical vector representations of two sample comments (T1 and T2) from our dataset.

The TF-IDF measures how relevant a word is to a document in a collection of documents. The TF-IDF score is calculated by multiplying two factors: the frequency of a

word in a document (TF) and the inverse frequency of the word across the collection of documents (IDF). The higher the score, the more important the word is in the document. The TF-IDF algorithm represents each word in a document as a numerical vector by assigning weights based on their importance. The TF-IDF algorithm is defined as in Equation (1)–(4), shown at the bottom of the page, where TF ($t_{w,d}$) denotes the frequency of the word, w in the document, d. The IDF calculates how frequent or rare a word is in the entire corpus by taking the log of the number of documents in a corpus, N divided by the total number of documents with the word, w. Table 2 illustrates the TF-IDF scores of two sample comments from our dataset.

Document embeddings are the use of pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT) to convert documents into dense vectors while preserving the semantic similarities between documents.

Original T1: > The real cost of dog ownership for me is easily over $50,000 a year as a single employable man in the prime of my life.Huh?.

Preprocessed T1: cost dog ownership man life

Original T2:If a child doesn't want to tell their parents about something crucial in their life (such as gender transition) there is probably a reason for it. This is fundamentally no different from requiring schools to out bi(sexual) students to their parents.

Preprocessed T2: child parent life gender transition require school bi(sexual) student parent

D. TOPIC MODELING

1) LATENT DIRICHLET ALLOCATION (LDA)

The Latent Dirichlet Allocation (LDA) is described as a three-level generative probabilistic model to discover hidden topics in a collection of documents [43]. LDA assumes that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [43]. The generative process for each document in a given corpus (collection of documents) to discover topics is as follows:

a. *For each document, d in a corpus, M*
   Choose the topic distribution per document, θ_m from Dirichlet parameter, α.

$$\theta_m \sim Dir\,(\alpha) \tag{5}$$

b. *For each word $w_{m,n}$ in the document:*
   i. Choose a topic $z_{m,n} \sim$ Multinomial (θ_m).
   ii. Find the distribution of words in topics, Φ_k from Dirichlet parameter, β.
   iii. Sample word, $w_{m,n}$ from a multinomial probability, $p\left(w_{m,n} \mid z_{m,n}, \beta\right)$ conditioned on the topic, $z_{m,n}$.

The graphical illustration of the model representation of LDA is shown in Fig. 5. The boxes are "plates" representing replicates. The outer plate, M denotes documents and the inner plate, N represents the repeated choice of topics and words within a document. The K plate represents the topics hidden in a collection of documents (corpus). Table 3 describes the notations used as reproduced from [41].

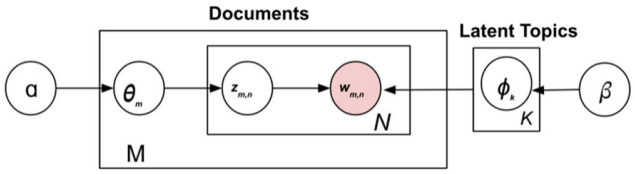

FIGURE 5. Graphical representation of the LDA model.

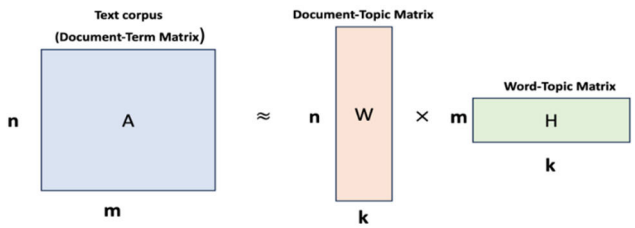

FIGURE 6. Graphical representation of the NMF model.

As LDA assumes that topics discovered in a collection of documents are a mixture of words, it adopts the BoW approach for text vectorization. BoW text vectorization ignores the order of words in a document, hence maintaining the word frequencies across the documents [43].

2) NON-NEGATIVE MATRIX FACTORIZATION (NMF)

The *Non-Negative Matrix Factorization (NMF)* is a linear algebraic algorithm that was first introduced as a positive matrix factorization by Paatero and Tapper [44]. The algorithm was later reintroduced by Lee and Seung [45] to learn the semantic features of texts. NMF is a dimension

$$W_{w,d} = TF \cdot IDF \tag{1}$$

$$W_{w,d} = t_{w,d} \cdot \log\left(\frac{N}{df_w}\right) \tag{2}$$

$$TF = \frac{n \text{ times a word occurs in the document}}{\text{total number of words in the document}} \tag{3}$$

$$IDF = \log\left(\frac{\text{number of documents in a corpus}}{\text{total number of documents in the corpus with the word}}\right) \tag{4}$$

TABLE 3. Description of notations used in LDA.

Notation	Description
M	Number of documents
N	Number of unique words
$w_{m,n}$	Words in document
$z_{m,n}$	Topic assigned to words
Φ_k	Word distribution in topics
θ_m	Topic distribution in the document
α	Dirichlet prior for θ
β	Dirichlet prior for Φ

reduction method that extracts meaningful data from high-dimensional data [46]. The goal of NMF is to find the lower-dimensional non-negative elements that approximate a given non-negative document term matrix [45]. As illustrated in Fig. 6, NMF decomposes a given text corpus, expressed as a non-negative $n \times m$ matrix A, into two non-negative factors, W, of $n \times k$ matrix and H, of $m \times k$ matrix such that:

$$A \approx WH \tag{6}$$

where A is the set of documents (corpus) to discover topics. A is expressed in the form of a document term matrix (DTM) while W denotes the document-topic matrix. H represents the word-topic matrix, and k represents the number of topics to be extracted from A. The document-topic matrix describes the relationship between the number of documents in a text corpus, n, and the extracted topics, k. The word-topic matrix describes the prevalence of words, m in a given topic, k.

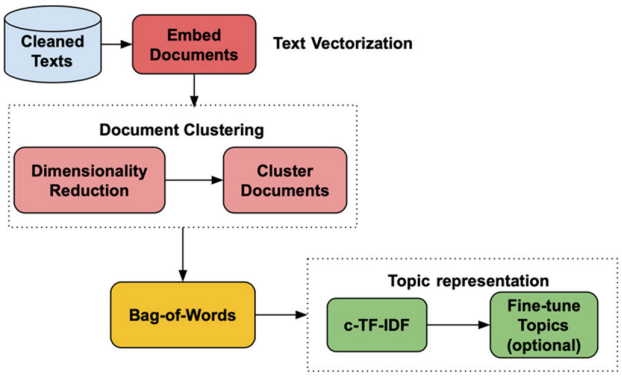

FIGURE 7. Generation of topics with BERTopic.

Unlike the LDA topic model, NMF can be applied to either BoW or TF-IDF transformed text corpus.

3) BERTOPIC

BERTopic employs transformers, clustering algorithms, and class-based variation of Term Frequency- Inverse Document Frequency (c-TF-IDF) to extract coherent topic representations [47]. The modular nature of BERTopic allows the exploration of different sentence transformers, clustering, and dimensionality reduction algorithms to design your topic model. Fig. 7 illustrates the topic generation process of BERTopic.

BERTopic extracts topics from documents through three steps.

a. *Embed Documents.* Documents are embedded using sentence transformers to convert sentences and paragraphs to vector representations. The default BERTopic architecture uses the Sentence-Bidirectional Encoder Representations from Transformers[2] (SBERT) framework for document embeddings.

b. *Document clustering.* Document embeddings are clustered into semantically similar documents. The dimensions of the embedding documents are reduced before clustering. This improves the performance of clustering algorithms in terms of accuracy and execution time [47]. As a default, BERTopic uses the Uniform Manifold Approximation and Projection (UMAP) technique to reduce the dimensionality of document embeddings and the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to cluster the reduced embeddings.

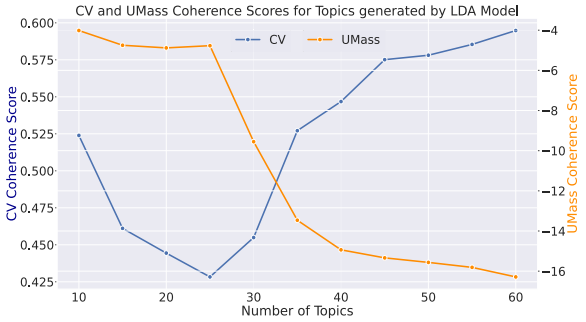

FIGURE 8. Coherence scores for LDA model.

c. *Topic Representation.* All documents in a cluster are merged into a single document to create a bag-of-words (BoW). The BoW computes the frequency of words in each cluster. BERTopic algorithm creates BoW on the cluster level and not on a document level. Topics are assigned based on the documents in each cluster. Each cluster is assigned a topic. BERTopic adopts a class-based TF-IDF (c-TF-IDF) strategy to extract the most important words in a cluster to generate accurate topic representations. As shown in equation 7, the c-TF-IDF is computed by taking the logarithm of the average number of words per class A divided by the frequency of term t across all classes [47]. Additionally, BERTopic employs several representation models to fine-tune the topics generated. Topics may be fine-tuned based on the semantic relationship between keywords, part of speech and to decrease redundancy. The fine-tuning step is optional.

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{f_t}\right) \tag{7}$$

[2]https://www.sbert.net/

TABLE 4. Training parameters for LDA and NMF.

Model	Parameters
LDA	number of topics = $k \in \{10 : 60\}$, corpus = BoWs, alpha = ' symmetric', beta = 'symmetric', random state = 100
NMF	number of topics = $k \in \{10 : 60\}$, corpus = TF-IDF, kappa = 0.1, random state = 42

TABLE 5. Training parameters for BERTopic models.

Model	Clustering		Dimensionality Reduction		Embedding Model
	Algorithm	*Parameters*	*Algorithm*	*Parameters*	
BERTopic_UMAP_HDBSCAN	HDBSCAN	minimum cluster size = 10, metric = 'Euclidean', cluster selection method = 'eom'	UMAP	number of neighbors = 15, number of components = 5, minimum distance = 0.0, metric = 'cosine', random state = 42	all-MiniLM-L6-v2
BERTopic_UMAP_Kmeans	K-means	number of clusters = 60, method of initialisation = 'k-means++', n_init = 10, maximum number of iterations = 300			

TABLE 6. The number of topics with their coherence score for the LDA model.

Number of Topics (k)	Coherence Score	
	C_V	UMass
10	0.5239	-4.016
15	0.4611	-4.7499
20	0.4444	-4.8812
25	0.4283	-4.7711
30	0.4549	-9.5521
35	0.5271	-13.463
40	0.5468	-14.9385
45	0.5751	-15.338
50	0.5781	-15.5639
55	0.5854	-15.8116
60	0.5948	-16.2867

where $tf_{t,c}$ is the frequency of word t in class c, f_t is the frequency of word t across all classes, and A is the average number of words per class.

E. THEMATIC ANALYSIS

A thematic analysis is performed to merge related topics into themes. We merged topics based on the occurrence of keywords, intertopic distance maps, and similarity matrix of topics.

We manually analyzed the top keywords of each topic to assign a label. Topics with common keywords and related concepts are merged to form a theme.

The intertopic distance map is a visualization of relationships between topics in two-dimensional space using dimensionality reduction techniques. The intertopic distance map of LDA and NMF topic models is computed using multidimensional scaling (MDS) while the BERTopic's intertopic distance map is based on the UMAP algorithm. In the intertopic distance map, the distribution of each topic is represented with a circle. The distance between the circles in the distance map determines the similarity between topics. Circles (Topics) that are overlapping or closer indicate that the topics are semantically similar. We check for topics that are overlapping in the intertopic distance map to identify related topics.

Additionally, for the BERTopic model, we visualize the similarity matrix of discovered topics to identify related topics. The similarity matrix is computed based on the cosine similarity of the topic embeddings (vector representations). The cosine similarity is a similarity measure that quantifies the similarity between two vectors. In BERTopic, the topic embeddings (vectors) are generated using sentence transformers (embeddings) and c-TF-IDF. The cosine similarity score ranges from 0 to 1. A cosine similarity score closer to 1 indicates a higher similarity. We set the cosine similarity score threshold to above 0.80 to determine related topics.

F. SENTIMENT ANALYSIS

We leveraged machine learning and lexicon-based techniques to identify the sentiments associated with each theme. For the lexicon-based sentiment classifier, we used the Valence Aware Dictionary and sEntiment Reasoner (VADER) [48] tool to assign sentiments. VADER uses a compound score between the range of -1 and 1 to label the sentiment of a given text as positive, neutral, or negative.

In the case of leveraging ML models for sentiment analysis, we employed a pre-trained transformer model, SiEBERT (Sentiment in English) [49], and trained four supervised models namely: Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost) on a manually annotated comments randomly sampled from our entire dataset used in this study. The SiEBERT model is a fine-tuned checkpoint of RoBERTa-large [50] and evaluated on 15 data sets from diverse text sources to enhance generalization across different types of texts.

IV. GENERATION OF TOPICS

We evaluated the performance of three topic models namely Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic in the extraction of topics from our collected comments from the Saskatoon subreddit. We leveraged a cleaned dataset of 85,113 to analyze the efficacy of the topic models. The experiments in this study were implemented in Python. We trained the LDA and

NMF models on a Mac operating system with an Intel Core i7 3.8 GHz 8-Core Processor with 32GB of memory. The BERTopic experiments were carried out on Google Colab notebooks with GPU runtime.

We used the topic coherence metric to evaluate the performance of the topic models in extracting hidden topics from the Saskatoon subreddit corpus. Topic coherence metric uses the semantic similarity of words in a topic to measure the meaningfulness and interpretability of topics generated by models. The UMass [51] and CV [52] metrics were used to evaluate the coherence of topics generated by the models. The UMass coherence metric is based on the co-occurrence of words in a collection of documents. The CV coherence metric uses the segmentation of top word subsets, a sliding window, and the aggregation of indirect confirmation measures based on normalized pointwise mutual information (NPMI) and cosine similarity of the top word subsets. A higher topic coherence score indicates that the generated topics are coherent and can easily be interpretable by humans.

A. TRAINING OF TOPIC MODELS

We performed a grid search with different numbers of topics (k), $k \in \{10 : 60\}$ where k ranges from 10 to 60 with a step of 5, on each model to determine the best number of topics in our corpus. Default parameters are used in training models to have a fair comparison.

The *Gensim*[3] Python library was used in training the LDA and NMF models. In the training of LDA and NMF models, we preprocessed the corpus by removing punctuations, tokenizing, lemmatizing, removing stopwords, computing unigrams and bigrams, and removing terms that appear in less than 5 documents. In training the LDA model, the bag-of-words text vectorization technique was used to transform the corpus into numerical representations. We set the document-topic distribution (alpha) to *'symmetric'* and the topic-word distribution (beta) to *'symmetric'* to train the LDA model.

The NMF model was fitted with a term frequency-inverse document frequency (TF-IDF) transformed corpus and gradient descent step size (kappa) of 0.1 for training. A minimum value is used for kappa to ensure that NMF convergences during training. Table 4 shows a summary of parameters used in training the LDA and NMF models.

For BERTopic, we trained two variations to discover topics. We label the first variation as *BERTopic_UMAP_HDBSCAN*, which is the default BERTopic architecture, and the second variation as *BERTopic_UMAP_Kmeans*. The two variations use the Uniform Manifold Approximation and Projection (UMAP) algorithm for document dimension reduction. The *BERTopic_UMAP_HDBSCAN* models use the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) for clustering. The *BERTopic_UMAP_Kmeans* leverages the K-means clustering algorithm to cluster documents. We utilized the bertopic[4] library implemented in Python to train the BERTopic models. For every number of topics (k), $k \in \{10 : 60\}$, we follow the steps detailed below to train the BERTopic models:

a) We created document embeddings by converting our corpus to numerical representations using the *"all-MiniLM-L6-v2"* sentence transformer.
b) Apply the UMAP algorithm to reduce the dimensions of the embeddings.
c) Apply the respective clustering technique of each variation to cluster documents.
d) We create BoWs using the *CountVectorizer() module* from the scikit-learn package to find the frequency of each word in each cluster. We set the *CountVectorizer()* to from unigrams and bigrams during the generation of BoWs.
e) The class-based TF-IDF (c-TF-IDF) is applied to the generated bag-of-words to create topic representations.

To speed up the dimensionality reduction and clustering step, we leveraged the cuML version of UMAP and HDBSCAN through GPU acceleration. The *BERTopic_UMAP_Kmeans* model uses the scikit-learn implementation of the K-Means clustering algorithm to cluster documents. Table 5 shows a summary of parameters used in training the BERTopic models. The *BERTopic_UMAP_HDBSCAN* uses its default parameters for training. In the case of the *BERTopic_UMAP_Kmeans,* all parameters are set to default values except for the number of clusters parameter which is set to 60 to cover the highest topic number in the range of predefined topic numbers. The number of clusters defined influences the number of topics to be generated.

TABLE 7. Number of topics with their coherence score for the NMF model.

Number of Topics (k)	Coherence Score	
	C_V	UMass
10	0.5855	-3.691
15	0.5517	-4.0324
20	0.5782	-3.9374
25	0.5373	-4.1141
30	0.5344	-4.4069
35	0.5346	-4.3166
40	0.5228	-4.4685
45	0.5136	-4.6136
50	0.5134	-4.5755
55	0.5063	-4.7874
60	0.4814	-4.926

B. EVALUATION OF TOPIC MODELS

Topic models were evaluated using the C_V and UMass topic coherence metric. The C_V coherence score ranges between 0 to 1. A C_V coherence score closer to 1 implies better coherence. The UMass coherence score ranges from negative to positive values. A UMass coherence score closer to 0 indicates better coherence.

Table 6 shows the C_V and UMass coherence scores for the LDA models for topics ranging from 10 to 60. In terms

[3]https://radimrehurek.com/gensim/index.html

[4]https://maartengr.github.io/BERTopic/index.html

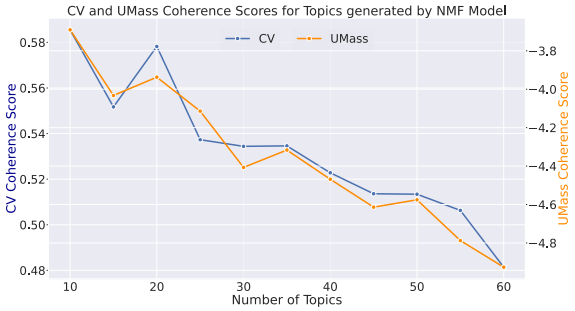

FIGURE 9. Coherence scores for NMF model.

of the C_V coherence score, the LDA model has the highest coherence score of 0.5948 for the 60 topics. In the case of the UMass coherence score, the optimal number of topics for the LDA model is 10 with a coherence score of -4.016. As illustrated in Fig. 8, the UMass coherence score of the LDA model decreases as the number of topics increases. This implies that as the topic number increases, the topics generated may be less meaningful and difficult to interpret. We experimented with generating 60 topics with the LDA model as it has the best C_V coherence score. In generating topics with optimal topics for LDA and NMF models, we used the scikit-learn Python package. From our experimental results, we observed that generating a higher number of topics with the LDA model yielded topic keywords that were less semantically meaningful and difficult to interpret.

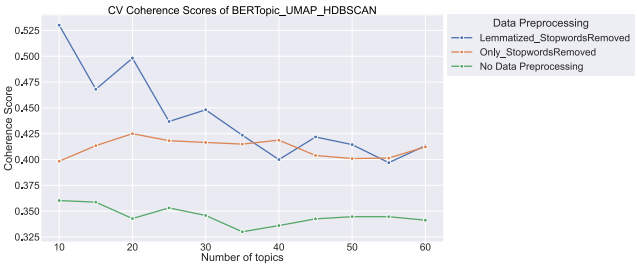

FIGURE 10. CV coherence score BERTopic_UMAP_HDBSCAN on the level of data preprocessing.

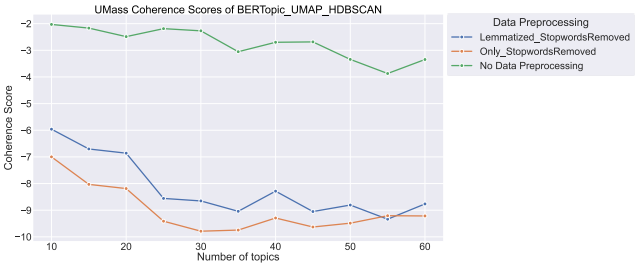

FIGURE 11. UMass coherence score BERTopic_UMAP_HDBSCAN on the level of data preprocessing.

The C_V and UMass topic coherence scores for NMF are shown in Table 7. The scores for both metrics show the best number of topics for the NMF model on our corpus is 10 with

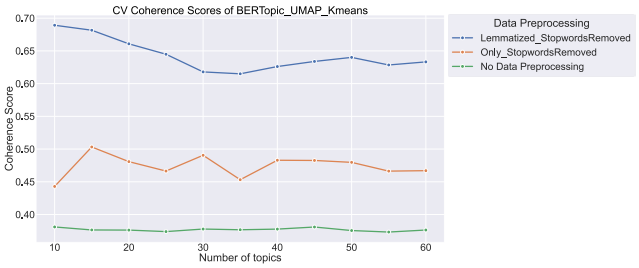

FIGURE 12. CV coherence score BERTopic_UMAP_Kmeans on the level of data preprocessing.

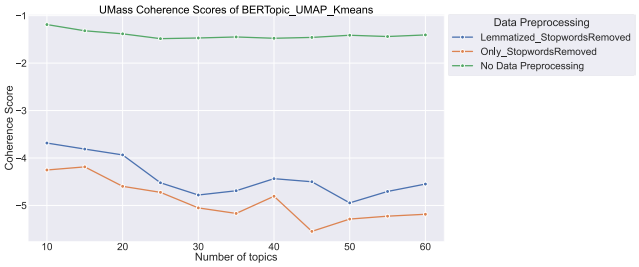

FIGURE 13. UMass coherence score BERTopic_UMAP_Kmeans on the level of data preprocessing.

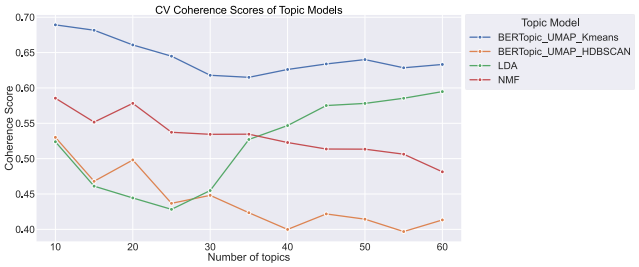

FIGURE 14. CV coherence score of all topic models.

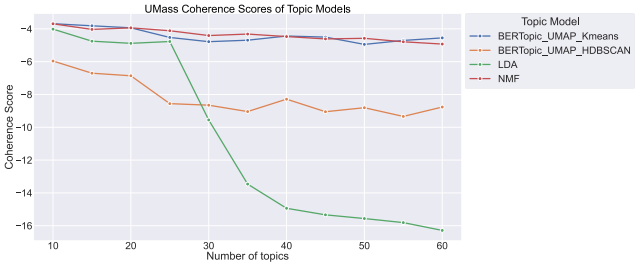

FIGURE 15. UMass coherence score of all topic models.

a C_V coherence score of 0.5855 and a UMass coherence score of -3.691. As shown in Fig. 9, the coherence score of both metrics decreases from topics 30 to 60. Hence, in the case of NMF, the best number of topics hidden in our corpus ranges from 10 to 25.

Table 8 shows the top 10 terms for 10 topics discovered by LDA and NMF models. We manually assigned labels to the topics using the weights of the top 10 terms. Common topics such as Education, Transportation, Humanity, and COVID-19 were discovered by both models.

The BERTopic model leverages BERT embedding models to capture the contextual information of the corpus to create accurate vector representations of the documents. Thus preprocessing steps such as stopwords removal and lemmatization are not encouraged. However in this study, we explored the performance of the two variations of BERTopic on three types of datasets; no preprocessed corpus, slightly preprocessed corpus and fully preprocessed corpus.

For the no preprocessed corpus data cleaning such as HTML links are removed as they do not contribute any meaning to texts. In the slightly preprocessed corpus, only stopwords and punctuations are removed. The fully preprocessed corpus involves preprocessing techniques such as the removal of stopwords and punctuations and lemmatizing words to only noun and verb forms. For each dataset, we computed the unigrams and bigrams and filtered out words that appear in less than 5 documents to train the two variations of BERTopic.

Table 9 shows the performance of the first variation, *BERTopic_UMAP_HDBSCAN* model on each dataset for topic numbers ranging from 10 to 60. The *BERTopic_UMAP_HDBSCAN* model yielded 10 topics as the optimal number of topics for all three datasets. For all three datasets, the coherence score for both C_V and UMass metrics decreases as the topic number increases. As shown in Fig. 10, the C_V coherence score of the *BERTopic_UMAP_HDBSCAN* model increases when additional data preprocessing is performed. However, the fully preprocessed data obtained the highest C_V coherence score of approximately 0.53. In the case of the UMass coherence score, as illustrated in Fig. 11, the no-preprocessed corpus has the highest coherence score. We observed that the UMass coherence score of the *BERTopic_UMAP_HDBSCAN* model decreases when further preprocessing is performed on the corpus.

Similarly, the C_V coherence score of topics generated by the second variation of the BERTopic, *BERTopic_UMAP_Kmeans* model as shown in Table 10 increases when additional preprocessing is done. The C_V coherence score of the *BERTopic_UMAP_Kmeans* model trained on each dataset is shown in Fig. 12. The fully preprocessed obtained a better C_V coherence score, with 10 topics presented as the best number of topics in our corpus. The UMass coherence score of the *BERTopic_UMAP_Kmeans* model as visualized in Fig. 13 shows that 10 topics in the optimal number of topics hidden in our corpus across all three datasets.

From our experiments, we observed that the Cv coherence score of the two variations of the BERTopic model increases when trained on a fully preprocessed corpus. However, the *BERTopic_UMAP_Kmeans* model achieved higher coherence scores compared to the *BERTopic_UMAP_HDBSCAN* model. Table 11 shows the top 10 terms for 10 topics generated for each dataset with the *BERTopic_UMAP_Kmeans* model. The terms generated for the slightly and fully preprocessed corpora are easy to interpret compared to the no preprocessed corpus. The no preprocessed corpus topic terms are full of stopwords and are difficult to interpret. Although preprocessing is not required for the BERTopic model, depending on the case study applied, the dataset must be preprocessed to achieve accurate topic representations.

C. COMPARISON OF TOPIC MODELS

In this study, we compared the results of the BERTopic models trained on a fully preprocessed corpus with the LDA and NMF models. Fig. 14 and Fig. 15 show the Cv and UMass coherence scores of all topic models respectively. In terms of C_V coherence score, the *BERTopic_UMAP_Kmeans* model obtained the highest score across all k topics. The NMF model was the second-best performing model for topic number ranging from 10 to 35. Nevertheless, the LDA model performed better than the NMF and *BERTopic_UMAP_HDBSCAN* models for topic numbers between 40 to 60. The *BERTopic_UMAP_HDBSCAN* model obtained the lowest Cv coherence score for topic numbers ranging from 30 to 60, however, achieved a higher Cv coherence score for topic numbers from 10 to 25 than the LDA model. We observed that the Cv coherence score of LDA increases as the topic number increases, whilst that of NMF, *BERTopic_UMAP_Kmeans,* and *BERTopic_UMAP_HDBSCAN* models decreases as the topic number increases.

In the case of the UMass coherence metric, the NMF slightly performs better than the *BERTopic_UMAP_Kmeans* model for most of the topic numbers. The LDA model is the worst-performing model for topic numbers within the range of 30 to 60.

All topic models used in this study, output 10 as the optimal number of topics hidden in the Saskatoon subreddit. However, from Tables 8 and 11, it is likely the topic models may not adequately capture all the diverse topics in the corpus. Comparing the C_V and UMass coherence scores as shown in Fig. 14 and Fig. 15, the optimal number of topics ranges from 10 to 30 for all the topic models.

We decided to use 25 as the best topic number to train the models to extract hidden topics. The top 10 terms for 25 topics generated by LDA, NMF, and BERTopic_UMAP_HDBSCAN are shown in Table 12. The top 10 terms for 25 topics generated by *BERTopic_UMAP_Kmeans* are shown in Table 13. The NMF model outperformed LDA and BERTopic_UMAP_HDBSCAN, but most of the topics generated by the model contain redundant keywords. For the LDA model, four out of the 25 topics generated could not be interpreted. These topics are given the label N/A. Overall the BERTopic variations generated meaning topic terms that easily be interpreted compared to LDA and NMF. This performance of BERTopic variations can be attributed to the use of embedding models to capture the contextual information to ensure that similar texts are placed in the same cluster and a class-based TF-IDF to improve the accuracy of topic representations. The first variation of the BERTopic model, *BERTopic_UMAP_HDBSCAN,* uses HDBSCAN as the clustering method which groups unrelated comments

TABLE 8. Top 10 terms of 10 topics generated by LDA and NMF.

LDA			NMF	
Topic ID	**Top 10 Terms**	**Topic Label**	**Top 10 Terms**	**Topic Label**
1	party, government, vote, help, protest, issue, support, province, country, drug	Governance	give, point, person, help, care, agree, issue, change, life, problem	Humanity
2	life, give, read, man, believe, woman, opinion, hate, person, church	N/A	kid, parent, child, teacher, parent kid, kid parent, drag, adult, kid school, family	Education
3	car, bike, road, walk, stop, winter, cyclist, drive, ride, lane	Transportation	car, bike, lane, stop, road, turn, driver, traffic, cyclist, vehicle	Traffic
4	mask, wear, case, covid, vaccine, health, doctor, hospital, number, risk	COVID-19	mask, wear, wear mask, mask wear, mandate, covid, face, virus, spread, distance	COVID-19
5	kid, school, child, parent, teacher, care, student, education, family, health	Education	love, hate, photo, beaver, friend, picture, winter, pic, love love, kid love	Pictures
6	money, tax, cost, cat, pay, spend, point, dollar, power, animal	Cost	call, name, police, phone, cop, call police, call call, name call, call cop, call name	Policing
7	downtown, area, find, build, show, move, house, play, building, remember	Development	find, hope, hope find, friend, price, buy, area, house, search, store	Housing
8	pay, business, property, service, job, taxis, restaurant, charge, store, wage	Services	pay, tax, money, taxis, job, cost, buy, pay taxis, price, property	Taxes
9	drive, driver, vehicle, lane, traffic, turn, truck, crime, person, stop	Traffic	drive, speed, circle, driver, truck, limit, circle drive, speed limit, vehicle, road	Transportation
10	call, buy, police, price, give, find, pay, phone, sell, garbage	Trading	school, teacher, student, church, school school, education, fund, child, teach, system	Education

TABLE 9. Number of topics with their coherence score for the BERTopic_UMAP_HDBSCAN model.

Number of Topics (k)	No Data Preprocessing		Only Stop words Removed		Stop words Removed and Lemmatized	
	Coherence Score		Coherence Score		Coherence Score	
	C_V	UMass	C_V	UMass	C_V	UMass
10	0.3602	-2.0296	0.3983	-6.9987	0.5301	-5.9598
15	0.3587	-2.1667	0.4134	-8.031	0.468	-6.7023
20	0.3428	-2.4872	0.425	-8.1878	0.4981	-6.8599
25	0.3531	-2.1874	0.4182	-9.4126	0.4368	-8.5579
30	0.3458	-2.2694	0.4165	-9.788	0.4481	-8.6525
35	0.33	-3.0532	0.4149	-9.746	0.4235	-9.0417
40	0.336	-2.6991	0.4187	-9.2929	0.3999	-8.2854
45	0.3425	-2.6849	0.4039	-9.6305	0.4218	-9.0505
50	0.3446	-3.341	0.4009	-9.4868	0.4144	-8.8066
55	0.3446	-3.8709	0.4014	-9.2069	0.3969	-9.3379
60	0.3412	-3.3456	0.4122	-9.2159	0.4133	-8.7653

TABLE 10. Number of topics with their coherence score for the BERTopic_UMAP_Kmeans model.

Number of Topics (k)	No Data Preprocessing		Only Stop words Removed		Stop words Removed and Lemmatized	
	Coherence Score		Coherence Score		Coherence Score	
	C_V	UMass	C_V	UMass	C_V	UMass
10	0.381	-1.1874	0.4429	-4.2529	0.6892	-3.6849
15	0.3764	-1.3185	0.5032	-4.1874	0.6816	-3.8119
20	0.3762	-1.3826	0.4808	-4.5971	0.6608	-3.9351
25	0.3739	-1.4844	0.4664	-4.7241	0.6449	-4.5226
30	0.3778	-1.4703	0.4905	-5.0518	0.6179	-4.7804
35	0.3767	-1.4488	0.4532	-5.1686	0.615	-4.6907
40	0.3777	-1.4762	0.4829	-4.8066	0.6261	-4.4359
45	0.3809	-1.4584	0.4826	-5.5465	0.6339	-4.5009
50	0.3755	-1.4127	0.4797	-5.2876	0.6401	-4.9465
55	0.3732	-1.4396	0.4663	-5.2264	0.6285	-4.7054
60	0.3763	-1.4059	0.467	-5.1865	0.6332	-4.5495

in our corpus as outliers to improve topic representations. These outliers are labeled as -1 and discarded from the topic representations. The *BERTopic_UMAP_HDBSCAN* model applied to our dataset classified about 62% of the comments as outliers. However, after manually inspecting the comments, we identified important comments that are worth exploring. The second variation of the BERTopic model, *BERTopic_UMAP_Kmeans* allows the selection of the number of clusters and includes every comment in our dataset in a cluster. Although the *BERTopic_UMAP_HDBSCAN* model discards outliers, the BERTopic*_UMAP_Kmeans* model generated topics with a higher coherence score and can easily be interpreted. The parameters of the HDBSCAN clustering algorithm can be optimized to achieve the best results, but the optimization of the clustering algorithm is out of scope for this study.

TABLE 11. Top 10 terms of 10 topics generated by BERTopic_UMAP_Kmeans for each dataset.

Topic ID	No Data Preprocessing	Slightly Preprocessed (Only stopwords removed)	Fully Preprocessed (Lemmatized + stopwords removed)	Topic Label for fully preprocessed
0	to, is, you, have, be, on, if, as, do not, their	children, catholic, vote, ndp, racist, anything, agree, drag, moe, rights	hope, find, play, call, name, life, video, remember, music, night	Entertainment
1	to, on, they, be, if, at, my, just, in the, it is	parking, lanes, truck, turn, bus, circle, limit, winter, intersection, bikes	vote, protest, support, province, agree, fact, hate, argument, ndp, race	Politics
2	to, you, have, people, be, but, covid, at, my, do not	downtown, arena, property, homeless, river, better, housing, taxes, build, areas	downtown, taxis, build, property, rent, arena, service, give, increase, income	Downtown development
3	to, is, have, saskatoon, be, people, we, as, would, city	covid, masks, vaccinated, care, cases, government, hospital, risk, spread, pandemic	lane, traffic, cyclist, bus, intersection, merge, pass, limit, sidewalk, circle	Road conditions
4	to, is, they, have, be, at, people, would, or, your	snow, yard, water, house, bins, beavers, pets, fence, waste, little	vaccine, covid, care, hospital, wear mask, case, spread, risk, mandate, system	COVID-19
5	to, and, is, you, on, have, do, at, be, his	radio, song, glad, rock, saw, facebook, new, christmas, family, listen	restaurant, tip, buy, grocery, shop, customer, coffee, money, pizza, give	Food Services
6	and, for, do, them, on, if, be, your, cat, out	costco, used, restaurant, grocery, stores, burger, pizza, local, chicken, prices	beaver, garbage, bag, fire, vet, bin, garden, compost, walk, litter	Garbage
7	is, for, they, have, food, at, like, do, if, so	service, tip, better, sasktel, paid, employees, minimum wage, food, taxes, union	kid, church, teach, drag, gender, support, abortion, change, government, read	Gender-related issues.
8	to, and, is, teachers, it, education, be, do, kids, if	cops, evidence, court, charges, custody, sps, call, victim, maybe, arrest	police, arrest, court, evidence, alert, break, punch, find, steal, area	Crime
9	beaver, beavers, the, and, the beavers, the beaver, photos, do, be, river	teachers, schools, public, funding, province, system, average, grade, kid, tax	winter, tire, wind, ice, spring, road, pool, lawn, boat, fall	Weather conditions

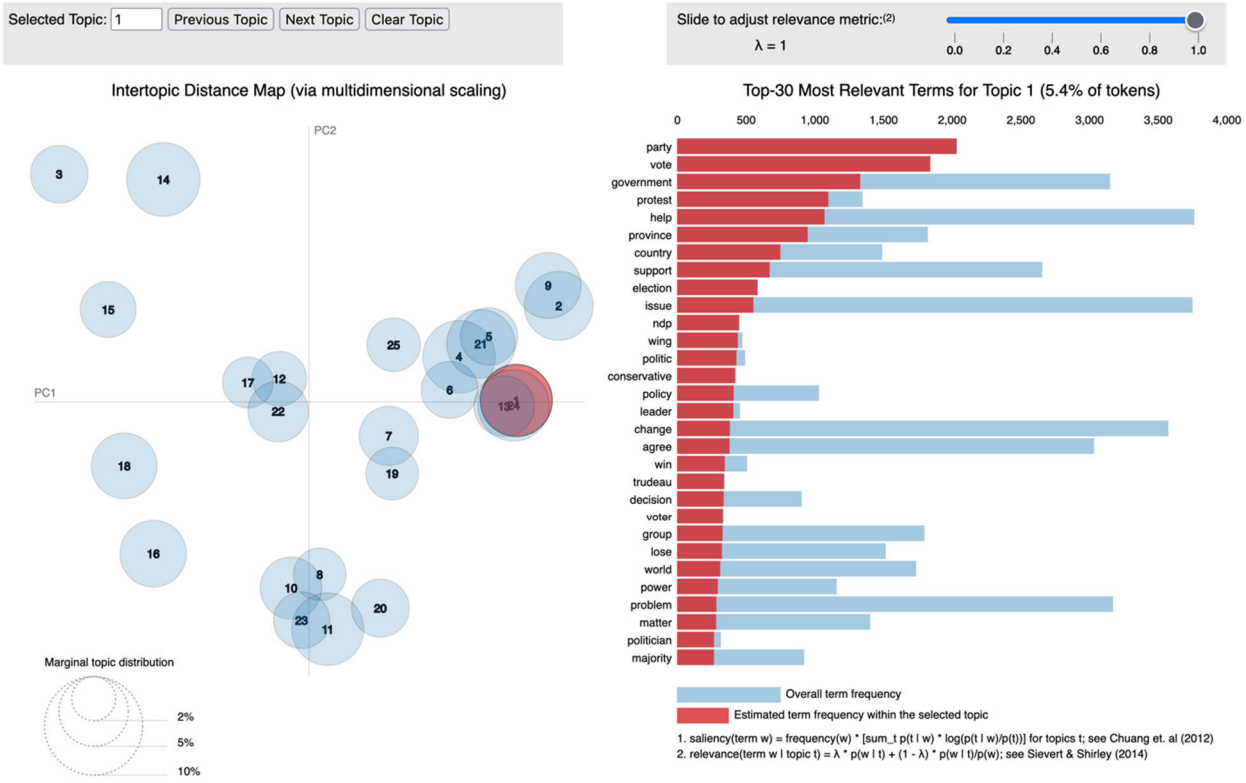

FIGURE 16. Intertopic distance map of LDA topics.

Comparing the C_V coherence score for 25 topics, the *BERTopic_UMAP_Kmeans* model achieved the best score of approximately 0.64. In terms of the UMass coherence metric, the NMF model performs best with a score of approximately -4.11. It performs slightly better than the *BERTopic_UMAP_Kmeans* model with a difference of

TABLE 12. Top 10 terms of 25 topics generated by LDA, NMF, and BERTopic_UMAP_HDBSCAN models.

LDA			NMF		BERTopic_UMAP_HDBSCAN	
Topic ID	Top 10 Terms	Topic Label	Top 10 Terms	Topic Label	Top 10 Terms	Topic Label
1	party, vote, government, protest, help, province, country, support, election, issue	Politics	help, change, issue, life, believe, problem, show, case, part, police	Public assistance	kid, teacher, protest, church, drag, party, support, call, gender, agree	Educational Influence
2	life, woman, man, give, believe, opinion, part, support, respect, race	Humanity	kid, parent, kid school, kid kid, parent kid, teacher, kid parent, adult, drag, school kid	Education	snow, addiction, tire, downtown, park, landlord, shelter, bridge, property, issue	Housing
3	turn, lane, speed, limit, move, traffic, pass, circle, merge, spot	Road Traffic	bike, lane, turn, road, driver, bike lane, cyclist, traffic, ride, speed	Road traffic	tax, tip, service, wage, realtor, climate, employee, increase, fuel, pay taxis	Services
4	mask, doctor, vaccine, risk, health, hospital, study, nurse, emergency, effect	Healthcare	mask, wear, wear mask, mask wear, mandate, face, covid, virus, spread, distance	COVID-19	smell, garbage, vet, restaurant, trap, litter, find, beer, bird, fry	Garbage
5	law, break, problem, health, person, deserve, case, point, rule, agree	violation	love, hate, photo, beaver, friend, picture, love love, pic, kid love, family	Pictures	lane, cyclist, traffic, turn, sign, bus, intersection, fly, flag, sidewalk	Road traffic
6	cat, hope, animal, die, joke, child, love, pet, kill, power	Animal	call, name, police, phone, cop, call police, call call, name call, call cop, call name	Police	covid, vaccinate, wear mask, hospital, care, risk, spread, mandate, flu, needle	COVID-19
7	job, name, miss, change, hate, list, location, remember, sub, find	N/A	find, hope find, friend, search, link, info, list, follow, facebook, stuff	Social media	bear, hair, moose, moe, jacket, buck, shoe, dress, pic, horton	Clothing
8	charge, dog, tip, property, experience, owner, service, staff, pay, abuse	Services	pay, tax, taxis, money, cost, pay taxis, property, attention, pay attention, service	Tax and services	camera, gallery, artist, pic, instagram, attention, vibe, dash, ad, billboard	Multimedia
9	call, word, person, evidence, crime, name, police, give, conversation, fact	Crime	drive, speed, driver, circle, truck, limit, vehicle, circle drive, road, speed limit	Road traffic	tent, rainbow, summer, morning, weekend, storm, date, moon, festival, couple	Seasons
10	police, price, garbage, phone, call, waste, find, report, beaver, rent	N/A	school, teacher, student, school school, church, education, fund, kid school, system, teach	Education	arena, radio, stadium, play, downtown, rock, concert, repeat, spend, library	Entertainment
11	money, pay, tax, cost, taxis, spend, dollar, gas, increase, save	Services	sound, winter, word, cat, wish, driver, sound problem, fun, mind, idea	N/A	trudeau, fsin, metis, cbc, move province, alberta, flag, bumper sticker, agree, hate	Politics
12	house, head, sign, cop, couple, person, call, stop, update, deal	N/A	downtown, move, store, buy, walk, area, house, park, grocery, food	Downtown	chime, neighbour, wave, ear, exhaust, sound problem, bark, idea, audacity, minute minute	Noise level
13	read, show, picture, medium, book, freedom, news, link, history, write	News	point, miss, view, point point, miss point, point view, prove, prove point, argue, agree point	Debate	cancel, restriction, culture, lift, censor, rating, ban, apply, choose support, block block	Social media restrictions
14	drive, driver, road, car, vehicle, stop, cyclist, traffic, rule, ticket	Road traffic	agree, disagree, snow, agree point, opinion, agree disagree, bit, protest, agree love, system	Protest	convoy, passport, invasion, world country, ottawa, imprison, cross border, conspiracy theorist, surprise	Protests
15	bike, snow, ride, stay, fly, sidewalk, winter, bike lane, lane, weather	Transportation	hope, hope to find, friend, hope help, share, family, help, bit, wish, stay	Family	campus, scrap, engineering, space, database, budget, resource, wifi, scale, idea	Education
16	buy, car, eat, stuff, business, restaurant, food, smoke, store, door	Services	give, money, ticket, give money, give ticket, break, vibe, credit, head, chance	Ticketing	booster, prize, congrat, weekend, testing, drive min, formula, kitty, hrs, age group	Awards
17	side, town, thought, weekend, enjoy, school, close, zone, market, buddy	Weekend	car, vehicle, winter, park, wash, car car, truck, hit, drive car, car wash	Transportation	fraction, hippie, stop spread, growth, scare, population control, baby boomer, terminology, increase number, argument	Aged population

TABLE 12. *(Continued.)* **Top 10 terms of 25 topics generated by LDA, NMF, and BERTopic_UMAP_HDBSCAN models.**

18	downtown, bus, parking, build, walk, building, store, minute, grocery, transit	Downtown	person, person person, turn, story, shame, matter, act, kill, internet, thought	Social media	ram, monitor, upgrade, express, gaming, truck, recommend, price quality, card, optimize	Pricing
19	water, issue, pick, parent, give, brain, area, page, call, family	N/A	vote, party, election, ndp, voter, voting, government, vote vote, province, conservative	Politics	datum, chart, approval, trend, percentage, gather, term term, lead, deflection, incline	Statistics
20	school, student, teacher, class, education, pay, teach, rate, sell, home	Education	read, book, story, write, read book, news, name, drag, read story, learn	News	menu, item, call fun, bar close, add list, excite, suspend, creep, category, uber	Food delivery
21	wear, mask, case, covid, spread, question, wear mask, mandate, vaccinate, restriction	COVID-19	stop, sign, stop sign, light, yield, stop stop, traffic, bus, intersection, law	Traffic	age group, toy, stagger, choice give, median, person report, validation, exist, booking, blow mind	Statistics
22	find, area, drug, night, park, crime, fire, issue, shelter, neighbourhood	Social problems	job, wage, lose, job job, hire, lose job, experience, tip, apply, worker	Employment	minimum wage, employer, standard, poverty, meet, manitoba, business pay, fair, rumor, earner	Income
23	business, wait, food, employee, event, line, pay, drag, company, wage	Services	care, health, health care, system, hospital, worker, care system, care care, dog, care worker	Healthcare	lens, arm leg, strap, twist, plot, tonight, sister, meeting, shape, map	Date
24	kid, child, parent, school, family, church, friend, adult, drink, group	Family	wait, minute, turn, line, list, emergency, doctor, bus, wait minute, room	Healthcare issues	zeller, target, close, haircut, expansion, launch, furniture, alternative, stock, management	Business
25	care, winter, point, love, watch, truck, photo, tire, learn, sound	Winter conditions	child, parent, teacher, abuse, parent child, family, child parent, drag, child abuse	Education		

approximately -0.61. However, comparing the topic terms generated by both models, the *BERTopic_UMAP_Kmeans* achieves better results than the NMF. The NMF model generates redundant topic terms (keywords) for almost all topics. This is because the NMF model considers terms with higher weights to represent a topic and does not consider the similarity and diversity of the terms in a topic. The BERTopic_*UMAP_Kmeans* model is selected as the final topic model to discover 25 topics from the Saskatoon subreddit comments.

Compared to the conventional topic models, NMF and LDA, the modularity nature of BERTopic allows the customization of the topic model. For instance, representation models for keyword extraction, part of speech, and diverse keywords can be leveraged to further fine-tune topic terms for accurate topic representations. To address the issue of redundant keywords (terms) found in topic representations, we utilized the maximal marginal relevance (MMR) as the representation model to improve diversity and reduce redundancy in topics generated with the *BERTopic_UMAP_Kmeans* model. We set the MMR diversity threshold to 0.5 to fine-tune topic representations to improve diversity in the topic terms.

Fig. 16 to 19 illustrate the intertopic distance map of 25 topics generated by LDA, NMF, BERTopic_UMAP_HDBSCAN, and BERTopic_UMAP_Kmeans model respectively. As illustrated in the figures, each circle denotes a topic. The size of the circle represents the prevalence of the topic within the collection of documents.

In Fig. 16 and 17, the bars represent the top 3- most relevant terms for a topic with their percentage in a topic. The blue bar represents the overall term frequency within the collection of documents. The red bar denotes the estimated term frequency within a selected topic. In Fig. 18 and 19, hovering over a circle gives the size of the topic and its corresponding top 5 terms.

The top 10 terms for 25 topics generated by the BERTopic_*UMAP_Kmeans* model are shown in Table 13. From Table 13, It can be observed that the BERTopic with K-means clustering generates coherent and meaningful topics compared to the other topic models. The BERTopic, by default returns topics discovered in a descending order based on their frequency. The topic ID starts from 0 to the highest. We manually analyzed the top 10 terms to assign labels to the topics generated.

Fig. 20 shows the distribution of comments for each topic generated by the *BERTopic_UMAP_Kmeans* model. The top 6 prevalent topics discovered in the Saskatoon subreddit are Topic 0 (Educational influence), Topic 1 (social media), Topic 2 (Road traffic), Topic 3 (Transportation), Topic 4 (Pictures), and Topic 5 (Protest). The educational influence topic (Topic 0) is the most dominant topic discussed, with 7.09% of comments. It highlights discourse on gender affirmation,

TABLE 13. Top 10 terms of 25 topics generated by BERTopic_UMAP_Kmeans on fully preprocessed Saskatoon Subreddit corpus.

Topic ID	Top 10 Terms	Topic Label	Proportion of Topic, (%)
0	kid, teacher, church, drag, gender, abortion, government, rainbow, belief, issue	Educational Influence	6,035 (7.09%)
1	read, point, laugh, miss, word, find, facebook, hate, person, attention	Social media	5,297 (6.22%)
2	traffic, stop, pedestrian, intersection, merge, speed limit, signal, pass, bike lane, yield	Road traffic	4,913 (5.77 %)
3	bus, truck, parking, ticket, transit, minute, walk, stop, tesla, damage	Transportation	4,891 (5.75%)
4	head, camera, pic, morning, light, turn, line, couple, shoe, memory	Pictures	4,688 (5.51%)
5	protest, racist, freedom, woman, support, flag, argument, violence, community, agree	Protest	4,572 (5.37%)
6	downtown, arena, park, stadium, winnipeg, library, money, apartment, centre, grocery store	Downtown development	3,935 (4.62%)
7	covid, spread, mandate, risk, rate, restriction, care, face, immunity, mask wear	COVID-19	3,846 (4.52%)
8	province, government, union, shelter, tax, benefit, degree, issue, staff, minimum wage	Minimum wage and taxation	3,822 (4.49%)
9	police, arrest, court, call, evidence, punch, alert, steal, break, find	Policing	3,816 (4.48%)
10	troll, conversation, name, report, internet, support, argue, lie, change, evidence	Misinformation	3,791 (4.45%)
11	nurse, emergency, addiction, patient, call, health care, needle, government, problem, appointment	Healthcare	3,599 (4.23%)
12	store, tip, grocery, customer, staff, fee, value, profit, village, wage	Pricing of Goods and Services	3,549 (4.17%)
13	property, rent, realtor, insurance, increase, income, road, pay taxis, mortgage, move	Housing	3,446 (4.05%)
14	restaurant, eat, drink, fry, noodle, breakfast, buy, menu, mcdonald, donut	Food Services	3,368 (3.96%)
15	call, move, review, event, moe, play, list, facebook, update, block	Social media	3,093 (3.63%)
16	family, expect, heart, kid, fear, wake, halloween, luck, cry, bubble	Family Units	3,064 (3.6%)
17	paint, pool, fence, garden, summer, prairie, area, build, park, trail	Housing facilities	2,796 (3.29%)
18	bag, waste, smell, compost, litter, throw, smoker, landfill, recycle, paper	Garbage and Recycling	2,696 (3.17%)
19	cat, beaver, trap, owner, yard, mouse, wind chime, animal control, leash, insurance	Animal control	2,687 (3.16%)
20	party, ndp, trudeau, politic, country, vote vote, give, platform, wing, majority	Politics and Election	2679 (3.15%)
21	winter, ice, winter tire, rain, shovel, bridge, vehicle, layer, paint, humidity	Winter weather condition	2,273 (2.67%)
22	listen, wonder, jazz, radio station, vibe, folk, spotify, song play, bombargo, festival	Radio and Entertainment	1,859 (2.18%)
23	garbage, winter, bag, pizza, smell, layer, plant, smoker, drone, litter	Garbage and Recycling	229 (0.27%)
24	laugh, cat, song, rock, listen, coverage, playlist, radio station, beaver, issue	Radio and Entertainment	169 (0.2%)

TABLE 14. Number of topics with their coherence score for the BERTopic_UMAP_Kmeans model.

Theme	Topic Number	Positive Sentiment (%)	Negative Sentiment (%)	Number of Comments, (%)
Online Community Engagement	1, 10, 15	4,614 (37.88%)	7,567 (62.12%)	12,181 (14.31%)
Transportation and Weather Conditions	2, 3, 21	3,829 (31.70%)	8,248 (68.30%)	12,077 (14.19%)
Cost of Living	8,12, 14	4,267 (39.73%)	6,472 (60.27%)	10,739 (12.62%)
Educational Influence and Family Values	0,16	3,189 (35.05%)	5,910 (64.95%)	9,099 (10.69%)
Healthcare	7,11	2,331 (31.31%)	5,114 (68.69%)	7,445 (8.75%)
Protest and Political Elections	5, 20	1,708 (23.56%)	5,543 (76.44%)	7,251 (8.52%)
Tourism and Entertainment	4, 22, 24	2,996 (44.61%)	3,720 (55.39%)	6,716 (7.89%)
Housing and Facilities	13,17	2,433 (38.98%)	3,809 (61.02%)	6,242 (7.33%)
Downtown Development	6	1,730 (43.96%)	2,205 (56.04%)	3,935 (4.62%)
Policing	9	845 (22.14%)	2,971 (77.86%)	3,816 (4.48%)
Garbage and Recycling	18,23	971 (33.20%)	1,954 (66.80%)	2,925 (3.44%)
Animal Control	19	1,073 (39.93%)	1,614 (60.07%)	2,687 (3.16%)

funding education, religion-owned schools, and the role of parents in education. Topic 23 (Garbage and Recycling), and Topic 24 (Radio and Entertainment) have the lowest proportion of comments at a value less than 1%.

V. THEMATIC ANALYSIS

The 25 topics generated by the best-performing model, *BERTopic_UMAP_Kmeans*, were categorized into 12 themes. Thematic analysis is performed to merge the related topics

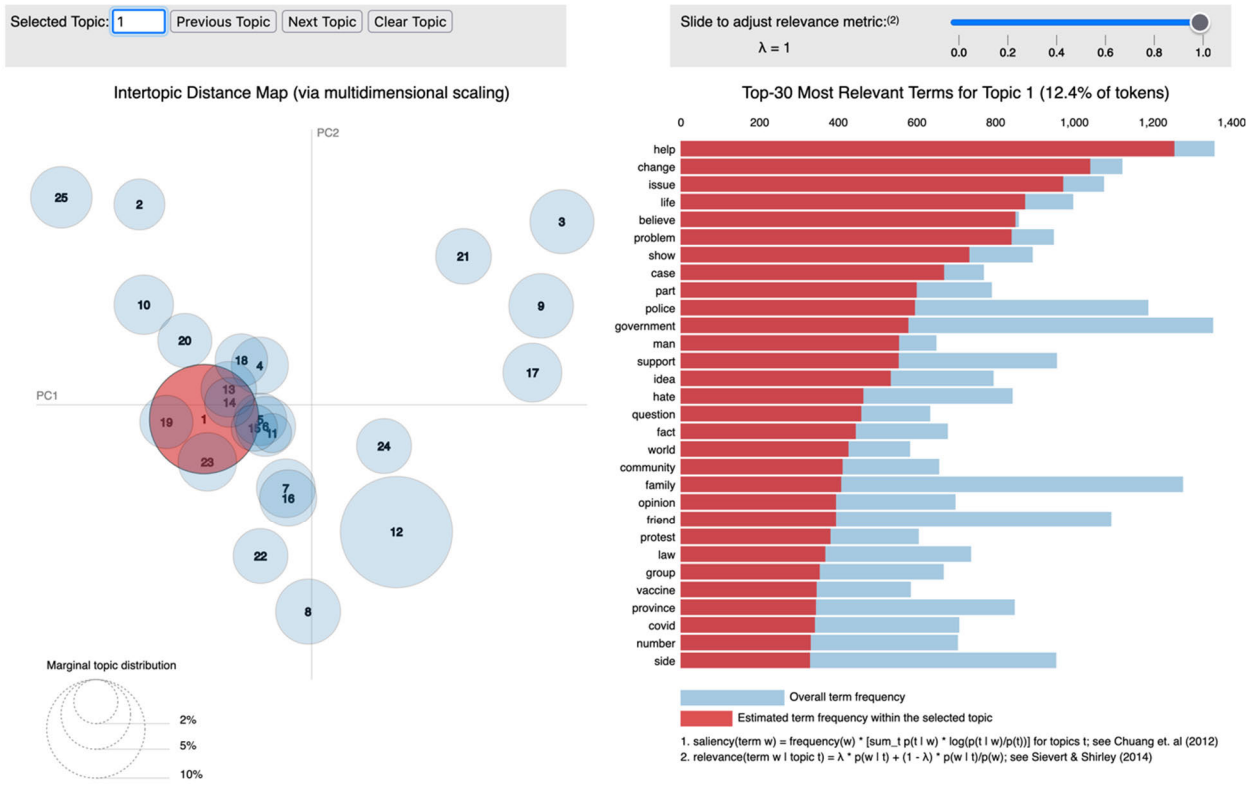

FIGURE 17. Intertopic distance map of NMF topics.

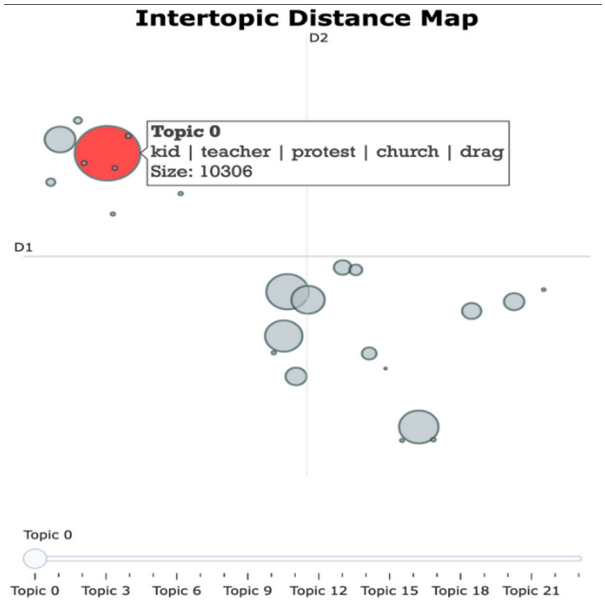

FIGURE 18. Intertopic distance map of BERTopic_UMAP_HDBSCAN topics.

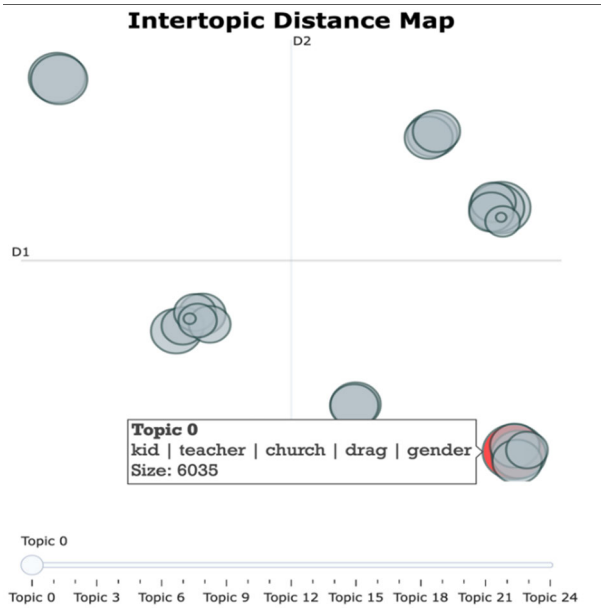

FIGURE 19. Intertopic distance map of BERTopic_UMAP_Kmeans topics.

into themes. As shown in Table 14, we merge the topics with the same label into a theme as they have common topic terms. For instance, Topics 22 and 24 have the same topic label. We analyze the topic similarity matrix (heatmap) to find the relationships between topics. The topic similarity matrix is computed by finding the cosine similarity scores of the topic embeddings.

A higher similarity score indicates a stronger relationship between topics. We consider topics with a similarity score

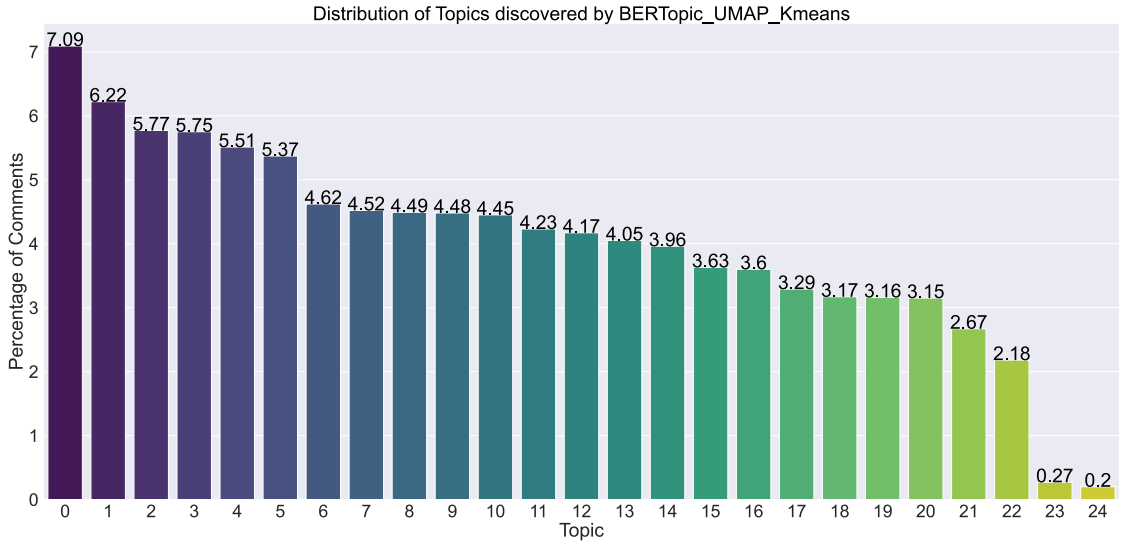

FIGURE 20. Distribution of topics generated by BERTopic_UMAP_Kmeans model.

above 80 as highly related. As shown in Fig. 21, Topics 18 and 23 are highly correlated with a similarity score of approximately 0.88. Fig. 22 displays the relationship between topics generated by the *BERTopic_UMAP_Kmeans model*. The BERTopic extracts the hierarchies of the topics generated to understand how topics are related. Related topics are grouped into the same cluster. Fig. 23 shows the hierarchical clustering of topics discovered by the *BERTopic_UMAP_Kmeans* model. The hierarchical clustering of topics is based on the cosine similarity scores between topic embeddings. In visualizing the hierarchy, topics in the same cluster are assigned the same color.

As shown in Fig. 23, the topics 2, 3, and 21 are in the same cluster. We merged these topics to form the *'Transportation and Weather Conditions'* theme. Although BERTopic automates the process of merging related topics, human interpretation is required as some combinations may not be logical. We manually analyze the top terms and representative comments of each topic to confirm how related the topics are. Table 14 shows the themes with their corresponding topics and the distribution of comments.

The top 5 themes discovered in the Saskatoon subreddit are online community engagement, transportation and weather conditions, cost of living, educational influence and family values, and healthcare. The top 5 themes collectively make up approximately 60.56% of the cleaned comments used in the study. The online community engagement theme is the most prevalent theme with 12,181 comments, followed by the transportation and weather conditions theme, which represents approximately 14.19% of the overall comments. The Garbage and Recycling, and animal control themes were the least discussed themes on the Saskatoon subreddit, with an overall distribution of less than 8%.

We observed that the discussions across all themes were minimal in 2019. Overall, there was a high level of engagement across all themes in 2023, except for the *'Healthcare'* theme. The discourse on the *'Healthcare'* theme was higher in 2021.

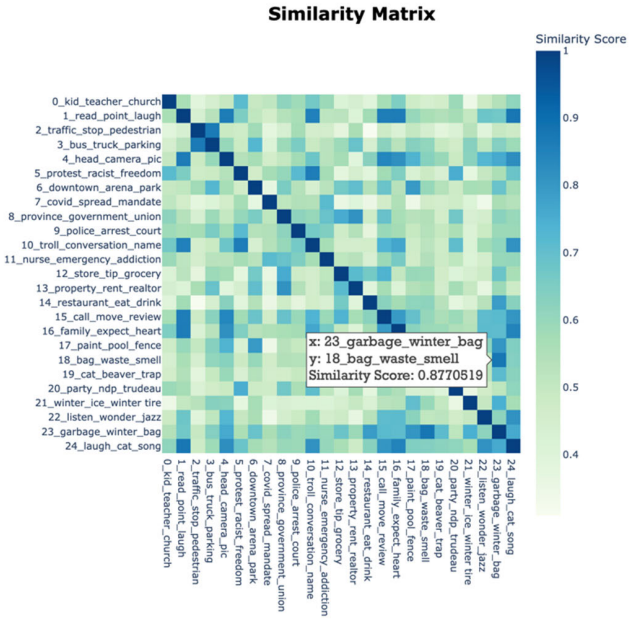

FIGURE 21. The similarity score between topics 18 and 23 of the BERTopic_UMAP_Kmeans model.

Fig. 24 illustrates the evolution of the ranks of the top 5 themes over the study period. In the first seven months of 2019, the *'transportation and weather condition'* theme was the hottest theme discussed on the Saskatoon subreddit.

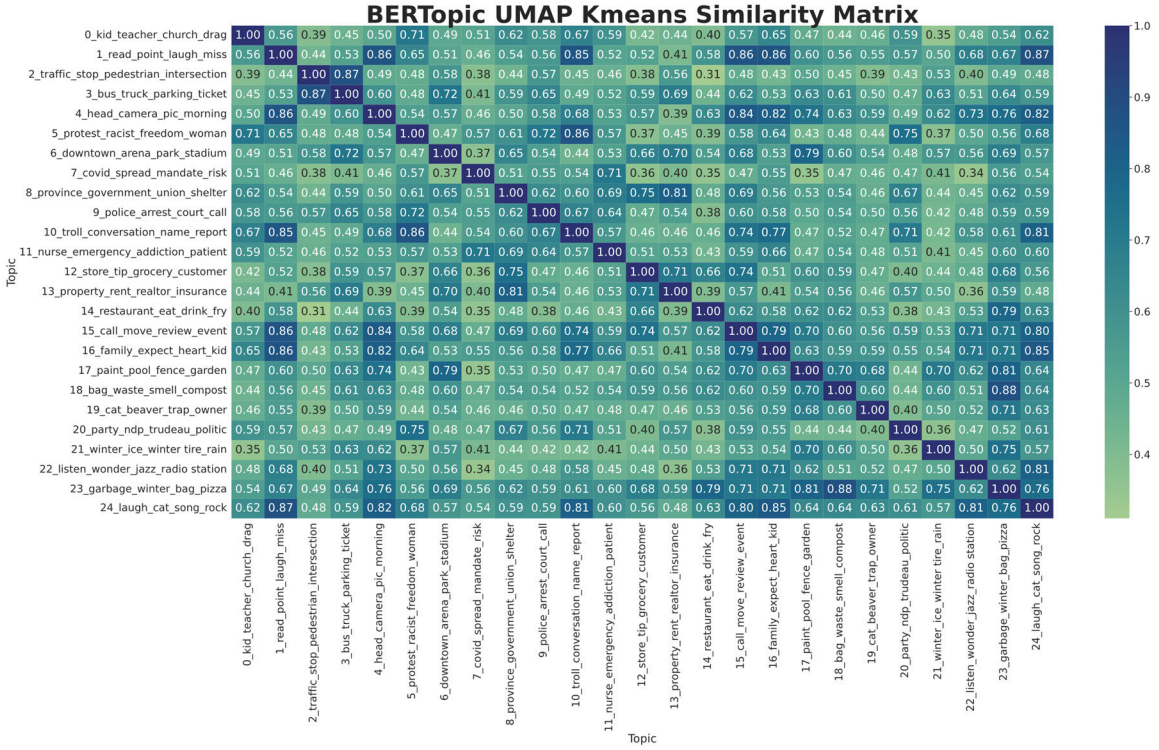

FIGURE 22. Similarity matrix of BERTopic_UMAP_Kmeans model.

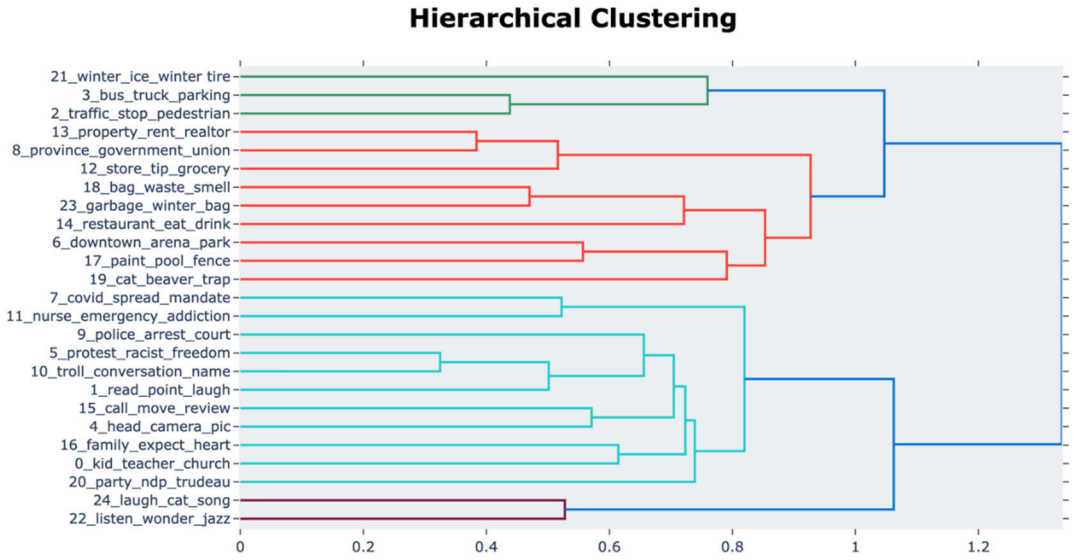

FIGURE 23. BERTopic_UMAP_Kmeans hierarchical clustering.

The *'cost of living'* theme remained at the second position from January 2019 to March 2019, and from July 2019 to September 2019. The *'cost ofliving'* theme ranked first in December 2019, and dropped to the fifth position in February 2020. The *'educational influence and family value'* theme started at the fourth position in the early months of 2019 and drastically moved to the first position in August 2019. The *'educational influence and family value'* theme remained at the fifth position from July 2021 to January 2022. The *'online communityzengagemen'* theme started at the third position in the first three months of the study period. The them's position fluctuated between first and second positions from September 2019 to November 2019. The *'online community engagemen'* theme was most prevalent in 2021 followed by

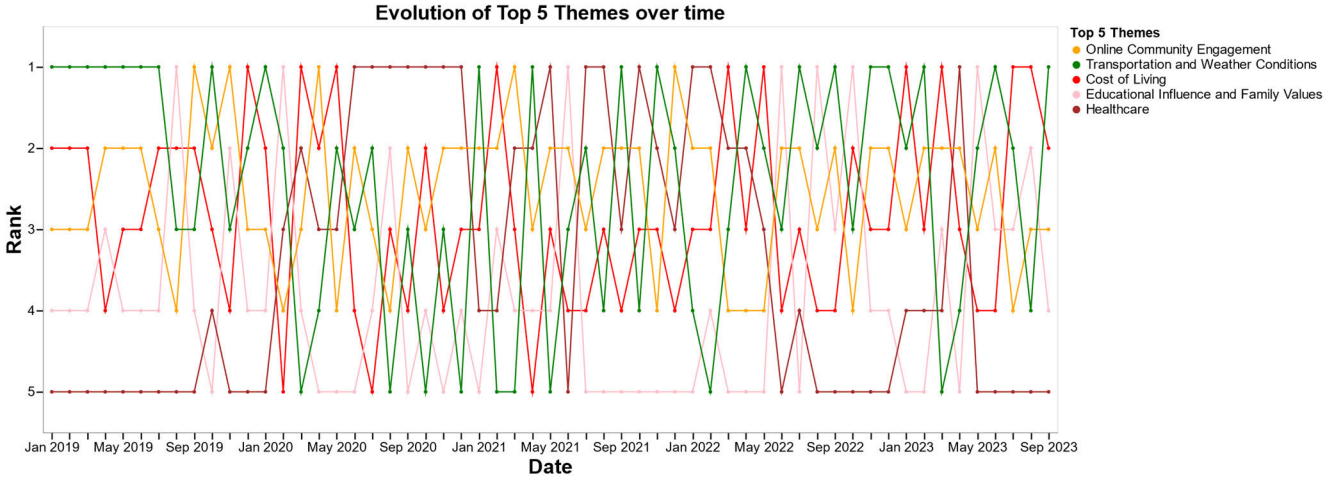

FIGURE 24. Rank of top 5 themes over time.

the *'healthcare'* ztheme. The *'healthcare'* ztheme remained at the fifth position in the first nine months of 2019.

The discourse on the theme steadily increased from February 2020, causing the *'healthcare'* theme to be prominent from June 2020 to December 2020. At the end of the study period the most prevalent theme was the *'transportation and weather condition'* theme followed by the *'cost of living'* theme. The *'online community engagemen'*, *'educational influence and family value*s, and *'healthcare'* zthemes were ranked the third, fourth, and fifth position respectively.

VI. SENTIMENT ANALYSIS

We used sentiment classifiers such as VADER, SiEBERT, and four supervised ML models namely Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost) to determine the sentiments expressed on the Saskatoon subreddit. We manually labeled 1,500 comments randomly sampled from the original dataset as ground truth to evaluate the performance of the sentiment classifiers. The labeled dataset is made up of 946 negative comments, 426 positive comments, and 125 neutral comments. We discarded the neutral comments as our focus is on identifying the positive and negative sentiments influencing the discovered themes.

For sentiment analysis with VADER, and SiEBERT, we maintained the stopwords and punctuations in the comments, as they contribute to the overall polarity of text in sentiment analysis [41]. However, for the supervised ML models, a fully preprocessed dataset is used in training.

We used the *"siebert/sentiment-roberta-large-english"* pre-trained BERT model from the Hugging Face's sentiment analysis pipeline[5] to predict sentiments of the comments with the SiEBERT model. In predicting sentiments with VADER, the compound score threshold for the positive and negative sentiments was set to +0.05 and -0.05 respectively.

[5]https://huggingface.co/siebert/sentiment-roberta-large-english

In the case of the supervised ML models, we transformed the labeled dataset to numerical vector representations using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique. The Synthetic Minority Over-Sampling Technique (SMOTE) is applied to handle our imbalance annotated dataset. We used a stratified 10-fold cross-validation to train and evaluate the ML models.

We used Accuracy, Precision, Recall, and F1-Score as evaluation metrics to assess the performance of the sentiment classifiers. The evaluation metrics are expressed in Equations (4) – (7).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - score = 2\frac{Precision * Recall}{Precision + Recall} \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where TP, FP, TN, and FN mean True Positive, False Positive, True Negative, and False Negative respectively.

The performance of the sentiment classifiers in each class is shown in Table 15. In the prediction of negative comments, SiEBERT was the best-performing model with a precision score of 94% and an F1-score of 94%. The Random Forest model performed best in the prediction of positive comments with a score of 90% and 84% for precision and F1-score respectively.

The overall precision, recall, and F1-score are based on the weighted average of precision, recall, and F1-score for each class of sentiment. The weighted average of each evaluation metric as expressed in equation (12) is computed by finding the average, where each metric's value is weighted by the number of instances for each sentiment class.

$$Weighted\ Average = \frac{1}{\sum_{l \in L} |y_l|} \sum_{l \in L} |y_l| \emptyset \left(y_l, \hat{y}_l\right) \tag{12}$$

TABLE 15. Performance of sentiment classifiers in each class.

Classifier	Sentiment	Precision (%)	Recall (%)	F1-Score (%)
VADER	Negative	90	48	62
	Positive	48	77	59
SiEBERT	Negative	94	90	92
	Positive	79	88	83
XGBoost	Negative	77	83	80
	Positive	81	76	78
Logistic Regression	Negative	84	85	86
	Positive	85	84	85
Decision Tree	Negative	76	72	74
	Positive	74	77	75
Random Forest	Negative	82	91	87
	Positive	90	80	84

TABLE 16. Overall performance of sentiment classifiers.

Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
VADER	77	57	61	57
SiEBERT	90	89	89	89
XGBoost	79	79	79	79
Logistic Regression	85	85	85	85
Decision Tree	75	75	75	75
Random Forest	86	86	86	86

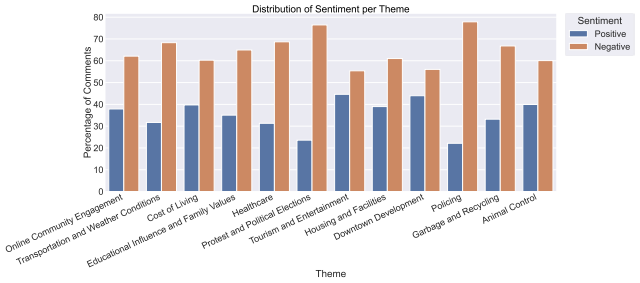

FIGURE 25. Distribution of sentiments for each theme.

where L is the set of sentiment classes, $\hat{y}$ is the predicted sentiment class, y is the true sentiment class, y_l is the subset of y with sentiment class l, and $\emptyset\left(y_l, \hat{y}_l\right)$ computes the precision, recall, or F1-score for the true and predicted sentiment classes that have the sentiment class l.

Table 16 shows the overall performance of the sentiment classifiers. The SiEBERT model outperformed the supervised ML models and VADER with a precision score of 90%. It achieved a score of 89% across recall, F1-score, and accuracy metrics. The lexicon-based classifier, VADER had the lowest performance with a score of 57% for both recall and accuracy and an F1-score of 61%. In terms of supervised ML for sentiment prediction, the Random Forest was the best model with a score of 86% across all evaluation metrics. SiEBERT is a transfer learning model based on BERT trained on diverse datasets in various domains to enhance its generalization on new data. It uses a self-attention mechanism to represent text as contextual embeddings to enhance its performance. VADER relies on pre-defined lexicons which may be outdated and not fully acknowledge modern vocabularies. On the other hand, supervised ML models require a large amount of annotated data to perform well. Supervised ML models are subject to overfitting which results in poor generalization to new data.

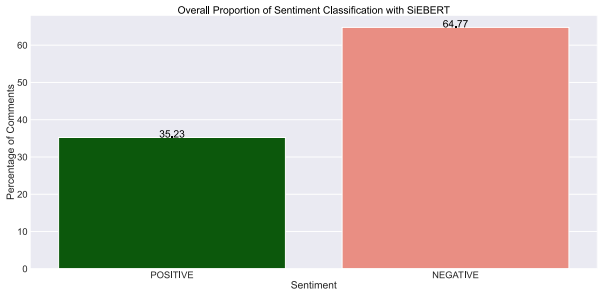

FIGURE 26. SiEBERT sentiment distribution.

We leveraged the SiEBERT model to determine the sentiments of the final dataset used in discovering the themes. The number of comments for each sentiment class for each theme is shown in Table 14. Fig. 25 illustrates the percentage of sentiment for each theme. Negative sentiment was prominent in each theme. However, positive sentiments were predominant for certain themes at various points in the study period. In the '*cost of living*' theme, positive sentiments were prevalent in May 2021, and from July 2023 to the end of the study period. For the '*tourism and entertainment*' theme, positive sentiments were highly expressed in the early months of 2019, February 2020, and June 2021. Similarly, positive sentiments were dominant in the early months of 2019 and towards the end of 2019, from the beginning of the period to March 2020, and from May 2021 to June 2021 for the 'educational influence and family values' theme. For the '*housing and facilities*' theme, positive sentiments were only highly expressed in January 2022. In the '*downtown development*'

theme, positive sentiments were prevalent in the last four months of 2019 and the spring of 2021.

The overall distribution of sentiments for the collected Saskatoon subreddit comments is shown in Fig. 26. The dominant sentiment in our Saskatoon subreddit dataset was negative with a proportion of 64.77%, and the distribution of positive sentiment was 35.23%. Fig. 27 shows the evolution of sentiments over time. Both positive and negative sentiments exhibit similar trends, but negative sentiments were prevalent throughout the study period. The total number of comments in January 2019 was below 400. It can be deduced that the level of Saskatoon citizens' engagement on the Reddit platform in 2019 was low compared to the successive years in the period of data collection. An upward trend of engagement was observed in March 2020. This was the year, WHO declared the COVID-19 pandemic. The number of negative and positive sentiments expressed increased from November 2020 but declined in December 2021. The number of comments for both sentiments fluctuated from January 2022 to June 2023. The negative sentiments dramatically increased in August 2023. The high number can be attributed to discourse on dangerous intersections, cyberbullying, the cost of goods and services, parental inclusion and consent in education, and housing and facilities. At the end of the period, which was the third week of September 2023, the number of comments for both positive and negative sentiments expressed was higher than the entire number of comments in 2019.

VII. DISCUSSION OF RESULTS AND RECOMMENDATIONS

A. MAIN FINDINGS ON TOPIC MODELS

In this study, we compared the performance of well-known topic models namely LDA, NMF, and BERTopic in the discovering of topics from Saskatoon subreddit comments. The topic models were evaluated based on their topic coherence score. The LDA model was trained on a BoW-transformed corpus and the NMF model was trained on a TF-IDF corpus. The BERTopic utilizes a pre-trained sentence transformer to create vector representations of the document. From our experiments, we observed that the LDA model generated incoherent topics compared to the other topic models. This LDA focuses on the frequencies of words across the document and ignores the semantic relationship between words. Due to this, LDA may generate dissimilar topic words, which makes interpretation difficult. On the other hand, the NMF exhibited a fair performance as most of the topic terms generated were redundant. This is because the NMF model considers terms with higher weights to represent a topic and does not consider the similarity and diversity of the terms in a topic. Redundant topic terms may lead to inaccurate topic representations.

The BERTopic leverages transformers, clustering techniques, and a class-based TF-IDF to generate coherent topic representations. The default BERTopic leverages the HDBSCAN for clustering and can automatically find the number of topics in a given corpus, unlike LDA and NMF which require the number of topics to be specified. However, in our studies, we observed that the default BERTopic generated many outliers, which were found to be relevant after manually analyzing them. Hence, to force every text in our dataset to be included in a cluster we replaced the HDBSCAN with K-means clustering. With traditional topic models, the dataset needs to be thoroughly preprocessed.

However, for BERTopic minimal preprocessing is required as it leverages sentence transformers to process texts. In this study, we evaluated the performance of the variations of BERTopic based on the degree of data preprocessing. We observed that the two variations of BERTopic performed well when trained on a fully preprocessed dataset. The BERTopic with K-means clustering obtained a higher coherence score than the default BERTopic model. To address the redundant topic terms, we leveraged the Maximal Marginal Relevance (MMR) algorithm to fine-tune topic representations to create diverse topic terms for the BERTopic with K-means clustering.

There are some limitations of this study. Topic Models used in this study perform better on short texts than long texts. Reddit comments are limited to 40,000 characters. Long text in our corpus may likely be inaccurately represented. In the future, we will evaluate the performance of topic models on longer texts and explore different truncation methods for longer texts.

B. DISCUSSION OF THEMES

1) ONLINE COMMUNITY ENGAGEMENT

The online community engagement theme has 37.88% and 62.12% positive and negative messages respectively. The theme represents 14.31% of all messages.

Overall, people see the online community of the city as a forum for sharing messages of hope and addressing others' concerns. Citizens pose questions and online community members respond quickly to the best of their knowledge. A user responded to a query…**C114**: *I would suggest contacting the city and asking. I provided a link, their contact info is on the right. I hope you can accomplish it so you can all rest a little easier. Best of luck. [2023-08-27].*

Members also provide information on how to access city services and resources through the website. Other important issues such as garbage collection dates and schedules are all discussed in these online communities. Another benefit for users is the recommendations for cheaper services and goods in the city as people share prices, product reviews, and their experiences with service providers.

However, there are more negative comments, a lot of which focus on city service improvement. Users complain about the difficulty of viewing certain information from the city website. According to a user…***C127****: a lot of the useful tools are somewhat obscured or hard to find. for example, the building permit lookup search tool is buried in a paragraph of text with a dark green, tiny hyperlink. or the pothole reporting map tool is decently buried. there is a lot of useless text and*

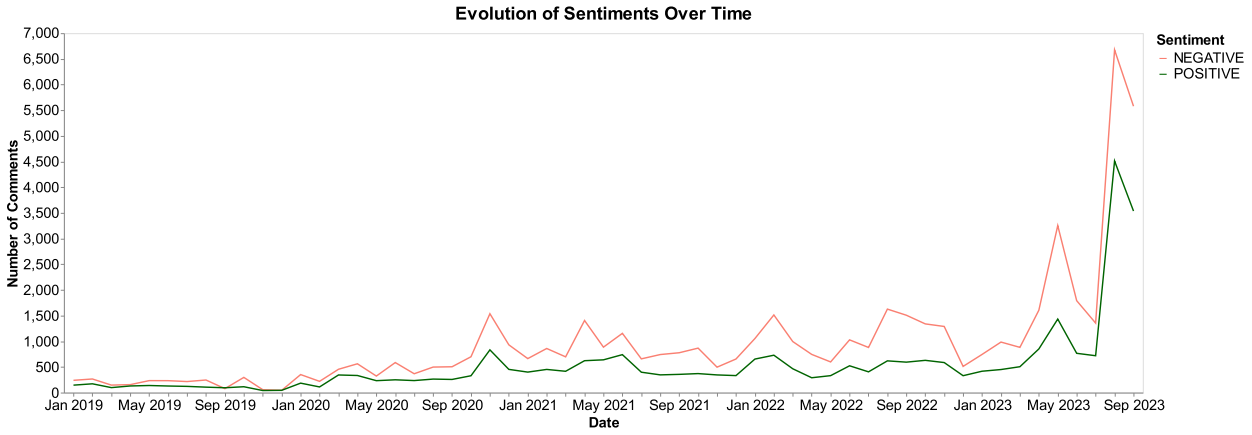

FIGURE 27. Evolution of sentiments over time.

information that could be cleaned up a lot and made to look more aesthetically pleasing and user-friendly. [2022-07-15].

There are also vindictive comments, intolerance exhibited by others, and trolling. Some people also spew misinformation on topics such as the city's energy supplier and COVID-19 vaccination concerns.

Recommendation: We recommend that the city improve its online outreach by revising and updating its website. Commonly accessed information and frequently asked questions should be readily available. Also, users are encouraged to provide honest reviews about the city and its services as that is vital for boosting tourism.

2) TRANSPORTATION AND WEATHER CONDITIONS

The transportation and weather theme dominated 14.19% of the total comments. Approximately 31.70% of the comments are positive and 68.30% are negative.

One major issue on this theme is dangerous intersections around the city although some residents are beginning to see an improvement. A user wrote…***C2113****: I remember when this intersection didn't even have a light - and the church and school had to fight just to get one put in. It's always been a dangerous intersection but it's much safer now than it used to be. [2023-08-29].*

However, there are several complaints about merging issues on highways in the city. Others have issues with the stop signs and the speed limits. A remark reads…***C2123****: part of the problem is that this interchange design was designed in the 40s and this particular one was built in the 60s when Saskatoon's population was around 110,000. the other part of the problem is the "slip lanes", where traffic weaves to enter is woefully short. [2023-08-29].* There are so many posts with people complaining about poor lighting of the roads and improving the road network to meet the population growth of the city.

A second issue of discussion is public transit in the city. While residents praised the courtesy of some drivers, others noted the scheduling and connections. A comment reads…***C2121****: My bus home only runs every 40 minutes - I always ask the downtown bus driver to ask the connecting bus driver to wait for me at Place Riel - they always do. [2023-08-23].*

But most people have issues with unruly riders towards older and disabled riders. Further, there is a need to add more benches to bus stops to support cane users and disabled people. Most riders experience unnecessarily long delays in bus schedules. A user said…***C2222****: I hate how a destination that is 8 to 10 minutes away by car is 40 minutes away via transit. [2023-06-01].* Others have concerns that the transit app malfunctions. Others must wait for hours in wintery conditions for the bus due to delays, and weekend schedules are worse.

In the winter season, there are suggestions that Saskatchewan should make winter tires mandatory. Additionally, it was suggested that drivers be given a rebate or reduced plate insurance costs if they provide a receipt for winter tires. This may entice people and would help lower the cost of winter collisions. A user wrote…***C2321****: It continues to blow my mind that winter tires aren't made mandatory by law like they are down east. [2021-11-18].*

Residents also discuss a lot regarding bike lanes in the city. There is excitement over the fact that some areas are seeing bike lane development. Some areas are also extending the bike lanes A comment read…***C2412****: Having worked for the city during the asinine year of those horrendous bike lane changes, this made my day! [2023-08-25].*

However, some residents are calling for better bike lanes. One user remarked…***C2421****: The west side needs better pedestrian/cycling infrastructure. Especially close to St. Paul's. [2023-09-07].* Others suggested that the city needs to make sure that bike lanes are separate from roadways for safety. And others remarked on how some cyclists tailgate and ride speedily.

Recommendation: The city must improve the design of road intersections or re-develop the trouble spots on the road. Putting streetlights where necessary as suggested by residents

is needed for the safety of motorists, riders, and neighborhoods. The public transit system also needs consideration, especially the scheduled times. Moreover, bike lanes can be added to the city but there must be strategic enforcement of rules for bike riders.

3) PROTEST AND POLITICAL ELECTIONS

The thematic discussions on protests and political elections cover 8.52% of the total comments analyzed. About 23.56% of the comments were positive while 76.44% were negative.

On the positive, some residents of Saskatoon were happy about the political changes and interventions. Political promises of ''parental inclusion and consent'' for children are welcomed by some parents. The deliberate effort for inclusivity is also praised by some. A comment reads…***C313***: *correct, they fell short of 2022 targets, but some gains were made. now the SPS will undergo a diversity, equity, and inclusion audit that will encompass the forces of human resources policies as it looks to improve the diversity of its staff. …I don't care what sex/race the hires fit into, I just want the most qualified candidate chosen. [2023-08-27].* Other people also acknowledged Canada's social democracy tenets and believe that both left and right issues can be collectively discussed. There are also discussions on the enactment of bills to have a protest-safe buffer zone around medical treatment facilities. This is the response to the NDP's proposal to ban pro-life protests near abortion clinics.

On the negative side, most residents were livid about the rollercoaster of confusion and uncertainty around the COVID-19 situation. Others want political parties to make their manifestos clear indicating what they will do for the city and province and end the culture of giving empty promises. Others have issues with provincial budget cuts and how surplus funds were allocated for unworthy projects. There are also concerns about how the NDP ignores other minority groups and immigrants in the public system in Saskatchewan. Others want to see more improvement in the failing rural healthcare and education across the province. Others also complain about political parties trying to clamp down on protests. A comment reads…***C3211***: *it's already illegal to harass someone. this is an unnecessary limitation on freedom of speech. whether or not an opinion is popular or even wrong shouldn't give the government the authority to quash demonstrations. hard pass. [2021-05-16].*

Recommendation: Though the government and political parties cannot please every citizen, they can certainly engage the citizens more. They can change their political undertones to be more inclusive and propose clear plans for the welfare of the citizens. Issues around taxation and budgetary allocations can be done transparently with proper justifications. This will increase confidence in the political parties by the citizens.

4) COST OF LIVING (GOODS, SERVICES, AND TAXES)

This theme has 39.73% positive sentiments and 60.27% negative sentiments. The theme constitutes 12.62% of total comments and is focused on the cost of goods, services, and taxes.

Some residents are committed to promoting local businesses online and are pointing people to the shops where they can get their preferred goods and services. These local businesses are willing to price match when necessary. Other recommendations are for local and small businesses to take advantage of mobile apps and online platforms to sell to reduce operating costs. Additionally, some residents are happy about the pricing of certain items. A comment read…***C413***: *they have some of the most surprising food items there. I got some very fancy flaked finishing salt there. decent prices generally. and a bunch of the seasonal stuff that's not so great a price always gets marked down when there is a ton left - have got some massive winter safe ceramic planters for like 5 and 10 dollars, which would cost at least 50 for that size at dutch growers as an example. [2022-01-23].*

On the downside, a lot of people bemoaned the rising cost of goods and services. People opined that they are being charged more for everything with less government assistance. A user posted…***C421***: *gas, groceries, leisure. everything costs so much more, and yet, our wages will never be on par with costs. I feel just as poor as I did making 11.50 in 2007 when I was like 20, now I'm making like 20$/hr and I still feel like I can't do anything, all I can do is budget and hope my paycheck gets me to the next. so sick of this. [2022-04-21].* Others also noticed that products keep reducing in size for the same prices and dubbed it as shrink-flation. Also, some of the local businesses find it hard to compete with major retailers such as Walmart on the pricing of goods. Others mentioned how they could get similar items from online e-commerce sites such as Amazon for half the price. There were also concerns about the worsening quality in the service industry, e.g., restaurants.

Some residents suggest that income is based on one's educational qualifications, especially offers outside of retail. A post read…***C1011***: *if you intend to come here as well make sure your degree will be worthwhile here. many people move with degrees, and they end up working in minimum wage jobs because their degree isn't recognized in Canada. [2023-04-16].*

Others bemoan that the province and city have the lowest minimum wage in the country. Some unions are seeking a 20% increment such as teachers. However, there is the expectation that other public sector workers will also have major justifications for huge pay increases. Others complain about the high taxes and how managers in retail cut their working hours.

Recommendation: This theme can be addressed through public-sector engagement and private-sector partnerships. It is inconceivable that the city will just announce an increase in minimum wage. However, the city can provide a clear roadmap on how to improve the minimum wage and taxation requirements for residents. The high prices of goods and services might not be peculiar to only the city of Saskatoon

but a reflection of a global economic issue. However, city and provincial leaders can look into some areas of improvement such as tax breaks, checking exploitation by retailers, and improving the standards across the service industry.

5) TOURISM AND ENTERTAINMENT

Approximately 7.89% of the total comments were on tourism and entertainment themes. About 44.61% were positive while 55.39% negative. A lot of posts discuss the beauty of the city and the Springtime Aurora borealis north of Saskatoon. A user said…***C5111****: I love the northern lights around here then I do further up north. Your beautiful picture reminds me how beautiful they are. [2023-04-15]*. Most of the other positive comments are on the radio programming in the city. The morning and afternoon radio programs are seen as soothing and informative. There are also several areas around the city where memorable pictures can be taken.

On the negative sentiments, some residents consider some outdoor events to have generated so much noise. A lot of the comments however focused on the need for improved programs on radio. A post says…***C5212****: if I'm gonna subject myself to an all repeat drive home show, I'll just burn my own CDs [2022-03-30*].

Recommendation: The city can continue to attract tourists by promoting events and places of interest. Radio and television programming can also focus on multiple demographics.

6) HOUSING AND FACILITIES

The housing and facilities theme represents 7.33% of the total data analyzed. About 61.02% of the comments were positive while 38.98% were negative.

Some residents are happy about government-funded low-income homes and reduced property taxes on such homes. New homes are being built by private citizens to address the housing deficit in the city. Realtors in the city have also been praised for their roles in getting people their preferred properties. A user wrote…***C614****: yup. I found Lacey Watson to be pretty honest and she did a good job for us. She made sure we had the information we needed and the mortgage broker needed so we could get the house we wanted at a fabulous price. like all of the comparables except for one were higher and that other comparable was on a slightly smaller lot so in the end we did well. our house appreciated over 70k since then. [2023-03-29.* Other residents agree that properties and taxes are cheaper in some neighborhoods. Some residents are happy about the fact that Saskatoon and Saskatchewan do not have land transfer tax, unlike most other Canadian provinces. Some people also opined that though their income is almost similar to what is earned in Ontario, their leaving cost is cheaper, which makes it affordable for them to leave in Saskatoon. People prefer the building types in Saskatoon with garages than in places such as Vancouver.

However, there are issues with pricing for others. Someone wrote…***C621****: a 400000 dollar house with a 5% down payment at the mortgage rates now is $2066 a month. insurance roughly 150, taxes 185, utilities around 300. comes out 2700 a month give or take a couple hundred. it's ridiculous to own a home now. [2022-05-28]*. Others want the city to solve the existing social problems that exist through class-based home segregation, which breeds resentment. Some residents also feel neighborhoods such as Fairhaven are being used as a social experiment and can hardly complain about theft and other vices. Some people also feel that the low commercial property tax in the city means that homeowners have to pay more to subsidize commercial landowners. Some people are also calling for rezoning of parts of the city. Some posts also suggest that their property tax always goes up by 5% a year.

On home energy consumption issues, residents are happy that the city controls the power supply. However, the heating cost keeps rising by 50-300% per home in the winter.

Recommendation: The cost of housing has been discussed extensively. This theme is multifaceted, and users have touched on different dimensions. The rising home cost was attributed to property taxes, energy bills, income taxes, the greed of landlords, etc. Certainly, discussions on increasing property taxes are not being received well by most citizens. Some realtors also need to be regulated because their actions have been described as an unregulated industry dominated by greed, manipulation, falsehoods, and lies.

7) EDUCATIONAL INFLUENCE AND FAMILY VALUES

The educational influence and family values theme represents 10.69% of the entire dataset. About 35.05% of the comments are positive while 64.95% are negative.

A lot of the posts commended the attendance during the rally in Saskatoon against new sexual education, and pronoun policies in provincial schools. These rallies were organized to support children's rights. Citizens were demanding that the government should properly fund education instead of ridiculous homophobic policies. Parents want to be involved in the decisions of their children at school. A post read…***C711****: …but if the school system is teaching your kids you don't have to tell your parents I feel like it's more detrimental to the parent-child relationship. but instead, should focus on communication with all parties. school and home should be a safe place if you are being abused and hurt at home or school communication is key to coming to a solution. parents should be included in the well-being/education of their children. As parents, we should choose to love everyone! even if you don't necessarily agree, unconditional love should not have any conditions on it (as the word states). [2023-08-28*].

A lot of users regret the consistent underfunding of educational programs. This is something that all political groups have to consider. A user wrote…***C721****: I think this shines another light on just how much the Sask party is bungling education. not only have they systematically underfunded public education, while simultaneously increasing funding to Christian schools, they are now willing to put vulnerable children at risk to pander to the far right wing. there will be a lot of overlap of people who find both situations abhorrent and they need to work collaboratively to see that our children and grandchildren get the educational experience they*

deserve. [2023-08-23]. Others call for adequate funding for the protection of children's mental health as currently, such funding is almost non-existent. Some posts call for equitable distribution of funding to all schools. Others opined that restrictions and limitations placed on sexual education will increase the risk of teenage pregnancy and STD transmission. Others want the schools to focus on teaching the kids taxes, math, science, and reading.

Recommendation: The educational influence issues bother parents wishing to be involved in the decisions of their children at school. Parents want to be informed when their children change their pronouns. Most parents oppose the idea of their children keeping their school life private from their home lives. This issue can be addressed with more open dialogues and community engagements on the part of school/city leaders.

8) DOWNTOWN DEVELOPMENT

The comments on downtown development-related issues account for 4.62% of the entire dataset. About 43.96% of the comments were positive sentiments while 56.04% were negative.

Some people express their happiness about the good transit system in downtown Saskatoon because there is limited parking space. One comment read…***C812****: good transit because downtown has bad parking I'd love a decent little arcade for siblings to check out and lounge in [2022-08-06]*. There were suggestions for the city to consider the development of underground parking facilities. There are dining options, pubs, arcades, and random fun things to do downtown.

While so many people posted that they want an arena to be built downtown, others opposed the idea suggesting that the money be used for something different. A comment reads…***C8212****: we have crumbling infrastructure, we run deficits for snow removal, we have a serious homeless problem, and we continue to lose money on the remainder. but yeah let's build a giant arena downtown. makes sense. [2023-07-18]*. Others complain about the high cost of goods and services downtown.

Recommendation: Residents want to see the addition of facilities downtown, but such initiatives require a huge initial investment cost. The city needs to weigh the impact such developments will have on existing infrastructure.

9) POLICING

Comments concerning policing dominate 4.48% of the entire dataset. The sentiments are 22.14% positive and 77.86% negative. Some residents are happy with crime reporting avenues in the city by the police service. Violent crimes are minimal. A resident wrote…***C1111****: most if not all police services in Canadian cities publicly post crime statistics on their websites including heat maps that display locations and types of reported crimes. you might even be able to compare statistics with where you live if they do the same. generally, many of the more violent crimes such as murder here tend to not be random events but are between people who know each other, either domestic issues or criminal conflicts. [2023-08-27]*. Other residents recall having pleasant conversations with the police in the city.

However, some residents complain of rampant petty crimes in their neighborhoods. Some discussed their bad experiences of being targeted by gangs. Others remarked about being failed by the police and justice system in the city. A domestic violence survivor complaint in a post as…***C1127****: Okay, so I have experience as a domestic violence survivor. the SPS officer did not want to hear about it unless I had some kind of evidence. Prince Albert police were more sympathetic but could not do anything as the crimes were out of their jurisdiction. [2022-08-10]*.

Recommendation: The crime-related issues can be investigated further, and police presence could be increased in hot spots across the city. Violent gangs should also be investigated and disbanded. The crime rate can be reduced through community policing.

10) HEALTHCARE (DRUG ADDICTION, REHABILITATION, AND COVID-19)

The healthcare theme contains 8.75% of the entire dataset. Approximately 31.31% were positive and 68.69% negative.

Most of the residents discuss the need for support for people with addiction(s). There are suggestions that the homeless and people battling drug addictions be offered a warm place and fed while they go through treatment. Others want the city to have a 24-hour minor emergency/medi-clinic. A user wrote…*C1217: See I think an education lesson on how care here works would benefit everyone. When to use an 811, a pharmacy, make an appointment, telehealth or digital health, medi-clinic, minor emergency, or an ER. Teach people how to help themselves navigate our healthcare system. Education would go a long way. [2023-09-12]*.

Users also want people to desist from chronically misusing emergency health services. The healthcare system within the province is also underfunded, understaffed, and overworked according to some comments. Others also listed neighborhoods in need of addiction centers. See…*C1228: Sure thing. Stonebridge, Erindale, and Wilowgrove are all lacking addiction facilities. Fair is fair. [2023-08-28]*.

Some users praised the city's push to develop advanced labs for pathogen testing. The city has a level-four containment facility that houses the world's most contagious and severe pathogens. Someone wrote…***C911****: Very cool. Before this one, there was only one BSL-4 lab in Canada, located in Winnipeg. Now we will have more people with access to pathogens that are pandemic and endemic-level to study them and discover ways to keep us safe and healthy! [2023-08-26]*. There is, however, general agreement that the city's medical system needs drastic improvement and nurses need incentives. So many users also discussed the need for vaccinations and mentioned how pharmacies are equipped to administer them.

Recommendation: The city is prepared for the COVID-19 pandemic and its aftermath from the comments. Residents

want to see more concerning public health education in addition to the availability of vaccines. Further, we recommend that the city prioritize addiction and rehabilitation issues. This has the propensity to reduce crime, decrease the burden on social infrastructure, and increase productivity.

11) GARBAGE AND RECYCLING

This theme represents 3.44% of the total dataset. About 33.20% of the comments are positive and 66.80% negative.

Some residents are happy with current city regulations that prevent people from burning fuel and polluting their neighbors. Also, the small bins accompanied by recycling and compost bins will divert some waste away from landfills says a user. Some have also recently realized that throwing cigarette butts around is illegal. A user wrote…***C1111****: it's not Saskatoon police service, it's bylaw but the only way to make people stop doing it is to slap them with fines. I appreciate you being a responsible smoker and keeping your butts. [2023-06-22].* Others praised the city for its efficient management of bio-waste.

Some residents want to see ashtrays to dump their cigarette butts. Others also discuss the influence of waste on wildfires. Other people noted how parts of the west side of the city are littered with trash.

Recommendation: The good initiatives on trash management by the city are commended by residents. However, public education is needed to address the problem of trash and dirt in certain parts of the city. There should be more trash cans stationed around public places.

12) ANIMAL CONTROL

This theme represents 3.16% of the overall dataset. Almost 39.93% of sentiments in the comment were positive and 60.07% negative.

Residents wrote positively about pet-friendly parks across the city. A user remarked…***C1517***: …*Corman Park has been amazing with my pets from spay/neuters right up until the end. They have an amazing staff, and knowledgeable staff, and the payment plan has been helpful. They're so good! [2022-03-14].* Some landlords are also willing to allow tenants with pets.

However, most of the concerns were about the cost of ownership. Someone wrote…***C1529****: but why is the cost so outrageous? it is just another thing that low-income people are being price-gouged out of. the benefits of having pets are incredible and proven to be therapeutic. I think it's very sad that pet ownership is not an attainable goal for most people now. [2023-07-04].* Some people also suggested that owners should train their pets to avoid causing damage to properties. Some animals outdoors should also be kept on a leash. Pet owners must do more to train their animals.

Recommendation: Residents who own animals should endeavor to take their pets to the vet when necessary and applicable. Also, owners must be responsible for the welfare of their pets. The city can do more in terms of managing how pets defecate in public places and who's responsible for cleanup.

VIII. CONCLUSION

Most cities are pursuing the smart city agenda to improve operational efficiency and offer better quality governance services to their citizens. In this regard, city leaders of Saskatoon, Canada want to know what the citizenry needs. Social media platforms are one of the sources of social discourse and public engagement.

Thus, we turned our attention to the Reddit platform to collect data from Saskatoon subreddit posts between the period of January 1, 2019, to September 20, 2023, to discover topics that emphasize the concerns of citizens. Three topic models were used for the analyses which include: Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and BERTopic with a K-means clustering algorithm. Our findings showed that BERTopic can discover coherent and diverse topics compared to LDA and NMF.

We discovered 12 underlying themes by merging related topics. Also, we leveraged SiEBERT (a pre-trained transformer model), which returned an accuracy of 89% compared to VADER and Random Forest which achieved an accuracy of 57% and 86% respectively in the prediction of sentiments. Based on the thematic analysis, we provided recommendations for each concern identified.

In summary, the paper made the following contributions to social science, social computing, and governance.

- Employed AI/ML, and social media data from the Reddit community to highlight the societal issues affecting the citizens of Saskatoon, Canada. Hence, pivot the vision of the city towards smart city design.
- Explored multiple AI/ML topic models such as LDA, NMF, and BERTopic to determine the best-performing model as applied to the Reddit data.
- The work further evaluated the performance of the BERTopic model (a) when no preprocessing is done, and (b) when data is pre-processed.

The research discovered factors for smart city engagement as applied to Saskatoon such as:

Online Community Engagement – The residents want to engage in online communities that will be used to recommend products and services. The government services can be made accessible via their websites.

Transportation and Weather Conditions – Residents want some of the road intersections to be redesigned. They also want to see better transit schedules.

Cost of Living – Some residents bemoan the rising cost of goods and services. They also noted shrink-flation.

Educational Influence and Family Values – On this theme, parents want to be involved in the decision-making processes of their children while in school. They are calling on the city and government to review existing educational laws and policies.

Healthcare – This theme focused more on calls for the city to address issues of drug addiction, rehabilitation, and other public health challenges such as COVID-19.

Protest and Political Elections – The city accepted some political and policy protests and residents were happy about that. However, they want to see more engagement and clear political manifestos that are more inclusive.

Tourism and Entertainment – The city should work more to attract tourists by promoting events and places of interest.

Housing and Facilities – Residents complain about the unbearable costs of housing and facilities. Other attempts can be made to reduce homelessness.

Downtown Development – Residents want to have more parking spaces downtown. They also suggested other infrastructural projects such as an arena.

Policing – Community policing strategies can be adopted to reduce the crime rate further in the relatively violent parts of the city. This is also to acknowledge that most areas in the city have a low crime rate and are safe.

Garbage and Recycling – Some areas in the city need more education and garbage cans stationed to keep the neighborhoods clean.

Animal Control – The residents are encouraged to keep their pets safe and controlled in public places such as the park.

In the future, we shall focus on the implications of emerging technologies on smart cities such as smart health, smart energy, smart agriculture, smart transit, smart buildings, etc. We shall use our findings to analyze climate change issues across North America.

REFERENCES

[1] Y. A. Argyris, K. Monu, P.-N. Tan, C. Aarts, F. Jiang, and K. A. Wiseley, "Using machine learning to compare provaccine and antivaccine discourse among the public on social media: Algorithm development study," *JMIR Public Health Surveill.*, vol. 7, no. 6, Jun. 2021, Art. no. e23105.

[2] V. A. Pitogo and C. D. L. Ramos, "Social media enabled e-participation: A lexicon-based sentiment analysis using unsupervised machine learning," in *Proc. 13th Int. Conf. Theory Pract. Electron. Governance*, Sep. 2020, pp. 518–528.

[3] M. R. Haupt, A. Jinich-Diamant, J. Li, M. Nali, and T. K. Mackey, "Characterizing Twitter user topics and communication network dynamics of the 'liberate' movement during COVID-19 using unsupervised machine learning and social network analysis," *Online Social Netw. Media*, vol. 21, Jan. 2021, Art. no. 100114.

[4] S. Chen, L. Zhou, Y. Song, Q. Xu, P. Wang, K. Wang, Y. Ge, and D. Janies, "A novel machine learning framework for comparison of viral COVID-19–related Sina Weibo and Twitter posts: Workflow development and content analysis," *J. Med. Internet Res.*, vol. 23, no. 1, Jan. 2021, Art. no. e24889.

[5] A. Adikari and D. Alahakoon, "Understanding citizens' emotional pulse in a smart city using artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2743–2751, Apr. 2021.

[6] M. Heidari, J. H. J. Jones, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS)*, Apr. 2021, pp. 1–5.

[7] K. Saraswathi, V. Mohanraj, Y. Suresh, and J. Senthilkumar, "Deep learning enabled social media recommendation based on user comments," *Comput. Syst. Sci. Eng.*, vol. 44, no. 2, pp. 1691–1702, 2023.

[8] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing," *Internet Things*, vol. 11, Sep. 2020, Art. no. 100222.

[9] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107440.

[10] S. Blasi, E. Gobbo, and S. R. Sedita, "Smart cities and citizen engagement: Evidence from Twitter data analysis on Italian municipalities," *J. Urban Manage.*, vol. 11, no. 2, pp. 153–165, Jun. 2022.

[11] P. K. Jain, W. Quamer, V. Saravanan, and R. Pamula, "Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis," *J. Ambient Intell. Hum. Comput.*, vol. 14, no. 8, pp. 10417–10429, Aug. 2023.

[12] H. Adamu, S. L. Lutfi, N. H. A. H. Malim, R. Hassan, A. Di Vaio, and A. S. A. Mohamed, "Framing Twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning," *Sustainability*, vol. 13, no. 6, p. 3497, Mar. 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/6/3497

[13] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence," *J. Infection Public Health*, vol. 14, no. 10, pp. 1505–1512, Oct. 2021.

[14] Y. Liu, C. Whitfield, T. Zhang, A. Hauser, T. Reynolds, and M. Anwar, "Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning," *Health Inf. Sci. Syst.*, vol. 9, no. 1, p. 25, Dec. 2021.

[15] M. Singh, A. Singh, S. Bharti, P. Singh, and M. Saini, "Using social media analytics and machine learning approaches to analyze the behavioral response of agriculture stakeholders during the COVID-19 pandemic," *Sustainability*, vol. 14, no. 23, p. 16174, Dec. 2022.

[16] M. Chau, T. M. H. Li, P. W. C. Wong, J. J. Xu, P. S. F. Yip, and H. Chen, "Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification," *MIS Quart.*, vol. 44, no. 2, pp. 933–955, Jun. 2020.

[17] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107242.

[18] S. Praveen, R. Ittamalla, and G. Deepak, "Analyzing the attitude of Indian citizens towards COVID-19 vaccine—A text analytics study," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 15, no. 2, pp. 595–599, Mar. 2021.

[19] T. Yigitcanlar, N. Kankanamge, M. Regona, A. R. Maldonado, B. Rowan, A. Ryu, K. C. Desouza, J. M. Corchado, R. Mehmood, and R. Y. M. Li, "Artificial intelligence technologies and related urban planning and development concepts: How are they perceived and utilized in Australia?" *J. Open Innov., Technol., Market, Complex.*, vol. 6, no. 4, p. 187, Dec. 2020.

[20] A. Kovacs-Györi, A. Ristea, C. Havas, M. Mehaffy, H. H. Hochmair, B. Resch, L. Juhasz, A. Lehner, L. Ramasubramanian, and T. Blaschke, "Opportunities and challenges of geospatial analysis for promoting urban livability in the era of big data and machine learning," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 12, p. 752, Dec. 2020.

[21] M. Tran, C. Draeger, X. Wang, and A. Nikbakht, "Monitoring the well-being of vulnerable transit riders using machine learning based sentiment analysis and social media: Lessons from COVID-19," *Environ. Planning B, Urban Anal. City Sci.*, vol. 50, no. 1, pp. 60–75, Jan. 2023.

[22] S. Milusheva, R. Marty, G. Bedoya, S. Williams, E. Resor, and A. Legovini, "Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, Art. no. e0244317.

[23] A. Hodorog, I. Petri, and Y. Rezgui, "Machine learning and natural language processing of social media data for event detection in smart cities," *Sustain. Cities Soc.*, vol. 85, Oct. 2022, Art. no. 104026, doi: 10.1016/j.scs.2022.104026.

[24] A. Elabora, M. Alkhatib, S. S. Mathew, and M. El Barachi, "Evaluating citizens' sentiments in smart cities: A deep learning approach," in *Proc. 5th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Sep. 2020, pp. 1–5.

[25] S.-U. Hassan, M. Shabbir, S. Iqbal, A. Said, F. Kamiran, R. Nawaz, and U. Saif, "Leveraging deep learning and SNA approaches for smart city policing in the developing world," *Int. J. Inf. Manage.*, vol. 56, Feb. 2021, Art. no. 102045.

[26] L.-C. Chen, C.-M. Lee, and M.-Y. Chen, "Exploration of social media for sentiment analysis using deep learning," *Soft Comput.*, vol. 24, no. 11, pp. 8187–8197, Jun. 2020.

[27] H. Fan, W. Du, A. Dahou, A. A. Ewees, D. Yousri, M. A. Elaziz, A. H. Elsheikh, L. Abualigah, and M. A. A. Al-Qaness, "Social media toxicity classification using deep learning: Real-world application U.K. Brexit," *Electronics*, vol. 10, no. 11, p. 1332, Jun. 2021.

[28] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, "Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113725.

[29] B. Zarouali, T. Dobber, G. De Pauw, and C. de Vreese, "Using a personality-profiling algorithm to investigate political microtargeting: Assessing the persuasion effects of personality-tailored ads on social media," *Commun. Res.*, vol. 49, no. 8, pp. 1066–1091, Dec. 2022.

[30] A. M. Guess et al., "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, vol. 381, no. 6656, pp. 398–404, Jul. 2023.

[31] A. Nistor and E. Zadobrischi, "The influence of fake news on social media: Analysis and verification of web content during the COVID-19 pandemic by advanced machine learning methods and natural language processing," *Sustainability*, vol. 14, no. 17, p. 10466, Aug. 2022.

[32] S. Bojjireddy, S. A. Chun, and J. Geller, "Machine learning approach to detect fake news, misinformation in COVID-19 pandemic," in *Proc. 22nd Annu. Int. Conf. Digit. Government Res.*, Jun. 2021, pp. 575–578.

[33] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets," *Inf. Syst. Frontiers*, vol. 23, no. 6, pp. 1417–1429, Dec. 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10796-021-10135-7

[34] U. Reisach, "The responsibility of social media in times of societal and political manipulation," *Eur. J. Oper. Res.*, vol. 291, no. 3, pp. 906–917, Jun. 2021.

[35] M. Alipour and D. K. Harris, "A big data analytics strategy for scalable urban infrastructure condition assessment using semi-supervised multi-transform self-training," *J. Civil Struct. Health Monit.*, vol. 10, no. 2, pp. 313–332, Apr. 2020.

[36] D. Alahakoon, R. Nawaratne, Y. Xu, D. De Silva, U. Sivarajah, and B. Gupta, "Self-building artificial intelligence and machine learning to empower big data analytics in smart cities," *Inf. Syst. Frontiers*, vol. 25, no. 1, pp. 221–240, Feb. 2023.

[37] C. Bono, M. O. Mülâyim, C. Cappiello, M. J. Carman, J. Cerquides, J. L. Fernandez-Marquez, M. R. Mondardini, E. Ramalli, and B. Pernici, "A citizen science approach for analyzing social media with crowdsourcing," *IEEE Access*, vol. 11, pp. 15329–15347, 2023.

[38] N. Kankanamge, T. Yigitcanlar, A. Goonetilleke, and M. Kamruzzaman, "Determining disaster severity through social media analysis: Testing the methodology with South East Queensland flood tweets," *Int. J. Disaster Risk Reduction*, vol. 42, Jan. 2020, Art. no. 101360.

[39] F. B. Biggers, S. D. Mohanty, and P. Manda, "A deep semantic matching approach for identifying relevant messages for social media analysis," *Sci. Rep.*, vol. 13, no. 1, Jul. 2023, Art. no. 12005.

[40] E. C. McClure, M. Sievers, C. J. Brown, C. A. Buelow, E. M. Ditria, M. A. Hayes, R. M. Pearson, V. J. D. Tulloch, R. K. F. Unsworth, and R. M. Connolly, "Artificial intelligence meets citizen science to supercharge ecological monitoring," *Patterns*, vol. 1, no. 7, Oct. 2020, Art. no. 100109.

[41] R. K. Lomotey, S. Kumi, M. Hilton, R. Orji, and R. Deters, "Using machine learning to establish the concerns of persons with HIV/AIDS during the COVID-19 pandemic from their tweets," *IEEE Access*, vol. 11, pp. 37570–37601, 2023, doi: 10.1109/ACCESS.2023.3267050.

[42] A. Schofield and M. Magnusson, "Understanding text pre-processing for latent Dirichlet allocation," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 432–436, doi: 10.1145/1378773.1378800.

[43] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, nos. 4–5, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[44] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, Jun. 1994, doi: 10.1002/env.3170050203.

[45] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

[46] D. Kuang, J. Choo, and H. Park, "Nonnegative matrix factorization for interactive topic modeling and document clustering," in *Partitional Clustering Algorithms*. Switzerland: Springer, Jan. 2015, pp. 215–243, doi: 10.1007/978-3-319-09259-1_7.

[47] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, *arXiv:2203.05794*. Accessed: Sep. 11, 2023.

[48] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, May 2014, vol. 8, no. 1, pp. 216–225, doi: 10.1609/ICWSM.V8I1.14550.

[49] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," *Int. J. Res. Marketing*, vol. 40, no. 1, pp. 75–87, Mar. 2023, doi: 10.1016/j.ijresmar.2022.05.005.

[50] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," Jul. 2019, *arXiv:1907.11692*. Accessed: Sep. 12, 2023.

[51] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Edinburgh, U.K., Jul. 2011, pp. 262–272.

[52] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 399–408, doi: 10.1145/2684822.2685324.

SANDRA KUMI (Graduate Student Member, IEEE) received the B.Sc. degree in computer science from the Kwame Nkrumah University of Science and Technology, Ghana, and the M.Sc. degree in computer science from Dongseo University, South Korea. She is currently pursuing the Ph.D. degree in computer science with the University of Saskatchewan, Canada. Her research interests include data trust, blockchain, cybersecurity, and machine learning.

CHARLES SNOW is currently pursuing the Bachelor of Science degree with The Pennsylvania State University, University Park, PA, USA. He is studying information science and technology. His research interests include software development, machine learning, and cyber security.

RICHARD K. LOMOTEY (Member, IEEE) is currently an IT Professor with The Pennsylvania State University, USA. He is also a Program Coordinator with the Information Sciences and Technology (IST) Program and the Cybersecurity Analytics and Operations (CYAOP) Program with Penn State Beaver Campus. He is researching the intersection of smart technologies, mobile computing, data science, machine learning, and cybersecurity, with a focus on how these technologies are transforming enterprises and policies.

RALPH DETERS (Member, IEEE) received the Ph.D. degree from Federal Armed Forces University, Munich, Germany, in 1998. He joined the University of Saskatchewan, as a Research Associate, in 1998. He is currently a Full Professor with the Department of Computer Science, University of Saskatchewan. His research interests include distributed ledger technology, the IoT, and cloud/edge computing.