Received May 16, 2021, accepted June 1, 2021, date of publication June 11, 2021, date of current version July 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3088410

Monitoring Cyber SentiHate Social Behavior During COVID-19 Pandemic in North America

FATIMAH ALZAMZAMI AND ABDULMOTALEB EL SADDIK, (Fellow, IEEE)
Multimedia Communication Research Laboratory, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
Corresponding author: Fatimah Alzamzami (falza094@uottawa.ca)

ABSTRACT With communications being shifted to online social networks (OSNs) as a result of travel and social restrictions during COVID-19 pandemic, the need has arisen for discovering emerging trends and concerns formed during the pandemic as well as understanding the corresponding online social behavior that reflects its offline settings. The online connectivity of devices through social media is one example of Internet of Things (IoT) in which a two-way communication between societies and officials, could be created. Therefore, it is possible to monitor people's behavior through OSNs, especially during pandemics, to prevent potential social and psychological instabilities that might lead to undesired consequences. This is particularly crucial for governmental and non-governmental organizations to ensure the stability and well-being in societies. In response, we propose a pandemic-friendly real-time framework for monitoring cyber social behavior by utilizing unsupervised and supervised learning approaches. Two BERT-based supervised classifiers are trained and constructed to analyze two types of online social behaviors, hate and sentiment. Unsupervised framework is proposed for OSNs data exploration and coherent interpretation that is used as a complementary tool to facilitate the analysis of online social behaviors during pandemics. Extensive experimentation and evaluation have been conducted to validate the effectiveness of the proposed work. Our results have shown superior performance of our BERT-based models in two classification tasks: 1) binary classification for hate behavior detection and 2) multi-class classification for sentiment behavior detection. In addition to our experimentation results, our large-scale analysis of COVID-19 pandemic has illustrated the capability of our unsupervised framework for concerns and trends discoveries using OSNs data, along with reliability in automatically and dynamically providing phrase-based interpenetration of the inferred trends and concerns. This paper provides a twelve-month comparison analysis of data discoveries and online social behavior between Canada and USA during COVID-19 pandemic.

INDEX TERMS Hate speech, sentiment, topic modeling, BERT, topic interpretation, phrase extraction, RAKE, online social media, online social behavior, Covid-19, tweets, twitter.

I. INTRODUCTION

With the COVID-19 global pandemic, the whole world has gone into an abrupt shift where every aspect of our lives has been impacted. This unprecedented crisis and the restrictions made to curb it have affected the way we live, behave, work, learn and even communicate. As a result, a new world-wide communication culture has been created. Virtual communication has been the new normal in the wake of Covid-19 pandemic in which millions of devices are connected between patients-doctors or citizens-officials. While Internet of Things (IoT) can capture a tremendous amount of data from these devices, it is not able to make sense of the data on its own. Integrating artificial intelligence (AI) with IoT however helps in making sense of data which in turn stimulating smart decision making. There have been rapid efforts in AI-enabled IoT researches during COVID-19 pandemic to combat the threat that pandemic imposes to societies [1]. AI-enabled IoT solutions have shown a major potential in controlling the spread of the virus and the monitoring of patients and public health [1], [2].

The spread of the corona virus has been overwhelming. The state of uncertainty and worry has caused people to feel stressed and anxious as they are unable to have a sense of control over their lives. To make things worse, propaganda and the spread of intensified information have particularly contributed in fuelling the crisis, which has posed massive challenges for the world. Compulsory quarantine, social

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

distancing, and lockdown measures, along with a severe economy collapse have caused psychological instabilities and health implications among people [3]. This has led all types of communications to pour in online social media, and a large pile of remote communication and conversations have been created, the most noticeable of which are those between citizens and officials across social media platforms namely Twitter. Around 4.4M original COVID-19 related tweets were published (i.e. according to our data collection) only during the period between the months of May and August from North America [4]. Thus, social media comes under the microscope to discover emerging trends and concerns as well as understand the corresponding social behaviors that reflect the actual settings nation-wide and world-wide. This is particularly crucial for relevant organizations like governments and civil society towards taking proper measures and providing adequate public responses.

It is vital to monitor public health during pandemics In view of the foregoing, it is vital to monitor public online social behaviors for the purpose of gaining a better understanding of their implications offline. Recognizing negative sentiments during the pandemic, for example, is a valuable tool for analyzing underlying issues and concerns to which a proper plan action could be provided. Being economically underprivileged and quarantined due to fear of infection have triggered high degrees of violence and hate. According to Facebook,[1] there has been around 134% increase in the number of hate posts from the first quarter to the second quarter of this year (i.e. 2020). The United Nation (UN)[2] has also released hatespeech guidelines related to COVID-19 pandemic. It has been reported that with the rise of COVID-19 cases and restrictions, an increase of a new wave of hate speech has been observed.

The extensive flow of information streams could efficiently be handled using AI powers to continuously keep track of public current states since it is nearly impossible to manually monitor huge loads of online data flow. In this context, this paper tries to answer five questions: (1) what high-level insights, trends, and concerns can be inferred from COVID-19 data?, (2) can the inferred trends and concerns be automatically and coherently interpreted?, (3) are there any trends or concerns that lead to emotional or hate spikes during COVID-19 pandemic?, (4) what and when does a behavior overtake the other during the pandemic?, (5) how does the spread of negative emotion impact the spread of hate behavior?.

Sentiment models trained using domain-specific data usually come with the concern that they do not generalize to other domains [5] since they latch to the information of the domain they have learned from [6]. However, the trend among recent sentiment analysis works related to COVID-19 have used either datasets collected with pre-defined topics and keywords or noisily-annotated datasets [7], [8]. This is understandable due to the limited resources of manually-annotated COVID-19 datasets which are expensive to construct in terms of time, efforts, cost, and human labor. Not to mention that the data size is preferable to be large especially for neural networks to be able to learn good feature representations from the data. An alternative effective solution is to opt for general (i.e. domain-independent) sentiment datasets that follow manual and domain-independent annotation protocols. Recent findings in [5] revealed that specific-domain sentiment models do not generalize well on domain-specific datasets especially on the negative class, which indicates that negative expressions are more specific to individual domains than positive expressions are. Unlike domain specific models, general sentiment models have shown to well adapt to specific-domain sentiment datasets for both positive and negative classes. In this work, we solve the domain-dependence issue by using our Domain-Free-Sentiment-Multimedia dataset (DFSMD) [9] that follow high quality data-collection and annotation protocols to meet the purpose of this study.

The UN[2] warns that the communications during COVID-19 pandemic could be exploited to instigate discrimination, stereotyping, stigmatization, racisms and xenophobia, all of which fall under the umbrella of hate behavior according to several universal definitions of hate speech[3], [10], [11]. Updating its policy guidelines, Twitter has warned against the use of hate language. It disallows promoting "violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease". Accordingly, it is understandable that hate behavior embodies violence, abuse, or harassment language which indicates that hate language consists of multifaceted contexts. Recent trends in hatespeech-related researches have focused on specific targets such as racism or aggression [6], [12]. Accordingly, the datasets have also been featured according to the targeted focuses, thus introducing a challenge of identifying universal patterns of hate language across social media. This approach makes hate detection a domain-dependent task. That being said, the requirements to build different models and datasets to capture different hate language phenomena should increase notably with the presence of limited resources of domain experts and high expenses of datasets manual annotation. The challenge of modeling domain-dependent hate language sheds light on the importance of learning general patterns of hate language. This general knowledge of models helps not only in capturing a wide spectrum of hate behavior across social media, but also in controlling and detecting the spread of hate contents regardless of their types. In this paper, we exploit the power of transfer learning combined with various phenomena of violence and hate languages as an attempt to build a generalized hate model capable of detecting general patterns of hate language phenomena on social media.

Given the critical situation of the pandemic and the continuous and heavy information flow during the pandemic,

[1] https://transparency.facebook.com/community-standards-enforcement
[2] https://www.un.org/

there is an urgent need to discover global and local concerns and issues in real time, yet it is nearly impossible to achieve this manually. The unsupervised learning nature of topic modeling methods makes it possible to achieve this goal fast and without prior human knowledge involved. However, these methods are sensitive to data noises, which is a normal phenomenon in OSNs data. It is well-known that social media data suffers from noises such as misspellings and the intense use of abbreviations due to the limited writing-space capacities. This challenge requires a careful handling of the OSNs data in order for the topic modeling methods to perform well. Given the issues of data noises and short lengths, proper techniques are needed to extract every informative piece of information from messages while synchronously removing the unnecessary noises residing within these messages. Although topic modeling methods are proven effective in capturing hidden insights from the social data, it is not capable of providing a coherent interpretation of the inferred insights [13]. However, studies [14], [15] still depend on using the top n words resulted from these topic models to interpret the topics (i.e. discovered insights or clusters) on social media. Another cheap alternative that has been considered is to interpret the topics manually. Manual topic interpretations requires human efforts and can be easily biased towards subjective opinions [16]. Both approaches are not applicable for pandemic-friendly systems which require accurate and instant interpretations in order to make proper decisions. According to the literature, people prefer phrases over single words to understand topics [13]. They claim that combining single words creates difficulties to comprehend the main meaning of topics while sentences are too specific and might miss other aspects of topics. Given the diversity of conversations on OSNs, finding the optimal length of phrases that best describe a topic is challenging. Current methods rely on a fixed sliding window for phrases which might limit the comprehension of topics. Some topics might use longer or shorter phrase-expressions than the other and this cannot be controlled on open platforms like OSNs that encourage unstructured data format.

This paper exploits the Natural Language Processing (NLP) techniques as pre-requirements to topic modeling in order to maximize the learning performance of topic models on OSNs. In addition, unsupervised learning approach is used to find phrases of dynamic sizes that provide coherent interpretations of the inferred topics automatically. We use the data explorations and interpretations as complementary tools to facilitate the understanding of online social behavior during and post pandemic. This work is an attempt to assist in catering to public safety and psycho-social needs towards providing measures for developing healthy coping strategies to reduce the psycho-social instabilities during and post pandemic. It could also create opportunities for tracing individuals or groups responsible for violent incitement as it has been proven that it is possible to infer this type of information through OSNs [17], [18]. We summarize the contributions of this work as follows:

- Design a framework for real-time monitoring of online social behaviors in online social networks (OSNs).
- Design a framework for real-time data exploration and interpretation using unsupervised learning approach.
- Develop a BERT-based sentiment classifier for general sentiment analysis using Domain-Free-Sentient-Multimedia dataset (DFSMD).
- Develop a BERT-based general hate classifier using wide range of violence and hate phenomena.
- Conduct a benchmark evaluation of RAKE algorithm on online social networks (OSNs) data and large scale analysis on COVID-19 data.
- Conduct comprehensive experiments to evaluate the performance of deep algorithms for sequence classification on two classifications problems: multi-class sentiment analysis and binary-class hate speech analysis.
- Conduct a large scale analysis of online social behavior during COVID-19 pandemic in North America for 12 months.

The rest of the paper is organized as follows. Section II presents in details the related work and Section III gives an overview background of the methods used in this work. Our proposed framework and methodology is presented in Section IV. Section V explains the datasets used in the modeling and analysis of this paper. The preprocessing steps are listed and described in Section VI. Section VII explains the experiment design and evaluation protocol followed in this work whereas the results and analysis are discussed in Section VIII. Finally, in Section IX we conclude our proposed work and findings and discuss future directions.

II. RELATED WORK

Research communities have witnessed ongoing efforts aimed at tackling the crisis of COVID-19 pandemic; smart health domain [19] utilizes artificial intelligence (AI) and Internet of Things (IoT) to lower the spread of corona virus and enhance health and well-being during the pandemic [1], [2], [4]. Health and patient monitoring is a popular application of AI-enabled IoT technology [20]. Hossain *et al.* [1] developed an AI-enabled IoT surveillance system that monitors mask wearing, social distancing, and body temperature in order to control the spread of the virus in communities. Social robots [2] were used to communicate with patients during quarantine time in order to reduce mental strain and keep track of their mental health. In this work, we utilize users' communications in social media to monitor online social behavior as an attempt to enhance public health and well-being during pandemic times.

Recent COVID-19 related works have focused on analyzing a single aspect of online social behavior; that is either sentiment [7], [8], [21] or hate [15], [22]. Other studies have focused on analyzing a single geographical location [7] or a single event [21] during the pandemic. Nevertheless few studies have utilized topic modeling methods for a wider-view analysis of the pandemic [14], [15], [23], [24]. However, to the best of our knowledge, none has addressed the automatic topic interpretation except for using top *n* words (i.e. resulting from topic models) or relying on human intervention to interpret the topics. In this work, we propose to monitor two online social behaviors: sentiment behavior and hate behavior during the pandemic (December 2019-Novemebr 2020). We utilize topic modeling methods to discover main issues, trends and interests emerged during the pandemic. We also provide automatic interpretation of the data discoveries through a set of phrases used by OSNs crowds during social conversations.

Bag-of-words (BOW) and Term-Frequency-Inverse-Document-Frequency (TFIDF) features at the *n*-gram level have been widely used with LDA and NMF algorithms for topic modeling in social media [25], [26] particularly in COVID-19 related social analysis [15], [23], [24]. Many of these studies have only accommodated removing canonical stopwords (e.g. ''the'', ''and'') during the pre-possessing step to construct features. However, removing canonical stopwords does not entirely solve the problem of the existence of common uninformative words, which will definitely affect the quality of the topic models. For example, LDA models trained without removing common words will produce topics with high probabilities of uninformative words. To overcome this issue, literature suggests removing domain-specific [27] and corpus-specific stopwords [28]. Such methods have been proven effective in enhancing coherence across topics. Authors in the work [29] took it further and showed that lemmatizing the corpus and limiting the vocabulary of news collections to only nouns, has improved the semantic coherence of topic models. However, reporting news is one part of social media data. OSNs platforms are open; hence the data flow spectrum is broad ranging from reviewing a product, expressing frustration, to reporting news. Ignoring other part-of-speech tags will result in throwing important information that can be found in nouns and verbs for example. Incorporating different part-of-speech tags for topic modeling [30] has shown to produce reasonable topics on a small Twitter dataset. In this paper, we apply the same approach but on a large scale datasets.

While topic models are proven effective in extracting latent patterns (i.e. themes or topics) out of social media data [31], they fall short in providing human-friendly interpretations for these topics [32]. Manual interpretation of topics is subject to human bias [13]. Moreover, given the diversity of OSNs contents and huge data volumes makes the availability of domain experts to annotate data for various problems, a difficult task. Early researches on topic labelling focused on exploiting external knowledge resources in order to automatically label topics of topic models. However, this approach is not applicable to OSNs data streams since the emerging social contents and events discussed in OSNs might not exist in these external resources in a timely manner [33]. Later, the focus redirected towards labelling topics with the most representative single words based on the output of topic models [16]. Single words provide generic meaning, which makes it difficult for users to create the main idea when single words of topic models are combined. In addition, single words may often be homonyms (i.e. they sound the same and have the same spelling but do not have related meanings) or polysemous (i.e. the word is used to express different meanings depending on the context). In this context, Qiaozhu *et al.* [13] proposed the use of phrase labels to automatically label LDA-style topics. The results of their questionnaire showed that people prefer phrases over words for topic comprehension. However, their approach depends on NLP techniques (i.e. chuncking, POS tagging, and *n*-gram) which is resource and time consuming. Additionally, the approach focuses on topics derived from static well-formatted documents (i.e. news articles and scientific article) which is the case in the work [34] as well. Recently, Amparo *et al.* [33] have tackled this issue and presented the topic labelling of Tweets as a summarization problem. The results demonstrated that the topic labels generated by their method showed that the use of summaries, as topic labels outperformed the use of top *n* words resulting from LDA model. However, the output summaries consist of single words computed using methods based on TextRank and TFIDF where the latter was shown to yield the best labels. Given the dynamic size of topics ranging from being small to large, TFIDF would fall short on small data sizes. Recently, a phrase-based topic labelling approach [35] has been developed based on OSNs activities parameters (i.e. views and likes). However, phrases of length two were only considered. The meaning of a sentence varies with the order and length of its constituting words (e.g. noun-verb-adjective phrase). In this work, we propose to use RAKE algorithm for automatic topic interpretation as it solves the mentioned issues of current topic labelling methods. To the best of our knowledge, this work is the first to address these issues and to utilize RAKE algorithm for automatic interpretation of LDA-style topics using OSNs data.

Recent studies on using BERT models for sentiment analysis have focused on domain-specific analysis [36], [37], [37], [38] and the polarity aspect of the sentiment while ignoring the objectivity part of texts [36], [38], [39]. Moreover, training and evaluation have been conducted on small datasets [36], [39] and many works have not handled the OSNs cultural language such as iconic emotions (i.e. emojis and emoticons) [38], [40], [41]. It has been repeatedly reported that training deep neural networks using large datasets yields better results than the training using small datsets [42]. Also, iconic emotions contain sentimental clues that would greatly contribute in sentiment learning [5], [43]. In addition, automatically or noisy annotated data has been used to train BERT-based sentiment model [8], [41] which in turn compromises the knowledge quality of the learning process. Even though cross-domain sentiment learning has been addressed, many studies focus on adapting sub-domain to one another while considering the main domain to be the same [36], [37], [44]. This approach would not generalize well on various domains since those models will latch on to domain-specific information [6]. This paper addresses all the mentioned issues. First, it considers the subjective and objective aspects of sentiment (i.e. positive, negative, neutral). Second, it provides a training dataset (DFSMD) of a decent size and of high quality psychologist-based manual annotation. Third, it provides a domain-free dataset (DFSMD) that was constructed free of restrictions to any domains or keywords. Fourth, it proposes a domain-free BERT-based sentiment model to bridge specific domains mismatches [5]. It also can be used to enhance learning of sentiment in the domain-specific problems by transferring the general sentiment knowledge instead of starting from scratch. To the best of our knowledge, this paper presents one of the first studies to build a general sentiment model based on BERT language model.

BERT models have shown state-of-the-art performance in detecting hate speech on social media [45], [46]. Marzieh *et al.* [47] trained BERT-based models for different hate speech categories: racism/sexism and hate/offensive. The overall results showed that BERT-based models yielded excellent performance. Offensive language was studied in [48] using BERT pre-trained model as a base for modeling offensive classifier using OffensEval-2019 dataset. BERT-based model was shown to outperform classical machine learning methods in identifying offensive language in tweets. However, iconic emotions (i.e. emojis and punctuation-based emoticons) were ignored. OSNs specific feature such as exclamations marks, words with repetitive characters (e.g looool) were ignored even though they contain strong sentimental insights [5]. Authors in [49] targeted the problem of aggression and misogynistic identification for three languages on social media. Their approach included using BERT pre-trained model yet no fine-tuning was conducted. They reported that BERT-based model had shown a better performance on binary classification than that of multi-class classification. Hind *et al.* [50] proposed using domain-specific word embedding with BERT model to detect white supremacist hate speech if it existed on social media. The evaluation was conducted on balanced dataset collected from Twitter and Stormfront forum. However, the size of the proposed training set is quiet small (i.e. 4588 messages) to be able to fine tune BERT architectures. This would raise concerns regarding model generalization and network's overfitting. The previous studies have shared one approach which is focusing on studying one aspect of the multifaceted hate language behavior (e.g. racism or aggression). Following the same approach, the hate datasets were designed and crafted to focus on a specific aspect of hate language. However, this focus makes it limited and difficult to identify general hate language across various events on social media. Authors in [12] investigated the abusive language generalization across datasets of different abusive focuses. Their findings and observations concluded that models trained using datasets with a broader coverage of phenomena are more robust in capturing a wider range of abusive language contents. LSTM-based models were shown to outperform models based on linear support vector classifier. Similar observation was found in the work [6] where authors claimed that supervised learning using domain-specific datasets performs poorly on cross-domain datasets due to the reason that they are attached to the domain-specific information. Their results demonstrated the effectiveness of using domain-independent abusive lexicon to detect abusive language in cross-domain social media datasets. Waseem *et al.* [51] confirmed the possibility of obtaining high-performance models to detect hate and abusive language when built using composite datasets. However, no considerations have been given to OSNs-specific features such as exclamation marks, words with repetitive characters, or iconic emotions (i.e. emojis and emoticons) that are capable of emphasizing the literal meaning of messages or even reversing it [46]. We follow the previous recommendation and we propose combining different datasets and binarizing them to expand the scope of hate behavior identification on social media. We also propose exploiting transfer learning using pre-trained BERT model as well as OSN-specific emotion hints like iconic emotions, in order to build our hate classifier.

III. BACKGROUND

This section presents a background overview of the methods used to design the proposed framework. The background is composed of three parts: topic modeling, phrase extraction, and deep sequence classification.

A. TOPIC MODELING

Topic modeling is an unsupervised learning technique that detects patterns of words and expressions within datasets, and automatically determines clusters of similar words and phrases that best characterize a set of texts. Recently, topic models have been increasingly used to explore and infer insights from social media data [31], [52], [53]. Latent Drichelet Allocation (LDA) [54] and Non-negative Matrix Factorization (NMF) [55] are examples of the most prominent algorithms for topic modeling in social media analysis.

1) LATENT DIRICHLET ALLOCATION (LDA)

LDA [54] is a probabilistic generative algorithm that utilizes Bayesian framework and Dirichlet distribution. It treats a collection of data as a mixture of latent themes or topics, where each topic is considered a multinomial distribution over a fixed vocabulary. LDA considers two matrices to determine the hidden patterns of topics: document topic density matrix θ and word topic density matrix ϕ. The word matrix ϕ has two dimensions K and V where K is the number of topics and V is the vocabulary size. Any value of $\phi_{k,v}$ represents the likelihood of word $v = 1, 2, \ldots, V$ belonging to topic $k = 1, 2, \ldots, K$. The document matrix θ has also two dimensions K and D where K is again the number of topics and D is the number of documents. A value of $\theta_{d,k}$ signifies the probability with which a topic $k = 1, 2, \ldots, K$ is likely to appear in a given document $d = 1, 2, \ldots, D$. Since LDA uses probability distributions from the Dirichlet family, it requires two Dirichlet priors; one for θ and another for ϕ. Each of the priors is governed by K (i.e. the number of topics) parameter and a prior parameter. It is referred to the prior parameter (i.e. α for ϕ and β for θ) as a model hyper-parameter which it affects the specificity of document-topic and word-topic distributions.

2) NON-NEGATIVE MATRIX FACTORIZATION (NMF)

Unlike LDA, NMF is a deterministic algorithm that uses a decomposition technique for multivariate data where non-negative constraint is necessary for learning topics. It factorizes a high-dimensional data matrix $X = (X_{j,i})$ into lower-dimensional matrices A and B such that $X \approx AB$. The aim of the factorization is to find hidden themes (i.e. topics) within data. The values of X, A, and B and their coefficients are non-negative. The X matrix is a term-document matrix with dimensions $D \times W$ where D is the number of documents and W is the number of words in the corpus vocabulary. $X_{j,i}$ represents the frequency of word j_{th} in document i_{th}. The frequency of words can be replaced by their corresponding TF-IDF weights. A is a document topic matrix with dimension $D \times K$ and B is a $K \times W$ word topic matrix, where K is the number of topics. A and B are computed by optimizing a loss function that is solved using gradient descent methods. Since we are dealing with large and unstructured datasets, we use the NMF algorithm developed by Renbo and Vincent [56]. The algorithm is optimized to handle the issues of processing large datasets and the existence of outliers.

B. PHRASE EXTRACTION

Phrase extraction is a process concerned with the automatic extraction of a set of representative phrases that express the aspects of textual contents [57]. Supervised and unsupervised methods have been widely used for phrase extraction [58]. In this work, we are interested in studying the unsupervised learning approach as it does not require annotated data. Manual data annotation for phrase extraction is prone to human subjectivity as well as it is inefficient; it not only takes a lot of time and requires a lot of effort, but it is also costly. Statistical and graph-based ranking approaches have been widely adopted to extract phrases from textual collections. TFIDF at n-gram [59] level is a well-known method used for statistical-based phrase extractions. However, one of its drawbacks is that it requires large data to produce good results. In addition, it needs to be combined with n-gram technique in order to process multi-word phrases, and this is computationally expensive and time consuming especially when using longer n-grams. Furthermore, n-gram considers n consecutive words but does not take into consideration the occurrences of words in a complete phrase or sentence. TextRank [60] and SingleRank [61] were among the first graph-based algorithms that were developed for phrase extraction. They use words co-occurrence information in order to find candidate phrases. Later on, SGRank [62] and PositionRank [63] algorithms proposed to incorporate statistical and positional information along with the information of words co-occurrences. These algorithms rely on natural language processing (NLP) techniques like POS and n-grams to form key phrases. They utilize POS tagging to use lexical units of specific part of speech limited only to nouns [60]–[63], adjectives [60]–[63], or verbs [62]. Given the short expression and multilingual nature of social media data, this introduces two limitations: (1) ignoring important information residing in different lexical units other than nouns, verbs, and adjectives, (2) increasing the resource cost of having different POS tagging system for different languages. Another limitation of the previously mentioned algorithms is that they analyze words co-occurrences within a fixed sliding window. This disables the flexibility of fine-grained measurement for words associations within a collection of data (i.e. individual data subsets of topics). Rapid Automatic Keyword Extraction (RAKE) [64], a domain-independent language-independent algorithm for phrase extraction, is able to overcome the limitations in the previous studies. It undoubtedly fits the scope of this study for four reasons. First, RAKE overcomes the TFIDF limitation on small datasets inasmuch as it is designed to perform on dynamic-size individual documents (i.e. topic data subsets) rather than on the entire corpus. Second, it is not constraint to specific language structure; this greatly fits the unstructured nature of social media data that does not follow grammar conventions and is full of misspelled words. Third, it reduces the computational overhead of NLP tasks such as POS tagging and n-grams. Fourth, its flexibility allows it to extract phrases of all possible lengths, free of fixed-sliding-window constraint and without the additional computations of n-grams. This is beneficial in exploiting every possible piece of information within short messages which, in turn, will improve the quality of online social data interpretation.

1) RAPID AUTOMATIC KEYWORD EXTRACTION (RAKE)

RAKE [64], a graph-based algorithm, was designed based on the assumption that a key word consists of multiple words that are rarely split by punctuation or stop words. Stop words could be canonical or uninformative words. The remaining words are assumed to be informative and are referred to as content words. RAKE takes two inputs: stop word list and punctuation list (i.e. word punctuation and phrase punctuation). The extraction process starts with splitting a given text into a set of candidate key words at the occurrence of pre-defined word delimiters. Next, the set of candidate keywords is split into a sequence of consecutive words at the occurrence of phrase delimiters and stop words. The consecutive words within a sequence together form a new candidate keyword (i.e. phrase). A graph of word-word co-occurrences is created to be used in computing the scores of the candidate words and phrases. Three scoring metrics were proposed: Word Degree $deg(word)$ calculates the words that have often occurrences in a document as well as in longer candidate phrases, Word Frequency $freq(word)$ computes the words that occur frequently without taking into consideration the word-word co-occurences, and Ratio of Degree to Frequency $\frac{deg(word)}{freq(word)}$. For the purpose of this work, we use the Degree of Words $deg(word)$ as a metric to compute phrases scores.

The score of a candidate phrase is calculated as the sum of its words scores.

2) TEXTRANK

TextRank [60] is a graph-based ranking algorithm used for keyword and phrase extraction. TextRank uses the word co-occurrence statistics to compute scores of words and extract phrases from texts. It uses the co-occurrence information to build a word graph. Two processes are applied before the graph is constructed: (1) words are filtered using POS mechanism and only nouns, verbs, and adjectives are used, (2) a sliding window value is defined. Each word in the graph represents a vertex and an edge between any two words is added if the two words co-occur within the pre-defined sliding window. A weight is assigned to every edge in the graph and its value represents the number of times a word co-occur within the sliding window. Each vertex is assigned with a score that reflects its importance and is computed in an iterative manner using PageRank algorithm. After convergence, the top n scored words are selected as keywords. Phrases are constructed if adjacent keywords are found in the resulted keywords.

C. DEEP SEQUENCE CLASSIFICATION

The advancement of deep neural networks has led to a decent improvement in several Natural Language Processing (NLP) tasks including sequence classifications [65]. Neural networks come with the capacity of mitigating the complexity of feature engineering and provide self-learning of data or feature representations. While memory neural networks with attention mechanism have been widely used to capture the sequential information of texts [66], CNNs [67], [68] have been less popular. Comparing transfer learning in computer vision to that in NLP few years ago, transfer learning in computer vision was by far more successful in performing computer vision tasks. Earlier NLP efforts had been put to exploit previous knowledge by using textual embeddings [69], [70] to avoid restarting training from scratch. Although these embeddings were trained on huge volumes of data, they still suffer from context-independence problem which means that word representations are the same regardless of their surrounding context. More recently, transformer-based language models such as BERT [71] and GBT-2 [72] have made a groundbreaking milestone in transfer learning in NLP. These models are capable of alleviating the complexity of feature engineering and overcoming the limitation of context-independence issue. BERT has achieved state-of-the-art results in learning semantics of textual expressions for various problems including sentiment [7], [36], [39] and hate speech analyses [6], [12], [51].

1) BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

BERT [71] is a multi-layer bidirectional Transformer encoder model. It uses Transformers' attention mechanism [73] to understands the inter-relationship among all words in a sentence. Transformers use an encoder to read input texts and a decoder to produce predictions of given tasks. For BERT, only the encoder is considered since its objective is to build language models. BERT model was built based on three concepts: (1) contextualized word representations, (2) transformers architecture, (3) pre-training language models on large corpus to be used for NLP task-specific fine tuning. Due to its deeply bidirectional contextualization, BERT provides a deeper sense of language contexts than single-direction models do. The contextual representation of a word considers both left and right contexts unlike single direction models that consider only the context of single direction. Two strategies are applied to learn the contextual representations: Mask Language Model (MLM) and Next Sentence Prediction (NSP). Before feeding input tokens into BERT, some percentage of the tokens are randomly masked by MLM model which then predicts the original value of masked tokens only based on the context of unmasked tokens. This way, the information about the predicted tokens is ensured not to leak to next layers. Next comes the role of NSP model where pairs of sentences (A, B) are selected from the data corpus. NSP model trains a binary classifier to predict whether the following sentence B is the actual next sentence of A. This is important to understand the relationship between sentences and to obtain language models that have a deeper sense of a language flow and context. The resulted high-level contextualized word representations are transferable to a downstream of NLP tasks (e.g. task-specific fine turning).

2) LONG SHORT TERM MEMORY (LSTM)

LSTM is an extension of standard Recurrent Neural Networks (RNN) which is capable of learning long-term dependencies between words in sequences. LSTM was designed to overcome the gradient vanishing issue that RNNs suffer from. It was designed with internal mechanism (i.e. gates) that regulates the flow of information. Its architecture consists of three gates (i.e. input gate, forget gate, and output gate) to decide how much information should flow in (i.e. to remember) and out (i.e. to forget) at the current time step. LSTM models process a sentence word by word and assume that each current state depends only on its previous one. LSTMs process a sequence of words in a forward direction where bidirectional LSTMs (biLSTMs) process the textual sequences in both backward and forward directions. This mechanism allows for more information to be available for the network to improve word contextualization. In this paper, we train an LSTM model using GLOVE embeddings on both sentiment and hate classification tasks for evaluation purposes.

3) CONVOLUTIONAL NEURAL NETWORKS (CNN)

CNN is a feed-forward neural network that is biologically-inspired variants of multilayer perceptions. It tends to recognize textual patterns directly from texts with minimum pre-processing applied before feeding sequence of words to the network. A CNN's hidden layer consists of a convolutional layer, pooling layer, and fully connected layer. CNNs

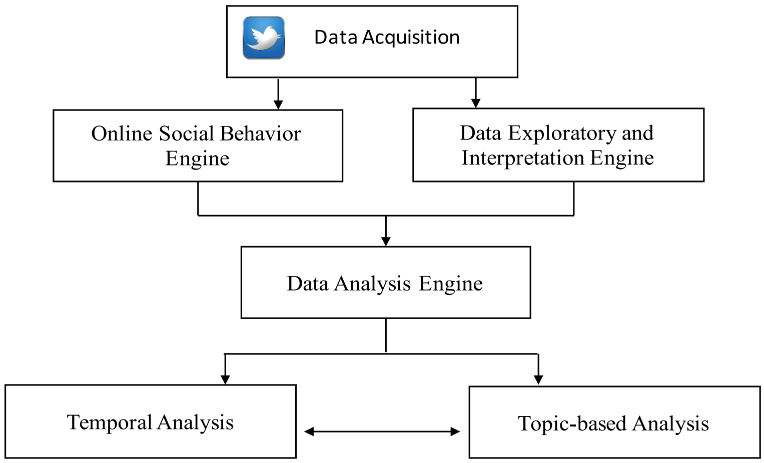


FIGURE 1. Proposed Framework for real-time monitoring online social behavior on OSNs.

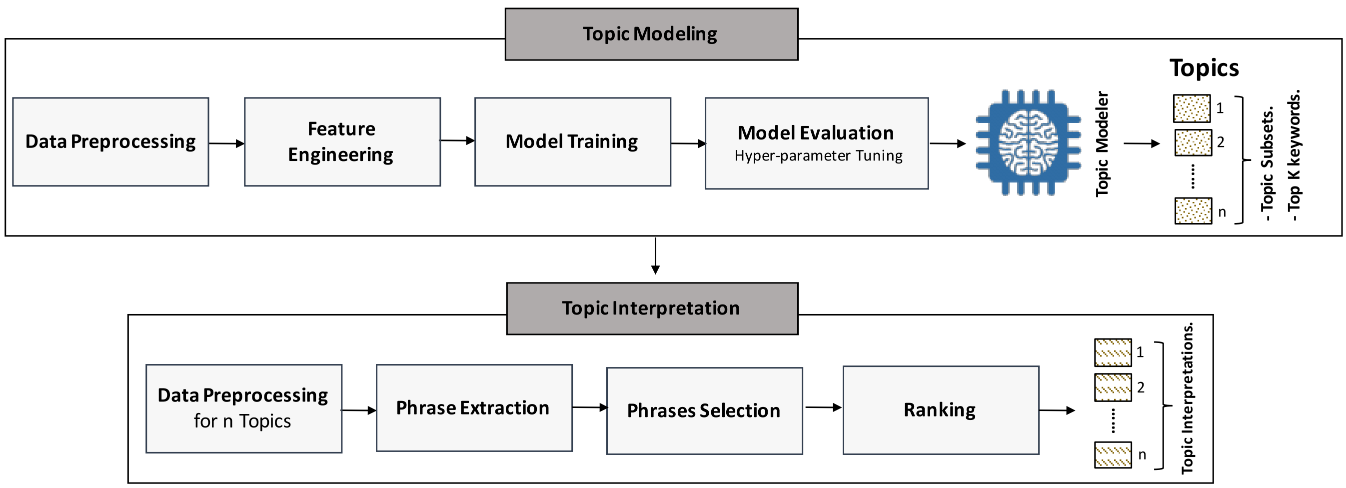


FIGURE 2. Proposed methodology for data exploration and interpretation.

strengthen their power from the convolutional layers that are stacked on top of each other with each one capable of extracting unique patterns independently of prior knowledge or human effort. The patterns could be expressions of multiple sizes (i.e. 2, 3, or 4 adjacent words). In this paper, we build a CNN model with GLOVE embeddings as input features for sentiment and hate classification tasks. We do this step for evaluation purposes.

IV. METHODOLOGY

Figure.1 presents the proposed framework for monitoring social behavior on online social networks (OSNs). The data collected from an OSN platform (i.e. Twitter in this work) is fed in parallel into the online social behavior (OSB) engine and data exploratory and interpretation (DEI) engine. In the OSB engine, the corresponding models- sentiment analyzer and hate analyzer- are developed using supervised learning approach. On the other hand, the unsupervised learning approach is used to develop the DEI models. The data analysis engine uses the outputs of OSB and DEI engines and generates an analytic story from two views: temporal and topic-based analysis. In temporal analysis, the online social behavior is illustrated in a time-line manner (i.e. over days, months, seasons, etc). Theme-based (i.e. topic-based) analysis provides non-linear analysis that is based on the themes and patterns found throughout the given datasets.

A. DATA EXPLORATION AND INTERPRETATION (DEI)

The main objective of this section is to explore and find patterns in social media data and then generate explainable interpenetration of these patterns. Topic modeling is one approach to explore these patterns in large datasets and discover latent patterns (i.e. topic within data). The general framework used in unsupervised learning for topic modeling is followed in this paper. In Figure. 2 - Topic Modeling, the methodology used to build our topic model is illustrated. The data pre-processing of topic modeling is designed based on the criteria to increase the topic relevance and minimize

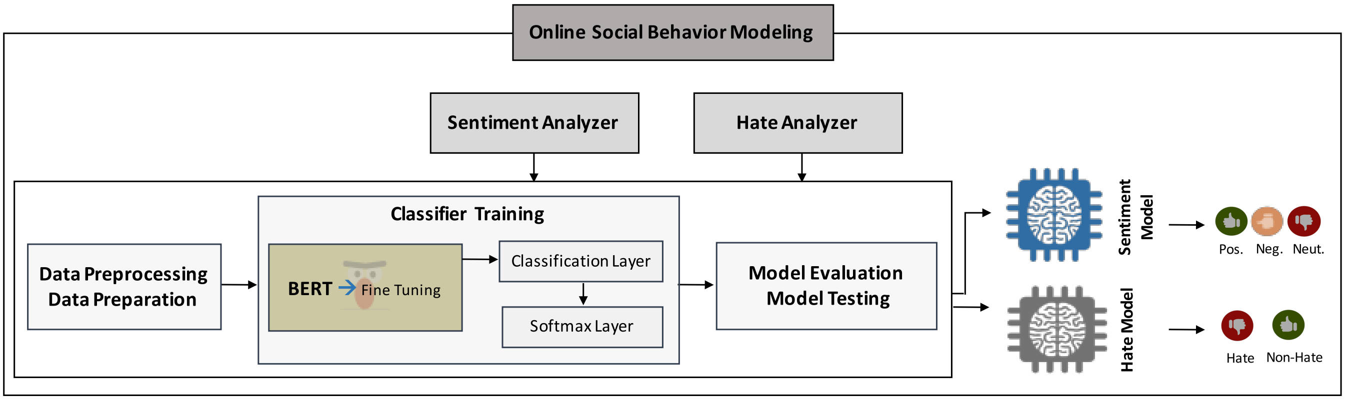


FIGURE 3. Proposed methodology for modeling online social behavior.

uninformative parts of the data. According to the literature [74], the data dominated by stopwords and general uninformative words is semantically uninterpretable as it reduces the reliability and utility of topic models. Accordingly, we consider removing two types of stopwords: (1) canonical words (''the'', ''or'') and dataset-specific words that have a very high and very low usage frequency. Vocabulary is limited to nouns, verbs, adjectives, and adverbs to increase the topic semantic coherence and to minimize the shortcomings of topic modeling algorithms like LDA and NMF, which treat all vocabulary words as having equal importance [29]. We adopt the suggestion made by Lau *et al.* [75] that lemmatizing data improves the topic coherence. The pre-processed data is then used to extract influencing features in the feature engineering component. Different types of features are investigated in the hope of finding effective ones that fit the short and unstructured nature of OSNs data (more details can be found in Section. VIII). The topic model is trained and evaluated using the engineered features. To explore and learn the topics, unsupervised learning approach is used. During the model evaluation, a series of sensitivity tests of hyper-parameter tuning are run in order to find the optimal set of values that produce the highest semantic coherence across topics. The output of the topic model will be k topics. For each topic, we consider the top n keywords and the corresponding subset of data.

After the themes (i.e. topics) have been inferred, they are fed into the topic interpretation component to facilitate the interpretations of the topics; hence providing us with a deeper understanding of the topics automatically. Using the top n keywords only is inefficient in interpreting the coherent meaning of topics [13]. A good interpretation of a topic should convey two characteristics: capturing the meaning of the topic, and distinguishing topics from one another. Single words fall short on these characteristics as they lack the context of phrases and sentences [13]. In more details, single keywords are too general and might miss the semantic relationship to form the main idea of the topics. Phrases, on the other hand, add context to single words, hence providing stronger coherence. Moreover, phrases by nature are broad so they are able to capture the overall meaning of topics [13]. In this paper, we propose the use of unsupervised phrases extraction approach. Automatic Rapid Keywords Extraction (RAKE) algorithm is utilized to find topic phrases. Figure. 2 - Topic Interpretation describes our proposed methodology to extract phrases of topics. First, the data subset for each topic is pre-processed independently. This process is similar to that of topic modeling; however, in phrases extraction stopwords are not removed and all the part of speech tags are not pre-processed (more details are found in IX). Second, RAKE algorithm is used to extract keywords and phrases from each topic data subset. The weights of the extracted keywords and phrases are computed using the degree of word metric $deg(word)$ that calculates the words that have often occurrences in a document as well as in longer candidate phrases. Third, keywords and phrases based on the top n keywords (i.e. resulting from our topic model) are selected. Before selecting RAKE keywords and phrases, the keywords (i.e. resulting from our topic model) duplication across topics are removed. The reason to remove the duplication is to make unique interpretations that distinctively represent each topic. Finally, the keywords and phrases are ranked according to the weights of the corresponding keywords and phrase degrees. The output phrases have various lengths with minimum of two. To choose the optimum length of phrases, the average length of phrases for each topic is calculated. After calculating the average, phrases with more general dimension and phrases with more specific dimension than the average are considered. This is done by selecting shorter phrases and longer phrases than the average length.

B. ONLINE SOCIAL BEHAVIOR (OSB) MODELING

Our study focuses on two types of online social behavior: sentiment and hate behaviors. Figure. 3 illustrates the proposed methodology followed in order to build sentiment and hate analyzers. Supervised classification approach is adopted in the modeling of OSB. The data pre-processing is performed independently of the DEI modeling. The processed data is fed into the training component. The Classifier Training component demonstrates the proposed neural network architecture.

The BERT layer consists of BERT pre-trained embeddings which are representations of words and their relation to each other in n-dimension. BERT pre-trained model is fine-tuned by training the entire BERT architecture on our datasets in order to alleviate possible biases resulted from pre-training on Wikipedia corpus [37]. BERT-base-uncased model is used in this work. It consists of twelve layers and uses 110M parameters. A feed-forward neural network layer used as a classification layer is appended to the BERT layer. The classification layer produces logits that indicate the likelihood of a tweet belonging to a class. Soft max layer is used to normalize the output logits and calculate the probability of classes. The training is conducted by back propagating the errors throughout our architecture and updating the weights of the pre-trained weights and the weights of the appended layer based on our datasets. The models are optimized using Adams optimizer. Early Stopping Approach is used to avoid over-fitting the neural network on the training data and improve the generalization of the models. Finally, we evaluate and test our models on the validation and testing sets before generating the final prediction results. The prediction of sentiment analyzer is one of three classes: positive, negative or neutral sentiments, whereas the hate analyzer predicts one of two classes: hate speech or non-hate speech.

V. DATASETS

This section lists and describes the datasets used for modeling, evaluating, and analyzing the online social behavior framework.

A. SUPERVISED LEARNING DATASET

1) SENTIMENT DATASET

Domain-Free Sentiment Multimedia dataset (DFSMD) [9] was used to train our sentiment classifier. DFSMD was collected using Twitter Stream API. The protocol followed to collect and annotate DFSMD makes it distinguished from other datasets as data collection process was not restricted to any keywords, domains, locations, or any predefined retrieval criteria. The annotation questions and annotators of the dataset were selected carefully to minimize any possible biases during the annotation. Moreover, the annotators of the dataset were selected on the basis of providing sentiment agreement with three expert psychologists. The DFSMD contains 14488 (46%) tweets; 6683 of which are positive, 4822 (33%) negative, and 2983 neutral (21%). The dataset is publicly available upon request. It was published in an earlier study [9].

2) HATE DATASET

In order to train our hate speech classifier, we use four available datasets published in previous studies:

- **HatEval 2019** The English tweet dataset [76] was constructed based on women or immigrants as targets of hate speech in this dataset. The tweet annotations did undergo two steps: (1) by non-expert annotators using crowd sourcing mechanism, (2) then two domain-expert annotators reviewed the annotated tweets. The inter-agreement in annotating the dataset scored 83%. The data set contains a total of 13000 tweets, out of which 5470 tweets are labelled hate speech.
- **OffensEval 2019** The dataset [77] was annotated for categorizing offensive/non-offensive language on twitter. The offensive language is defined as insult or threat contexts. If offensive language is directed towards individuals, groups, or others, it is annotated as hate speech. The annotation was done by domain experts using crowd sourcing approach. The dataset consists of 13240 tweets, 4400 of which are labelled offensive and hate speech.
- **Antigoni Dataset 2018** The tweet dataset [78] was manually annotated using the crowd sourcing approach.
  It consists of hate, abusive, spam, and normal labels. The results have shown that there was confusion between abusive and hate labels during the annotation process, so we decided to combine both under the hate label. We removed the spam label which resulted in a total of 60702 tweets; 28587 of which are normal and 32115 are hateful.
- **Waseem and Hovy 2016** The tweet dataset [79] was annotated for racism and sexism types of hate language. The tweets were reviewed by the authors [79] and then by domain experts.

The labels of these datasets were binarized into two labels: hate and non-hate. This approach was adopted in previous studies [12], [50] and it has been proven effective. The number of hate tweets in our dataset is 39593 where the normal tweets are 46753, which brings the overall total size to 86346 tweets.

UN[2] warns that COVID-19 related posts could be exploited to instigate discrimination, stereotyping, stigmatization, racisms, or xenophobia. All these categories imply abusive and hate behaviors according to several universal definitions of hate language as follows:

– ''bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics'' - Almagor [10].
– ''all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin'' - The European Court of Human Rights [11].
– ''public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin'' - The Code of conduct between European Union and companies.[3]

[3]https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135

– "Hateful conduct: You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease" - Twitter recent social behavior rule.[4] Twitter refers to harassment and abuse rules as "Abuse/harassment: You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm" where it refers to violence rules as "Violence: You may not threaten violence against an individual or a group of people".

B. UNSUPERVISED LEARNING AND ANALYSIS DATASET

Two COVID-19 datasets were collected from Twitter. One dataset was collected for USA and the other for Canada. COVID-19 related keywords were used to retrieve the data. The list of the keywords includes covid-19, covid19, covid, corona virus, corona, and virus. The collection was conducted in three periods over a duration that extends from December 2019 to November 2020: period-1: December 2019 to April 2020, period-2: May 2020 - August 2020, period-3: September 2020 - November 2020. We used geo-location coordinates to define the geographical regions to retrieve the tweets from.

The two COVID-19 datasets were used for modeling the unsupervised learning components of DEI and analyzig the COVID-19 pandemic over the duration between December 2019 to November 2020.

C. PHRASE EXTRACTION EVALUATION DATASET

Tsix dataset [80] was used to evaluate phrase extraction using RAKE algorithm on social media data (i.e. tweets). Tsix dataset consists of 32970 tweets that were categorized into six topics: brexit, election, isis, nobel, note7, and spacex. Each group of tweets (i.e. belonging to a topic) was assigned into a cluster. Each cluster was assigned a summary reference which composes of candidate sentences selected by two human annotators.

VI. DATA PREPROCESSING

Pre-processing data is a very essential step in machine learning in general. It prepares the resource knowledge for the machine models to learn from. High quality pre-processing ensures the quality of the learning process. The objective of pre-processing data is to remove excess noise, which could affect the learning performance, and retain useful information. This work consists of two main components: data exploration and interpretation (DEI) and online social behavior modeling (OSB). Each component is processed independently as each requires a different pre-processing mechanism. Following are the pre-processing steps we propose to use:

1) **Removing HTML encoding symbols**
2) **Removing user mention**
3) **Removing URLs**
4) **Removing Retweets**
5) **Removing extra whitespaces**
6) **Converting text to lower case**
7) **Expanding abbreviations**: to replace abbreviation words with sequence-of-words format. This step processes contraction words "e.g. we're", negation words (e.g. "don't"), and slang words "e.g. ppl, bro".
8) **Tokenizing text**
9) **Fixing repetition**: to remove character repetition and replace it with a single character. For example, word 'niiice' will be replaced by 'nice'. We believe in the importance of this step as it ensures the generality of learning.
10) **Converting iconic emotion into textual format**: to convert emojis and emoticons into textual representations [81].
11) **Removing special characters and numbers**
12) **Removing stop words**: We use two types of stop word lists: (1) standard list by standard libraries like NLTK, (3) customized list that is constructed manually by empirical experiments. According to the literature, this step has a major impact on topic predictions.
13) **Removing words with high and low frequencies**: to remove words of frequencies greater than 60% and less than 10 occurrences per the dataset.
14) **Tagging Part of Speech**: labeling words with grammatical description. We do this step to include only adjectives, adverbs, nouns, proper nouns, and verbs. The aim is to increase the efficiency of topic modeling performance.
15) **Removing short words**: We assume that words with a single character does not have an independent meaning, hence, they do not contribute to the learning process. words of length less than 2 are removed.
16) **Lemmatization:** lemmatization to change each word into its original form. The objective is to reduce the size of vocabulary by conflating terms with related meaning.

Steps 1-6 are applied to the two components, DEI and OSB. Steps 7, 8, 9, 15 are applied to DEI sub-components: topic exploration (e.g. topic modeling) and topic interpretation (i.e. phrase extraction) where steps 11-16 are applied to the topic modeling part and step 11 is partially applied to topic interpretation to keep few punctuation marks (e.g. period). Step 10 is applied only to OSB sub-components. Note that all the pre-processing steps were implemented using regular expressions NLTK, spaCy toolkit. Bert tokenizer was used for the tokenization of OSB sub-components.

VII. EXPERIMENT DESIGN & EVALUATION PROTOCOL

We evaluate the validity of the proposed framework in inferring and interpreting topics of online social media data, as well as in analyzing online social behaviors of social users. Our proposed framework consists of two types of learning:

[4]https://help.twitter.com/en/rules-and-policies/twitter-rules

(1) unsupervised learning for the topics inferences and interpretation, (2) supervised learning for online social behavior analysis. Accordingly, different evaluation protocols are required.

A. TOPIC EXPLORATION AND INTERPRETATION

The objective of the topic inference (i.e. exploration) experiments is to find a topic modeling algorithm and feature type that yield the best performance to find the optimal number of topics within given OSNs datasets. Two types of experiments were conducted for this purpose: (1) studying two topic modeling algorithms: LDA and NMF. (2) studying two types of features: BOW and TFIDF. We investigated the performance of the LDA and NMF with both features BOW and TFIDF on two datasets (i.e. COVID-19 collected during the period 1 - December 2019 to April 2020): (1) COVID-19 – Canada, (2) COVID-19 – USA. This yields a total of eight experiments for the topic modeling. We run a series of sensitivity tests to determine the best values of model hyper-parameters as summarized in Table.1. The tests were performed in sequential manner; one parameter at a time by keeping the others constant and then we run them over the datasets. We used coherence score [82] as a metric for our performance comparison. A coherence score for a topic is calculated by measuring the degree of semantic similarity between high scored words within the topic.

TABLE 1. **The selected hyperparameters that are used for tuning before training LDA and NFM models.**

Model	Parameter	Value
LDA	Number of topics: K	Range from 2 to 14
	Dirichlet hyperparameter: alpha α	Range between 0.01 to 1.
	Dirichlet hyperparameter: beta β	Range between 0.01 to 1.
NFM	Gradient Descent step size: kappa	Range between 0.1-0.5
	Number of topics: K	Range from 2 to 14

For training, words with ≤ 10 occurrences and $> 60\%$ of occurrences in a dataset were filtered out. To assess the topic model performance in finding the optimal size, coherence score was used as an evaluation metric.

For topic interpretation (i.e. phrase extraction), we found out that the phrase length of three has the highest average frequencies in all the topics for Canada and USA datasets. We considered the lengths of two and four to add a more general dimension before and a more specific dimension after, than the phrases of length three. We evaluated RAKE algorithm on Tsix dataset using ROUGE metric [83]. ROUGE, Recall-Oriented Understudy for Gisting Evaluation, is a well-known metric used for evaluating automatic summarization of texts. It works by comparing the automatically generated summaries to human-generated summaries. ROUGE-N metric computes the overlap of n-grams between the automatic and human summaries. In this work, we are interested in evaluating the recall of our phrase extractor by examining the percentage of n-grams, in our generated phrases, that exist in the reference phrases. In addition, we compare RAKE algorithm to other two phrase extraction algorithms, TextRank and TFIDF using the same dataset. In this paper, we set the n-gram to 4-gram and therefore the evaluation was performed using ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-4. In this paper, we use keywords (i.e. single words) and phrases of lengths 2-4 and hence the decision to use 4-grams in this experiments. Accordingly, TFIDF and TextRank models were built at the 4-gram level. We were able to fix the sliding window for TextRank to maximum 3 due to hardware limitation.

B. ONLINE SOCIAL BEHAVIOR MODELING

Sentiment and hate are the two social behaviors studied in this work. We evaluated the performance of BERT for sequence classification algorithm using DFSMD dataset for sentiment learning and Hate dataset for hatespeech learning. We then conducted evaluation comparisons of the BERT-based models with other algorithms for sequence classification including LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM, using the same datasets. Figure. 4 illustrates the architectures that we have used in our experiments.

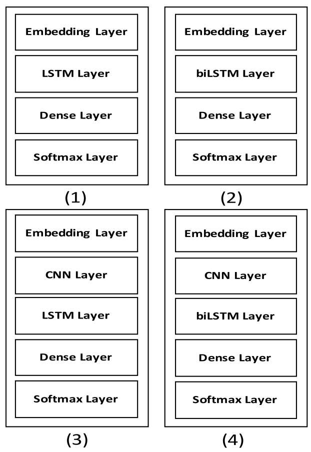


FIGURE 4. **The neural network architectures used for comparison with our BERT-based models. (1) for LSTM, (2) for biLSTM, (3) for CNN-LSTM, and (4) for CNN-biLSTM.**

For BERT pre-trained model, BERT-base-uncased was used. It consists of 12 blocks of transformers, 768 hidden layers, 12 attention heads, 110M parameters. During the fin-tuning process, the learning rate was set to 2e-5, epochs were set to 6, and the batch size was 16.

GloVe [69], a pre-trained word-embedding, was used to train LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM models. The data preprocessing steps and experiment setups were the same as those of the BERT-based models.

The DFSMD and Hate datasets were used for training, validation, and testing sentiment and hate models, respectively. The datasets were randomly split into 70% for training, 15% for validation and 15% for testing. We used accuracy, precision, recall, and F-score, commonly used for classification evaluation, as evaluation metrics. Precision, recall and F-score give a better view of model performance than accuracy alone does.

VIII. RESULTS AND ANALYSIS

This section presents the results of the experiments conducted according to the design explained in Section VII. It will also present online social behavior analysis during COVID-19 pandemic for Canada and USA. Two types of COVID-19 analysis will be provided: temporal and topic-based analysis.

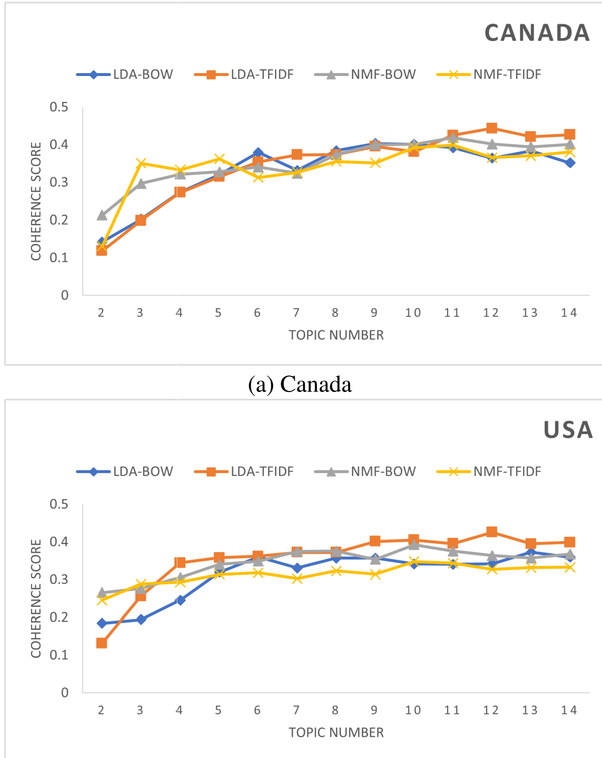
A. TOPIC MODELING

We present the results of the methodology we follow for modeling key topics using OSNs short texts. For this experiment, we use two datasets related to COVID-19; one for Canada and another for USA. It is worth mentioning that stopword filtering has shown a major impact on the overall topic modeling learning. The following experiments were conducted on all the data after removing stopwords.

1) LEARNING FEATURES

To find the best representative features, we have considered using three types of textual features: Bag-of-Words (BOW), Time-Frequency-Inverse-Document-Frequency (TFIDF) and *n*-gram. Our empirical experiments have shown an improvement in learning when *uni*-gram and *bi*-gram were combined together. To find the optimal number of topics, two experiments were conducted by combining uni-bi-gram with BOW and TFIDF. Figure.5 illustrates the performance results of LDA and NMF models using BOW and TFIDF in order to find the best topic model for OSNs data. The BOW and TFIDF features were constructed based on uni-bi-gram features. According to the results, LDA model performed better with TFIDF features than it did with BOW features in both datasets. Unlike LDA model, NMF model performed better with BOW than it did with TFIDF.

Overall, LDA model trained using TFIDF features outperformed NMF model trained using BOW features. For Canada dataset, LDA-TFIDF model maintained the highest scores for topic sizes of 11 to 14. It also scored the highest for topic size of 7. Similarly, LDA-TFIDF model maintained the highest scores for topic sizes of 4 to 14 in USA dataset. This shows that TFIDF method works well with OSNs texts since they are huge in volume while short in length and hence the content is limited per message post. As a result, the need to spot influencing words, to learn different topics, arises. TFIDF method, which plays at word level, measures the relevance of words but not the frequency; it represents documents about 'computers', for example, far from documents about 'batteries'. This gives it the advantage of choosing influencing vocabulary and reducing the complexity of training since using the entire vocabulary in training [84] is expensive. By using the TFIDF weights, the chance that rare words are sampled would increase (i.e. which is the goal to improve the topic learning on short texts documents of a large size). This results in making them have a stronger influence on topic



(a) Canada

(b) USA

FIGURE 5. Performance of four topic models to find the optimal topic size for Canada and USA, in terms of coherence score.

TABLE 2. The optimal hyperparameters values for LDA and NFM models, resulted from hyper-parameters tuning process using two COVID-19 datasets.

Model	Parameter	Best Value
LDA	Number of topics: K	6-14
	Dirichlet hyperparameter: alpha α	**0.01 for both BOW and TFIDF**
	Dirichlet hyperparameter: beta β	**0.51 for both BOW and TFIDF**
NFM	Gradient Descent step size: kappa	**0.2 for BOW, 0.3 for TFIDF**
	Number of topics: K	7-14

assignment. This is the same reason why it is recommended to remove stop words before training an LDA model.

2) TOPICS SIZE

Table.2, shows the results of the hyper-parameter tuning for the LDA and NMF models. The best values of the model hyper-parameters have been chosen based on the top 10 highest coherence scores. The set of hyper-parameters values that has the majority of the highest coherence score are selected. The values of LDA alpha α and beta β and NMF kappa are shown to be fixed for the top 10 highest coherence scores. For the number of topics parameter K, it ranges between values of 6 and 14. However, no exact value was given. Therefore, we selected the best hyper-parameters values as suggested in Table.2 to determine the exact number of topics for topic modeling training. Note that we select 2 as the minimum number of topics and 14 as maximum.

TABLE 3. **The optimal number of topics for Canada and USA datasets for three periods during the pandemic. The optimal topic sizes were determined based on the highest coherence scores for each dataset.**

	Canada		USA	
	Topic Size	Coher-Score	Topic Size	Coher-Score
Period-1	12	0.44	12	0.43
Period-2	13	0.4	10	0.4
Period-3	10	0.43	11	0.5

We then run another process of tuning with respect to the number of topics. Figure.5 illustrates the performance of LDA and NMF models over a range between 2 and 14 topics with the hyper parameters alpha α and beta β fixed at 0.01 and 0.51 for LDA models, and kappa fixed at 0.2 and 0.3 for NMF-BOW and NMF-TFIDF models, respectively. Twelve topics are shown to score the highest coherence by LDA-TFIDF model for both datasets before the performance decreases and flattens out. Similar observation was claimed by Yuxin *et al.* [85] where LDA showed more robust performance on full sentences than NMF did. To find the optimal number of topics for the other two periods (i.e. May 2020 - August 2020, September 2020 - November 2020), we follow the same steps mentioned previously using LDA-TFIDF. Table. 3 summarizes the optimal number of topics resulting from LDA-TFIDF models for Canada and USA datasets. The results represents three periods during the pandemic: December 2019 - April 2020, May 2020 - August 2020, September 2020 - November 2020).

Table.4 shows the results of our LDA-TFIDF model on Canada and USA datsets. We select the top 30 keywords in this paper but due to a limited writing space, we only list the top 20 keywords of sample topics inferred by our LDA-TFIDF model during for three periods during the pandemic. Note that the keywords listed in the table are before removing the duplications across topics.

3) TOPIC INTERPRETATION

Table. 5 shows the performance of phrase extraction models using a tweet dataset (i.e. TSix). The evaluation was conducted on the top 30 phrases resulted from each model. At 4-gram level, RAKE model has been shown to outperform TFIDF and TEXTrank models in terms of execution time and ROUGE-n recall scores. RAKE model was able to recall 80% of the single keywords existed in reference sentences whereas TextRank and TFIDF recalled 60% and 50% of the keywords presented in the reference sentences. At 2-gram level, RAKE model was the winner in recalling 60% of phrases with length 2, followed by TextRank model with 40% of length-2 phrases recall and 25% recall by TFIDF model. Similarly with phrases of lenth 3 and 4, RAKE model yielded the best performance followed by TextRank and TFIDF models. Even though TextRank was shown to be the second best, it required $\approx$ 12 more times (i.e. 972 seconds) than RAKE

TABLE 4. **Top 20 keywords for sample topics inferred by LDA-TFIDF model, for Canada and USA datasets, during three periods of the pandemic.**

	Canada	USA
	Period 1	
Topic 1	love, quarantine, life, hope, home, stayhome, pandemic, walk, safe, dog, work, fun, hair, isolation, toronto, enjoy, song, kid, ontario, normal, friend	quarantine, nyc, lockdown, song, york, stayhome, pandemic, quarantinelife, love, youtube, stayathome, brooklyn, movie, home, street, walk, elon_musk, photo, fun, staysafe
Topic 2	trump, china, americans, chinese, president, world, response, america, blame, bad, death, die, country, lie, government, kill, pandemic, vaccine, fauci, believe, white_house	trump, china, americans, president, obama, lie, blame, response, death, chinese, die, election, country, america, pandemic, world, hoax, government, state, kill
Topic 3	canada, news, cure, world, petition, big_banks, canadian_authoritie, publish, media, expert, chronic_ink, elon_musk, poor, investigation_stock, fight, warrant_senator, superbug, contact_trace, conspiracy_theorie, cargill_close	spread, sick, die, mask, stop, home, bad, work, wear_mask, stupid, safe, kid, open, wear, test, life, wearing_mask, understand, risk, crazy
	Period 2	
Topic 1	case, death, test, mask, report, health, number, die, spread, vaccine, positive, ontario, canada, school, flu, wear, cdc, testing, risk, news, patient, confirm	pandemic, vaccine, test, health, death, patient, work, school, case, risk, impact, cdc, testing, learn, die, business, crisis, spread, report, care
Topic 2	hair, beautiful, pretty, wear, idk, mask, cut, hurt, eye, face, head, black, white, love, brain, nose, ask, point, cough, blue, lady, cat	nan, love, eat, bad, food, taste, chicken, thread, pizza, dinner, karen, cook, slogan, cake, meat, ready, boo, lunch, yup, ugh
Topic 3	trump, die, china, americans, death, president, kill, country, lie, america, dead, bad, vote, stop, world, response, state, blame, pandemic, hoax, election, biden	trump, death, die, americans, china, kill, president, country, dead, america, lie, biden, spread, vote, pandemic, response, state, rnc, mask, stop
	Period 3	
Topic 1	trump, biden, election, vote, president, americans, die, stop, joe, lose, country, america, lie, win, disappear, believe, plan, death, rally, task_force, president_elect, hoax	trump, die, americans, biden, country, kill, lose, dead, president, lie, care, american, china, vote, economy, life, america, election, job, death
Topic 2	pandemic, business, impact, work, support, health, job, community, learn, pay, challenge, market, canada, program, money, plan, care, crisis, service, recovery, economy, government	mask, wear, spread, celebrate, rally, trump, stop, super_spreader, event, biden, party, crowd, celebration, safe, social_distancing, street, social_distance, forget, supporter, sick
Topic 3	vaccine, pfizer, effective, news, early, trial, pfizer_biontech, cure, study, test, mink, denmark, science, announce, infection, mutate, human, antibody, research, candidate, develop, data_signal	vaccine, pfizer, test, positive, effective, ben_carson, news, election, disappear, pfizer_biontech, chief_staff, mark_meadow, trump, trial, early, result, white_house, announce, percent_effective, analysis

TABLE 5. A comparison between RAKE algorithm and TFIDF and TextRank algorithms for phrase extraction using Tweet TSix dataset. The performance is evaluated in terms of execution time and ROUGE-n recall metric; where $4 \geq n \geq 1$.

	TFIDF	TextRank	RAKE
ROUGE-1	0.5	0.6	**0.81**
ROUGE-2	0.25	0.41	**0.6**
ROUGE-3	0.14	0.31	**0.44**
ROUGE-4	0.06	0.21	**0.32**
Execution Time (seconds)	82	972	**81**

(81 seconds) did. Other limitations of TextRank are fixed sliding window and phrase length boundaries; it is not able to dynamically extract phrases of varying lengths. This is shown in its ability to recall single words (60% of recall) while its performance deteriorated for phrase extraction. Like TextRank, TFIDF model also suffer from the limitation of phrase length boundaries; n-gram has to be fixed prior phrase extraction. In addition, TFIDF model has shown to have a weak performance in extracting keywords and phrases from small-sized data. The average data size per a cluster (i.e. document) in TSix dataset is 36 tweets. This is not surprising as TFIDF algorithm needs a decent amount of data in order to find influencing words with respect to a whole document.

These results show the effectiveness of RAKE model for phrase extraction on OSNs data. Its advantages lay in its dynamic ability to extract phrases of varying lengths (i.e. not constrained to fixed sliding window nor fixed phrase length) and it works at a message level (i.e. it is not constrained to either small nor large datasets).

Table.8 and Table.9 list the results of our RAKE-based phrases extracted from our COVID-19 datasets. The tables show a sample of topics keywords and phrases for Canada and USA during the three periods defined in this paper. The keywords in the Tables are the output of our LDA-TFIDF model after removing the duplicates across the topics and then ranking them based on their RAKE weights. We can detect an improvement in the quality of the keywords after removing the duplicates. For example, in Canada topic 3 - Period-1 (Table. 4), we see that the top 4 keywords are common in other topics. This duplication not only degrades the quality of the topic description but also adds ambiguity. After removing the duplications we notice that the comprehensibility of the topic has enhanced (Table.8). RAKE produces scores based on the frequency of phrases in each topic. Therefore, the high frequency of phrases of a certain topic (i.e. after choosing the most influencing words resulted from the LDA-TFIDF model) indicates that people are indeed talking about this topic. Thus, it has become clear that the topic 3 discusses Elon Musk and Tesla. Additionally, we see that the inferred phrases (Table.8) take it further and show that topic 3 is about Elon Musk and Tesla in the Canadian news (e.g. "social media bbc news", "cure cbs news") and that Elon Musk might reflect negative vibes (e.g. "elon musk threaten", "fight elon musk tweet") during the pandemic in Canada.

Looking at the keywords of topic 4 in Canada (Table.8), we are unable to understand what the main idea of the topic is. They only provide names of prominent figures like British Prime Minister "Boris Johnson", American government official "Stephen Miller", and public events like "UFC" (Ultimate Fighting Championship). Thanks to the topic interpreter that has provided details about the keywords by adding contexts, which facilitated the understanding of the topic. For instance, phrases of length 3 have added some information about Boris Johnson; the phrases "boris johnson die" and "save boris johnson" bring forth the idea that Boris Johnson is undergoing some critical situation. The idea of the topic is wrapped up in phrases of length 4; now we know that Boris Johnson has been tested positive "boris johnson tests positive", and this explain the phrases "boris johnson die" and "save boris johnson". In the case of Stephen Miller, however, we are able to conclude that he was tested positive "stephen miller test positive" in a phrase of length 4 only while no information was mentioned in phrases of lengths two and three (i.e. top 10 phrases of length 2 and 3). With respect to UFC, phrases of length 2 are about cancelling ufc: "ufc cancel". Phrases of length 3 add more details to phrases of length 2; we now know that the fight game cancellation is associated with the player Jacare, a Brazilian mixed martial artist,: "fightcancelled jacare ufc". Also UFC has something to do with testing positive "ufc card test positive" (i.e. a phrase of length 4). We already have the knowledge, from phrases of length 3, that UFC game cancellation might be associated with the player Jacare. Put together we can infer that Jacare might have been tested positive and that is why the game was cancelled.

Another example can be seen in USA Topic 4 - Period-1 (Table.9). Phrases of lengths 2, 3 and 4 add more dimension to the topic readability and comprehensibility than single keywords do. In phrases of length 2, we notice that sport season is the topic talked about the most. Phrases of length 3 explain further and mention spring sport season, high school sport, sports season during pandemic and NFL/NBA season, for instance. They reveal also that the topic is about travel season and future travel plans. Phrases of length 4 add up more details to the season, sport and travel that they were cancelled. Additional information was also revealed, in phrases of length 4, that baseball was among the sport activities in high schools. It is worth noting that such details are not highlighted when using LDA's single keywords exclusively.

Phrases of maximum lengths are discarded since long phrases (i.e. sentences) does not provide the overall meaning of the topics. instead it only shows one part of the whole topic. A case example can be seen in Canada topic 4 - Period-1 (Table.8). The maximum-length phrase says "prime minister british prime minister boris johnson tests positive". The long phrase only shows one part of the topic and did not show the whole picture that the topic was talking about famous figures who tested positive. This finding is supported by the

TABLE 6. The performance of five deep learning algorithms for sequence classification for hate classification in terms of accuracy, precision, and recall.

	LSTM			biLSTM			CNN-LSTM			CNN-biLSTM			BERT		
	Accuracy	85		Accuracy	85		Accuracy	85		Accuracy	85		Accuracy	**86**	
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Normal	0.83	0.9	0.87	0.85	0.88	0.86	0.84	0.9	0.87	0.86	0.86	0.86	0.88	0.84	0.86
Hate	0.87	0.79	0.83	0.85	0.81	0.83	0.87	0.8	0.83	0.84	0.84	0.84	0.85	0.89	0.87

results obtained by Qiaozhu [13] that sentences might not be accurate to capture the general meaning of a topic as they might be too specific.

The results show the effectiveness of our topic models and phrase extractors in automatically identifying and interpreting topics inferred from OSNs data. It has big benefits in minimizing the human intervention in identifying and interpreting topics which in turns facilitates the real-time topic modeling and interpretation with minimized need to human approvals.

B. ONLINE SOCIAL BEHAVIOR

The results of modeling two online social behaviors, sentiment and hate, are demonstrated and discussed in this section. Following, a detailed analysis of the sentiment and hate online behaviors during the pandemic, is provided for the duration between December 2019 and November 2020 for both Canada and USA. The analysis presents two views: temporal and topic-based analysis.

1) HATE BEHAVIOR

Table.6 summarizes the learning performance of four algorithms for sequence classification using pre-trained GLOVE embeddings as features, and compares them with our proposed BERT-based classifier using its pre-trained embedding as features. It is important to mention that the embedding features were fine-tuned during the training. In terms of accuracy, it is shown that BERT-based classifier performs the best in detecting normal and hate contents in social messages compared to the other four algorithms. The other four algorithms show equal performance in terms of accuracy. However, LSTM and biLSTM classifiers show a bias towards learning the majority class (i.e. normal class). This can be seen in the relatively high variance between the recall values of normal and hate classes with the normal class value being the higher. Also, the lower value of normal precision and hate recall explains that the LSTM and biLSTM models have over learned the majority class (i.e. normal), and hence started to introduce a degree of confusion in correctly classifying hate class (i.e. the minority). Adding a layer of CNN to LSTM and biLSTM has helped solve the issue of bias learning and improve the overall learning performance in detecting hate speech in texts. Combining CNN as a mechanism to find important features, and bidirectional learning mechanism has shown to enhance the hate learning process, more than when combining CNN with a single-direction LSTM. The variance

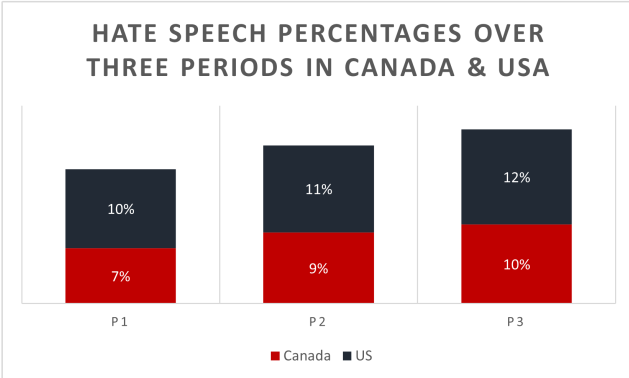


FIGURE 6. Percentages of hate behavior in Canada and USA during periods 1 (Dec 2019 - Apr 2020), 2 (May 2020 - Aug 2020), and 3 (Sep 2020 - Nov 2002) of COVID-19 pandemic.

between the recall values of normal and hate classes has decreased while still maintaining high scores of recall and precision for both classes. In comparison with CNN-biLSTM classifiers, BERT calssifier has shown more robust capabilities in tackling the issue of bias learning towards the majority class. The attention and bidirectional mechanisms adopted by BERT algorithm have proved their effectiveness in improving the quality of learning the two classes, more than CNN and bidirectional mechanisms could achieve. From Table.6, we can see that the hate F-score of BERT model has improved from 0.84 (of CNN-biLSTM) to 0.87 and precision scores for both classes have improved from 0.86 and 0.84 (of CNN-biLSTM) to 0.88 and 0.85, respectively.

From the results, we can see that the attention and bidirectional learning mechanisms adopted by BERT show more efficiency in textual sequence classification than CNN combined with bidirectional learning mechanisms do.

We have utilized our BERT-based hate classifier to analyze the hate speech behavior during COVID-19 pandemic in North America. Figure.6 shows the overall hate behavior detected in Canada and USA during three periods of the COVID-19 pandemic. We see that hate behavior in USA is slightly higher than it is in Canada.

In Figures. 7 and 8, an exploratory analysis of hate behavior in both Canada and USA is demonstrated. We provide a detailed analysis and interpretation from two views: temporal analytic view and topic-based analytic view.

Figure.7 illustrates online social hate behavior at the very early signs of the virus and during the pandemic

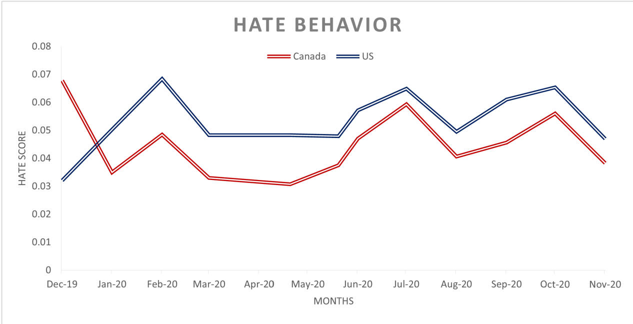


FIGURE 7. Temporal comparisons of hate behavior over twelve months between Canada and USA before and during COVID-19 pandemic.

(December 2019 - November 2020). Generally speaking, online hate behavior seems to have been lower in Canada than in USA except for the month of December 2019; the people in Canada seem to have been more upset about the virus. However, the number of COVID-related tweets during this month was very low (154 tweets) compared to the number of tweets afterwords (4M+ tweets). In USA, hate behavior marked the lowest in December 2019, and that was when the news started to talk about the novel Coronavirus just before the outbreak of the pandemic. Then, the hate behavior started to increase (in USA) till it hit its highest in February 2020 which marked the beginning of the pandemic. It also increased during February 2020 in Canada. By this time, the corona virus had been all over the news and people started to panic. Surprisingly, during the start of the quarantine and lockdown (i.e. March), the hate behavior decreased in both Canada and USA. Later from April till May 2020, it slightly increased in USA during April and then it decreased again in May. In Canada, it almost flattened out after March. From June till November, it can be seen that hate behavior was higher than it was before May especially for USA; two spikes were found in the months of July and October for both Canada and USA as seen in Figure. 7.

Out of the twelve topics discussed during period 1 (Dec 2019 - April 2020) of the pandemic in North America (Figures. 8a and 8b), we have noticed that Canada has a single hate spike (i.e. topic of 'Trump&China') while USA has two hate spikes (i.e. topics of 'Social Distancing' and 'Trump&China'). The inferred topics for Canada and USA are listed in Table.8 and Table.9, respectively. Note that a sample of the topics are reported due to the limited writing space in the paper.

'Trump&China' topic in Canada and USA focus mainly on Trump's blaming China for the COVID virus outbreak. On this topic, USA shows higher hate score than Canada does. From our extracted topic phrases, we can see that among the mostly used phrases in this topic are: 'lie trump', 'trump blames china', 'trump wartime president', 'president trump shame', and 'president trump beat china'. The second highest spike in USA is on the topic related to social distancing and wearing mask. Phrases like 'social sick, 'social crazy', 'fear mask', 'stupid trump' provide a hint that people were not happy with social distancing and the policy of wearing masks. Again, we see hate content in topic 'Face-Masks&Food-Stores' that talks about wearing masks while shopping at stores. In Canada, We observe a less hate behavior for wearing mask. Apparently, wearing masks policy has an association with hate content during the pandemic. Moreover, People in USA did not show hate speech in the topic of 'Community-Health Support' While in Canada we detect a slight increase in the hate behavior on the 'community support' topic (i.e. as compared to USA). A sample of tweets related to this topic (i.e. in Canada) is given below:

- "don't forget, he has a habit of repurposing emergency funds to his wall. Watch for the noise to start about immigrants carrying the virus and showing up at the southern border".
- "Child care staff who are overworked and grossly underpaid will once again be left holding the bag. Close childcare centers, too!".
- "How stupid you guys looked when @jkenney announced medical emergency in Alberta. Please hold your horses now and see why it was important to do emergency you idiots".

COVID-19 death toll topic did not show a sign of hate speech in Canada while the hate speech on the same topic increased in USA. The tweets related to this topic were mostly news. From this, we observe that the Canadian news have less sharp reporting tone than that of the American's. Quarantine and staying home topic shows a very low hate behavior in both Canada and USA. Actually, our topic model and interpreter showed that people enjoyed being quarantined; keywords like 'song', 'movie', 'fun', 'dance, 'dog', 'walk', 'laugh', 'stayhome', 'staysafe', are indicators that the quarantine could have been associated with spending good time while staying home. Cancelling sport game events, a topic talked about during the pandemic, seems to have upset people in USA more than it did with people in Canada. The effect of the pandemic on economy was also discussed in Canada and USA. People in both countries showed low hate behavior. 'Pandemic-State in Quebec' topic in Canada was related to the home care in Quebec province. We observe a slight increase in the hate behavior in this topic and this might be associated with the high number of cases that hit Montreal city particularly.

It is observed that people were more adapted to the quarantine during period 2 (May 2020 - Aug 2020) more than they were during period 1 of the pandemic; topics were mostly discussing songs, movies, tv shows, birthday, hair, and food as seen in Figures. 8c and 8d. These topics have mostly shown low hate behavior compared to topics directly related to the pandemic such as 'Trump&China' and 'Sympathy Attitude' where people have expressed blames and resentments as a result to friends, family or jobs loss for example. "OMG! THATS SO, SO HORRIBLE.. I'M STUNNED. WTF KIND OF WORLD IS THIS? SO MANY CONDOLENCES TO ALL FAMILY, FRIENDS AND THE

FIGURE 8. **Comparisons of Hate behavior between Canada and USA over topics inferred during periods 1 & 2 & 3 of COVID-19 pandemic. P1 (Dec 2019 - Apr 2020) is depicted in (a) and (b), P2 (May 2020 - Aug 2020) is dipected in (c) and (d), P3 (Sep 2020 - Nov 2020) is depicted in (e) and (f).**

MANY ANIMALS LEROY TOUCHED SUCH A WASTEFUL, STUPID LOSS'' and ''I lost around 12 friends to the virus be it online friends, lose them it's painful. You feel anger, sadness; great loss, you go through the stages of grief. Trump would be acting way differently if he lost friends?'' are examples of how resented people were. Similarly, 'Hair' topic has shown a slight sign of hate speech as a result of barber shops being closed (e.g. ''i need someone to get rid of the hair on my scalp i hate it'', ''I hate his hair'').

The curve in Figure. 8e illustrates that the hate behavior in Canada is more relaxed than it is in USA (Figure. 8f). Topics inferred during period 3 (Sep 2020 - Nov 2020), from USA, reflect the political situation (i.e. elections) in the area. five out of eleven topics are related to Trump, Biden, and election with Trump related topics representing the highest hate signs. Similar behavior is found in ''Trump&Biden Election' topic in Canada as well. Interestingly, community support related topics, in Canada, have shown lower hate scores in period 2 and 3 compared to its score during period 1.

2) SENTIMENT BEHAVIOR

In Table.7, we report the performances of five deep models for sentiment classification. The first four models were trained using pre-trained GLOVE embeddings as features and BERT-based model was trained using its pre-trained embedding as features. It is important to mention that the embedding features were fine-tuned during the training.

Table.7 shows that BERT-based classifier outperforms LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM in sentiment learning. BERT model yields by far the best learning results for the three classes with F-scores of 0.81, 0.82, 0.48 for positive, negative, and neutral, respectively. The nature of BERT algorithm, that is using attention and bidirectional learning mechanisms together, allows it to boost the learning performance across the three classes in the presence of imbalance class distribution. BERT algorithm has improved the overall learning of majority class (i.e. positive) by $\approx 2\%$ and by $\approx 8\%$, $\approx\%12$ for negative class and neutral class (i.e. minority class). The overall learning improvement

TABLE 7. **The performance of five deep learning algorithms for sequence classification for sentiment classification in terms of precision and recall.**

	LSTM			biLSTM			CNN-LSTM			CNN-biLSTM			BERT		
	Accuracy	71		Accuracy	71		Accuracy	70		Accuracy	71		Accuracy	**75**	
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Positive	0.79	0.79	0.79	0.73	0.86	0.79	0.78	0.78	0.78	0.78	0.8	0.79	0.81	0.81	0.81
Negative	0.72	0.8	0.76	0.74	0.76	0.75	0.73	0.77	0.75	0.73	0.79	0.76	0.81	0.83	0.82
Neutral	0.46	0.38	0.41	0.49	0.28	0.36	0.46	0.43	0.44	0.48	0.39	0.43	0.5	0.47	0.48

is evaluated in terms of F-score metric. It is strongly able to distinguish between classes especially for positive and negative. Both precision and recall scores for both classes are very high. Neutral class has been always challenging in sentiment classification [5]. However, BERT classifier was able to correctly recall 47% of the neutral instances at 50% of precision. The corresponding confusion matrix shows that the BERT model predicts 31% of neutral class as positive and 21% of neutral class as negative. This is not surprising since negative expressions tend to be strongly subjective. Therefore, positive expressions would be closer to neutral than negative sentiment [5].

We have observed that biLSTM introduces bias in learning the majority class. The positive recall score is higher compared to the negative and neutral recall scores, which makes the variance between them high as well. The corresponding confusion matrix shows that the model predicted 15% of the negative class as positive and 8% of the negative class as neutral. Similarly, with neutral class, 46% was predicted as positive and 25% as negative. We have observed the same behavior in biLSTM when modeling the hate behavior (Table.6).

Using CNN as a filtering mechanism along with biLSTM has demonstrated good performance in reducing the sensitivity to class imbalance shown by biLSTM and in improving the learning of the minority class (i.e. neutral). The neutral F-score has improved from 0.41 (i.e. when using LSTM only), 0.36 (i.e. when using biLSTM only) to 0.44, 0.43 when CNN was combined with LSTM and biLSTM, respectively. This shows evidence that CNN was able to find important features and filter out unimportant ones. Feeding the important features to LSTM with bidirectional mechanism has proven to slightly enhance the learning of sequence classification for sentiment analysis on imbalance dataset.

According to our results, we have found that BERT algorithm provides robust capabilities for sequence classification for sentiment and hate speech. It has shown excellent performance in learning binary classification and multi-class classification. In addition, it has been proven effective in dealing with class imbalance as discussed previously in the results.

Figures. 9 and 10 illustrate the sentiment predictions of our BERT-based sentiment classifier and provide two views of exploratory analysis, temporal and topic-based. Overall during the pandemic, the sentiment tends to be more negative in USA than it is in Canada (as seen in Figure. 9).

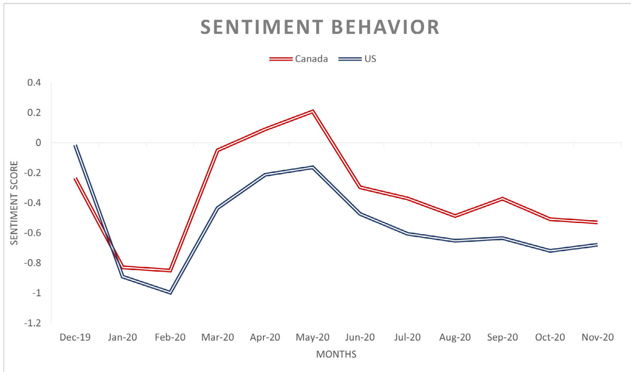


FIGURE 9. **Temporal comparisons of sentiment behavior over twelve months between Canada and USA before and during COVID-19 pandemic.**

The temporal analysis of COVID-19 in North America has shown a negative behavior since the very beginnings of the pandemic that is back in December 2019 when the negativity is shown to be higher in Canada than in USA. However, the number of COVID-related tweets during this month was very low (154 tweets) compared to the number of tweets afterwords (4M+ tweets). The negative behavior increased during the months of January and February 2020. Then it decreased over the next three months with Canada exiting the negative zone and entering the positive zone, as depicted in the figure. After May 2020 we again witnessed an increase in the negative behavior for both Canada and USA.

To facilitate the understanding of the sentiment behavior over time, we provide a deeper analysis about the topics that people were discussing during the pandemic on OSNs platforms (Figure. 10). Having the topics at hand, the reasoning of temporal analysis can be achieved. In other words, we can understand the reasons and causes of behavior changes over time. This will help clarifying the story of events. Quarantine topic has shown a high positive behavior in both Canada and USA, for periods 1, 2, and 3, with Canada showing more positive vibes. This is compatible with the increase in the sentiment positivity in the months of March, April, and May 2020. During these three months, the discussed topics were mostly related to quarantine and staying home. This also explains the low hate behavior that we found in these topics. Our topic model and interpreter confirm this by inferring topics including watching TV, reading books, cooking and baking, as well as providing a set of keywords and phrases (e.g. 'fun', 'laugh', dance, 'favourite love song

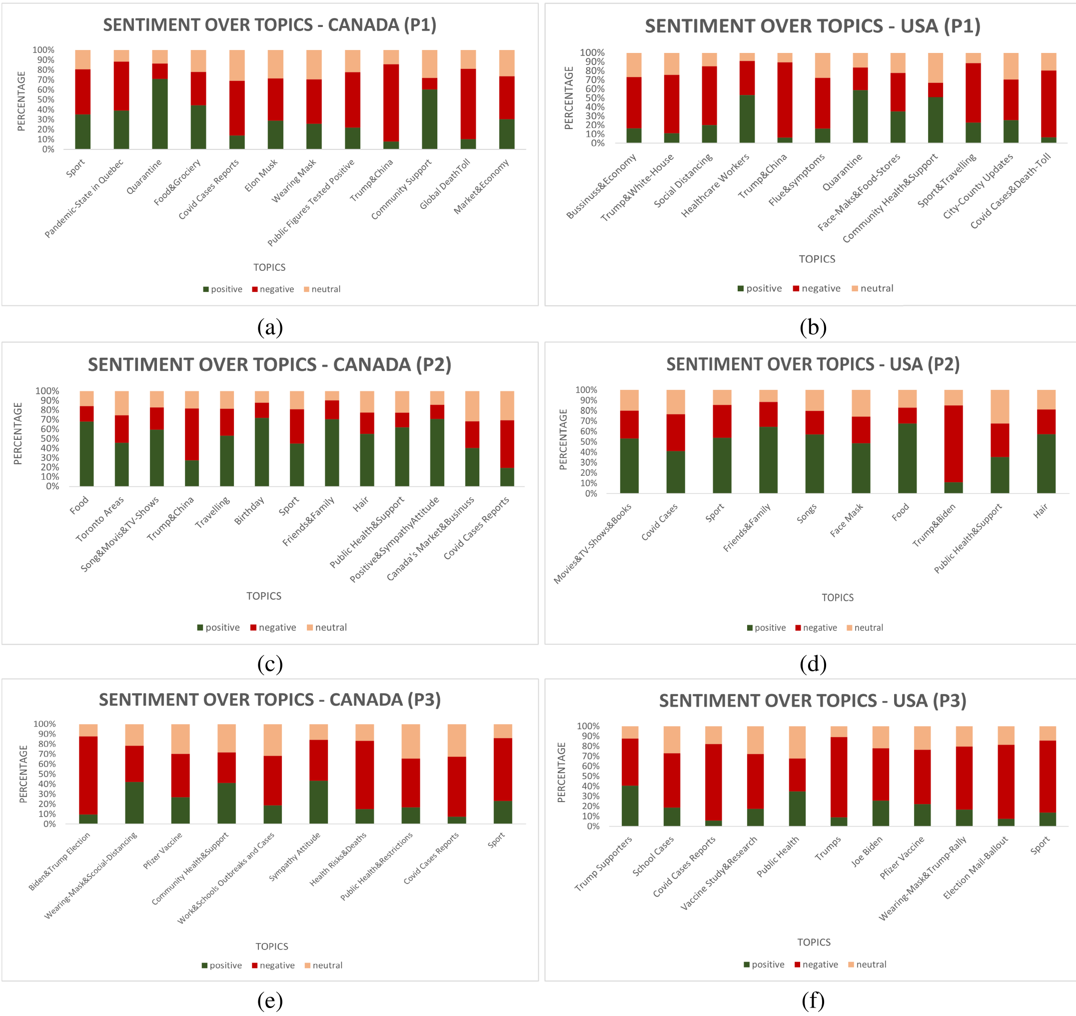


FIGURE 10. **Comparisons of sentiment behavior between Canada and USA over topics inferred during periods 1 & 2 & 3 of COVID-19 pandemic. P1 (Dec 2019 - Apr 2020) is depicted in (a) and (b), P2 (May 2020 - Aug 2020) is dipected in (c) and (d), P3 (Sep 2020 - Nov 2020) is depicted in (e) and (f).**

listen', 'fun food easy food', 'stay safe', etc) that people used in the conversations related to these topics. Community support and Health care is another topic that shows high sentiment positivity. This is not surprising since many of the conversations were related to community and health support including mental health, as shown in the extracted keywords and phrases in Table. 8 Topic 2 - Period 3 and Table. 9 - Topic 1 - Period 2. Health-care and front-line workers have provided and received great support through OSNs in USA. This is depicted in the high positive sentiment seen in Figure. 10b - 'Public-Health and Support' topic. Some of the supportive tweets were praying to workers and some were from families or friends who have members working in hospital. This is confirmed by the keywords and phrases extracted by our topic model and summarizer. "Pray", "happy", "love family", "hospital worker pray" are examples of the positive vibes that people embraced while interacting with this topic. A high level of negative sentiment is shown in the conversations related to Trump and China, for both Canada and USA during periods 1, 2, and 3. Phrases like "trump hoax", "trump blame china", "president trump beat china" reflect a high negative vibes inferred by our sentiment model. Reporting COVID-19 cases and death toll has also shown negative sentiment throughout the topic discussions with USA showing more negative behavior than Canada did. In addition, topics related to wearing masks have shown a degree of negative behavior in Canada and USA. Moreover, discussing the economy and financial situation during the pandemic appears to have a more negative impact on people in USA than those in Canada.

During period 1 (i.e. Dec 2019 - May 2020) in Canada, our topic model inferred the topic "Public Figures Tested Positive" (Figure. 10a) that discussed mainly famous figures that tested positive. UFC fans, apparently, were not happy about the fighting games being cancelled because of the player Jacare being infected with COVID-19. According to the results, Stephen Miller does not seem to have enough fans in Canada; people showed neither sympathy nor support when

TABLE 8. Top keywords and phrases extracted using RAKE based on LDA-TFIDF top keywords. The keywords and phrases are ranked based on RAKE - Canada.

		Keywords	Phrases-L2	Phrases-L3	Phrases-L4	Phrases-Lmax
Period-1	Topic 1	stayhome, quarantine, isolation, staysafe, stayathome, quarantinelife, beautiful, dog, fun, walk, song, anxiety, hair	stayhome quarantine, quarantine stayhome, socialdistance stayhome, stayhome socialdistancing, canada stayhome, stayhome canada, stayhome toronto, home stayhome, toronto stayhome	socialdistance quarantine stayhome, stayhome quarantine pandemic, stayhome staysafe quarantine, stayhome staysafe socialdistancing, staysafe stayhome socialdistancing, toronto stayhome ontario, home staysafe stayhome, stayhome socialdistance flattenthecurve, pandemic stayhome toronto, stayhome stayhome selfquarantine,	quarantine quarantine stayhome staysafe, quarantine social distancing stayhome, pandemic mentalhealth quarantine stayhome, stayhome isolation quarantine ottawa, untiltomorrow stayhome quarantine ontario, pandemic stayhomesavelive stayhome socialdistancing, quarantinelife toronto staysafe stayhome, stayhome staysafe socialdistance besafe, stayathome stayhome socialdistancing washyourhand, distance socialdistance stayhome plankthecurve	canada health government protectyourself stayhome socialdistance life lifelesson staysafe montreal toronto ottawa canada usa nyc california europe greece
	Topic 2	white_house, president, chinese, america, americans, american, fauci, obama, racist, blame, believe, election, bad	president trump, trump americans, trump chinese, trump president, chinese trump, trump america, lie trump, americans trump, america trump, obama trump,	china trump chinese, trump blame china, white house trump, government president trump, president trump chinese, break president trump, updates president trump, america president trump, trump wartime president, stop president trump,	trump americans trump trump, trump president trump shame, trump america save trump, trump trumpvirus trump bad, raise china president trump, trump stop blame china, trump blame china plan, donald trump china blame, buddy trump blame china, president donald trump miracle,	trump americans trump trump
	Topic 3	elon_musk, tesla, cure, contact_trace, true, future, poor, expert, petition, chronic_ink, infectious, publish, superbug	elon musk trump true, true canada, canada expert, expert world, trump cure, tesla cure, official tesla, tesla elonmusk, paraphrased tesla,	tesla elon musk, elon musk tesla, cure cbs news, elon musk threaten, random elon musk, elon musk reopen, elon musk furious, elon musk defy, elon musk lash, reopens tesla factory,	petition government canada canada, elon musk tesla factory, elon musk reopen tesla, cure minutes cbs news, social media bbc news, elon musk threaten move, fight elon musk tweet, elon musk champion defy, trump cure injecting disinfectant, bloomberg elon musk lash,	publish canada news media news
	Topic 4	boris_johnson, ufc, stephen_miller, murder_hornets, fraud, national_exhibition, whistleblower, scam, question, joke, meme, covidiot, answer	ufc jacare, boris johnson , cancel ufc ufc card, card ufc, fbi ufc, ufc cancel, government whistleblower, ufc fighter,	canadians question trump, fight cancel ufc, fightcancelled jacare ufc, murder hornets canada, save boris johnson, boris johnson die, dana white ufc, fight later ufc, boris johnson advice, boris johnson message,	ufc card ufc ufc, boris johnson tests positive, ufc card test positive, stephen miller test positive, contracts ufc ufc release, ufc card jacare souza, semen trump stephen miller, british boris johnson lay, whistleblower warn darkest winter, bank canada ctv news,	prime minister british prime minister boris johnson tests positive
Period-2	Topic 1	case, public, testing, ontario, news, risk, vaccine, hospital, number, high, patient, report, flu, cdc, infection, confirm, symptom	public health, death case, case death, case high, case birx, ontario health, case study, canada case, infection health	hit case case, toronto public health, case flu case, health case recover, case health cbc, additional case health, henry public health, ontario public health, case health unit, total case case	case death case death, prison health public health, case public health unit, positive case positive case, case deaths total death, case death active case, total number case health, case death total case, case report department health	case health authority vancouver coastal health fraser health island health interior health northern health
	Topic 2	hair, black, white, face, beautiful, blue, pretty, body, head, woman, cut, hot, cat, eye, brain, perfect, nose, lady, pain, idk, imagine, point, hurt, ask, haircut, cough	hair love, black hair, care hair, white hair, white love, hair mask, hair face, cut hair, hair blue, hair feeling	black hair salon, black woman hair, white hair head, love short hair, yoongi black hair, love hair colour, black hair dye, naturally black hair, cut hair love, love black woman	white black face mask, pink hair blonde hair, cut hair cute love, curler hair hair thick, black white green red, swing blue radio love, mask white black combo, friend dye hair red, blame love silver hair, personality hair white tooth	black trans woman black woman white gay black woman simple math
	Topic 3	trump, president, china, america, biden, americans, government, stop, dead, american, world, vote, country, response, economy, election, lie, control, kill, blame, protest, hoax	president trump, trump china, trump americans, trump president, trump america, biden trump, trump stop, job trump, trump world	trump anti trump, trump trump imbecile, donald trump trump, trump election trump, understand trump trump, trump trump republicans, trump blame trump, trump hoax trump, trump supporter trump, trump riot trump	trump blame china trump, elect president trump trump, america trump presidency trump, donald trump donald trump, trump administration donald trump, trump bad trump responsible, ivanka trump trump warrior, trump family trump hiding, trump kill nonsense trump, trump divide trump failure	trump killer murderer trump killer murderer trump killer murderer trump killer murderer trump killer murderer trump killer murderer trump killer murderer
Period-3	Topic 1	biden, president_elect, president, vote, americans, mail, america, joe, task_force, china, dead, american, rally, country, lie, believe, hoax, disappear	biden trump, trump biden, president trump, vote trump, trump president, americans trump, trump vote, trump americans, mail trump, american trump	trump rally trump, trump lie trump, trump trump lie, election trump biden, biden trump election, biden president trump, trump president biden, trump biden president, biden vote trump, trump biden vote	trump biden concede trump, trump biden blame trump, vote trump trump negligence, trump gop rally trump, trump rally trump rally, trump biden election election, impressive election trump biden, usa vote trump biden, compare president trump biden, die vote trump biden	president trump black matter trump trump
	Topic 2	support, community, market, economic, online, global, mental, canadian, recovery, service, impact, program, learn, economy, industry, crisis, job, pay, money, challenge, fund, company	mental health, community health, health online, health economic, economic health, online health, global health, health program, health crisis, pandemic support	pandemic mental health, mental health support, health care community, support mental health, health care service, health care job, canadian public health, public health service, work mental health, global health crisis	health care mental health, health issue mental health, financially health mental health, strengthen health care support, health care social service, support child mental health, pandemic affect mental health, economy health care education, relief health care job, support employee mental health	believe science believe health care professional believe alternative care professional believe mental health care expert
	Topic 3	vaccine, pfizer_biontech, pfizer, effective, study, science, mink, early, denmark, trial, research, human, fauci, drug, treatment, antibody, scientist, develop, candidate, cure, announce, mutate, prevent	study vaccine, pfizer vaccine, vaccine fauci, vaccine pfizer, vaccine news, vaccine trial, news vaccine, vaccine effective, vaccine trump, vaccine shot	vaccine quickly vaccine, vaccine secondly vaccine, vaccine pfizer drug, vaccine pfizer pfizer, vaccine news pfizer, pfizer vaccine news, pfizer vaccine effective, vaccine pfizer effective, effective pfizer vaccine, effective trump vaccine	pfizer vaccine mrna vaccine, vaccine effective vaccine bramptonnorth, pfizer early vaccine data, vaccine vaccine merely protection, vaccine candidate effective pfizer, pfizer early data vaccine, amazing news pfizer vaccine, vaccine early trial news, vaccine news cnbc news, pfizer vaccine effective preventing	vaccine early data pfizer vaccine

he tested positive. Hence, we see the negative reflection in our results. This supports the detected hate behavior found in this topic. 68% of the tweets related to Boris Johnson implied negative sentiment. This again shows an unwelcoming attitude among people in Canada. Topics inferred during period 2 (i.e. May 2020 - Aug 2020) have shown dominant positive sentiments in both Canada and USA (Figures. 10c and 10d). However, the number of tweets related to COVID-19 (i.e. mentioning covid related keywords) has decreased during June to Aug and the tweets were mostly discussing politics and publish health restrictions. This can be shown in the high negative behavior during months of June till Aug 2020. On the contrary to period 2, period 3 (i.e. Sep - Nov 2020) has generally shown high negative behavior in comparison to the sentiment during period 2 of the pandemic as seen in Figures. 10e and 10f). Further, topics discussed in USA during this period have mostly shown dominant negative behavior except for "Public Health" and "Trump Supporters" topics showing relatively high positive sentiments. Canada on the other hand, enjoyed more positive vibes, than USA did during period 3, across its topics with three of which being dominated by positive conversations.

The main limitation of this work is the non-dynamic topic modeling that does not analyze the evolution of topics over time. Despite this limitation, our proposed topic models have shown good performance results in discovering patterns and inferring topics from the challenging noisy unstructured-formatted OSNs data. Combining TFIDF with NLP techniques and carefully preparing our data and crafting our features have successfully contributed in building topic models that are capable of handling the OSNs data, as reported in our results and analysis. Another limitation is that this work focus on predicting explicit hate and sentiment contents from OSNs messages, however, it was not designed to detect hidden hate or sentiments contained in sarcastic messages. Despite this fact, emojis and emoticons (he iconic features) were considered as an attempt to assist in recognizing hate and sentiments in this case. An additional limitation of this paper is data imbalance found in our sentiment dataset (i.e. especially for the neutral class). However, the proposed BERT-based sentiment model provide excellent performance results even on neutral class. We believe that the attention and bidirectional mechanisms adopted by BERT algorithm have minimized the bias towards the majority classes and have shown a very acceptable performance on recognizing the minority class (i.e. neutral). This can be shown in the results reported in this paper as well as the large-scale analysis provided in this paper.

IX. CONCLUSION AND FUTURE WORK

This work socializes Internet of Things (IoT) by utilizing social media communications and artificial intelligence (AI),

TABLE 9. Top keywords and phrases extracted using RAKE based on LDA-TFIDF top keywords. The keywords and phrases are ranked based on RAKE - USA.

		Keywords	Phrases-L2	Phrases-L3	Phrases-L4	Phrases-Lmax
Period-1	Topic 1	nyc, quarantine, stayhome, brooklyn, staysafe, quarantinelife, manhattan, stayathome, elon_musk, street, beautiful, song, youtube, murder_hornets, fun, walk, dance, movie, photo, enjoy, wearamask, tesla, laugh	quarantine nyc, nyc quarantine, nyc stayhome, stayhome nyc, stayhome quarantine, quarantine stayhome, socialdistance nyc, nyc socialdistance	quarantine stayhome quarantine, nyc socialdistance stayhome, nyc lockdown nyc, quarantine quarantinelife nyc, quarantinelife quarantine nyc, stayhome quarantinelife nyc, stayhome quarantine pandemic, nyc nyc nyclockdown, quarantine quarantine quarantinelife, nyc lowkey nyc	quarantine stayhome stayhome staysafe, quarantine quarantine stayhome bored, pandemic newyorkcity nyc quarantine, socialdistance nyc stayhome doyourpart, quarantine quarantinelife quarantinelife quarantine, quarantine quarantinelife stayhome staysafe, love socialdistance pandemic nyc, quarantine socialdistance stayhome feetapart, quarantine socialdistance stayhome stircrazy, quarantine stayhome socialdistance wip	quarantine quarantine quarantine quarantine quarantine quarantine quarantine boy game throne
	Topic 2	china, americans, america, chinese, obama, biden, american, hoax, election, blame, lie, dead, vote, believe	trump china, china trump, trump obama, trump america, trump americans, america trump, trump chinese, obama trump, hoax trump, biden trump	president trump china, china president trump, donald trump china, trump china response, ask china trump, trump ask china, blame china trump, trump trust china, president trump chinese, china trump lose	lie trump president trump, trump biden trump biden, trump america maga trump, trump america save trump, trump supporter believe trump, trump praise china china, lab china president trump, china option president trump, president trump beat china, president trump blame china	presidential race head head biden trump battleground trump biden handle economy trump biden handle health care biden trump handling
	Topic 3	white_house, anthony_fauci, task_force, rand_paul, penny, greta_thunberg, fauci, cdc, donald, administration, cnn, staff, senate, official, negative, aide, result, warn, pence	trump fauci, trump pence, pence trump, donald trump, trump administration, negative trump, trump penny, penny trump, fauci trump, administration trump	trump white house, white house trump, president donald trump, president trump pence, trump pence test, trump fauci penny, penny trump fauci, trump fauci warn, trump fauci play, trump fauci warning	white house president trump, idiot trump trump penny, trump trump fauci present, trump pence test positive, trump administration trump gop, trump administration trump create, white house trump remark, trump pence white house, trump aide infect trump, white house doctor trump	vice president mike pence press secretary test positive donald trump personal valet test positive
	Topic 4	season, travel, sport, college, summer, game, cancel, event, play, flight, baseball, player, fan, class, football, mlb, airline, postpone, concert, graduation, fall, refund, wedding, fly, ruin	season sport, sport season, season team, nba season, school sport, travel school, season player	spring sport season, high school sport, trump nfl season, pandemic college sport, pandemic season game, season travel europe, fall sport season, nba season cancel, future travel travel, pandemic sport event	high school sport season, njsiaa spring sport season, cancel spring season sport, cancel spring sport season, spring high school sport, high school spring sport, high school baseball season, high school college student, future high school sport, sport recent major sport	nba season suspendednhl season suspendedmlb season delay daysus government self quarantine affect
Period-2	Topic 1	public, business, support, vaccine, risk, community, high, patient, crisis, study, impact, learn, flu, result	public health, health public, business health, risk health, health risk, community health, health community, health study, health crisis, patient health	pandemic public health, nyc public health, public health big, vaccine public health, support mental health, public health business, public health infrastructure, public health community, work public health, public health emergency	health care public health, impact health health care, community health mobile health, global public health pandemic, pandemic public health crisis, public health pandemic crisis, pandemic public health expert, public health care crisis, support health care worker, pandemic community health center	pay exam pay essay pay stat pay assignment pay paper pay final paper research pay homework pay economic pay finance pay accounting pay economic pay project pay online class pay business
	Topic 2	chicken, food, eat, pizza, dinner, bread, cake, lunch, cooking, meat, taste, cook, fat, ready, finger_lickin, karen, ugh, kfc, dip, thread, agree, boo, pound, yikes, sigh, slogan, nan	love chicken, fried chicken, eat chicken, love food, chicken sandwich, pizza chicken, chicken salad, chicken dinner, homemade chicken	cream cheese chicken, chicken chicken riggie, love cooking chicken, fast food chicken, love chicken beef, chicken cheese steak, eat fry chicken, love blue chicken, butter chicken pizza, fried chicken pizza	chicken grill chicken rice, fry chicken chicken thigh, love fried chicken universal, chicken cheese steak fry, menu chicken cheese steak, chicken bread chicken cutlet, chicken tender fast food, dinner chicken alfredo dinner, chicken cheese steak sell	addict gummy bears tacos raw cookie dough burritos nachos cake pie popsicle fry chicken cheesecake sour patch kid reese peanut butter cup string cheese sour gummy worm oreo cookies pizza fruit snacks cheese wheels rotisserie chickens
	Topic 3	trump, biden, president, americans, america, china, vote, dead, election, country, economy, lie, rnc, protest, kill, believe, hoax, blame	president trump, biden trump, trump biden, trump americans, americans trump, trump china, trump president, trump desperate, trump america	trump president trump, president trump trump, trump america trump, trump china trump, donald trump trump, trump hoax trump, hurt trump trump, trump election trump, trump country trump	blame trump trump trump, trump trump trump insert, biden trump biden trump, president trump president trump, donald trump president trump, donald trump america trump, face president trump trump, biden trump hillary trump, hydroxychloroquine trump trump biden, delusional trump president trump	donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail
Period-3	Topic 1	americans, america, dead, china, american, lie, kill, save, blame	americans trump, trump americans, american trump, trump dead, trump america, dead trump, trump china, trump american, lie trump, america trump	trump trump lie, blame trump trump, trump americans dead, trump america biden, americans dead trump, trump america americans, dead americans trump, biden trump kill, americans infect trump, americans trump responsible	america trump trump incourage, trump lie trump promise, americans dead president trump, donald trump president americans, dead americans child trump, biden blame president trump, trump biden usa america, americans dead elect trump, matter biden trump america, donald trump dead americans	trump hate group stand trump unfit president trump violent action hate group votetrumpout large number trumpisaracist americans dead
	Topic 2	social_distancing, super_spreader, social_distance, wear, rally, event, distance, wearamask, crowd, party, sick, stupid, protect, celebration, idiot, street, inauguration, joke, forget	trump wear, rally trump, trump rally, wear mask, event trump, mask wear, trump event, rally mask, mask rally, sick trump	wear mask trump, trump wear mask, trump rally trump, wear mask mask, trump rally mask, mask wear mask, wear mask wear, social distancing trump, wear mask rally, biden wear mask	trump rally trump rally, trump rally trump supporter, wear mask wear mask, wear mask trump rally, trump lie wear mask, melania trump wear mask, afraid trump wear mask, protect trump wear mask, recklessly mask trump rally, trump supporter wear mask	trump refuse wear mask trump hold indoor rally
	Topic 3	pfizer_biontech, chief_staff, mark_meadow, pfizer, ben_carson, percent_effective, effective, hud_secretary, carson, operation_warp, result, break, trial, early, speed, negative, analysis, candidate, develop, disappear	trump pfizer, pfizer vaccine, vaccine pfizer, pfizer trump, pfizer positive, pfizer early, news pfizer, vaccine trial, early vaccine, president pfizer	pfizer vaccine pfizer, vaccine pfizer early, vaccine news pfizer, vaccine effective trump, vaccine effective pfizer, pfizer data vaccine, pfizer news trump, pfizer vaccine news, positive vaccine result, vaccine pfizer drug	pfizer vaccine effective trump, vaccine trial pfizer vaccine, trump chief staff positive, vaccine pfizer early signal, break trump positive test, vaccine candidate effective pfizer, break news president trump, vaccine ready pfizer effective, vaccine pfizer early peek, trump chief staff meadow	trump administration disaster white house chief staff mark meadows trump aide test positive

and proposes to build a real-time framework for monitoring online social behaviors during the COVID-19 pandemic. Unsupervised and supervised learning approaches are adopted in the design of the framework. Online hate and sentiment behaviors are the two aspects that this paper focuses on and hence it proposes to build two BERT-based classifiers for two supervised sequence classification tasks: binary-class hate model that predicts hate and non-hate contents, multi-class sentiment model that predicts positive, negative, and neutral contents from OSNs data. The results show that BERT-based models yield superior performance in learning all classes for both classification tasks, in comparison to the learning performances of LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM models in both tasks. We have found that BERT-based models are less sensitive to class imbalance (i.e. class imbalance friendly) when compared to LSTM and biLSTM models. LSTM and biLSTMS sensitivity to class imbalance has been shown to introduce a degree of confusion and bias towards majority classes as reported in our results. However, our findings have revealed that combining CNN with LSTM especially with biLSTM could be used to relax the sensitivity towards class imbalance in text sequence classification. The understanding of online hate and sentiment behaviors are facilitated through our proposed unsupervised framework for data exploration and interpretation. We propose using topic modeling and phrase extraction methods for discovering hidden patterns and inferring topics, trends, and concerns formed during the pandemic as well as automatically providing coherent interpretation of the inferred topics without human effort involved. The results show that TFIDF-LDA topic model produces more semantically coherence performance across topics than BOW-NMF topic model does, in terms of coherence score. This finding sheds light on the effectiveness and capability of TFIDF technique in signifying the importance of influencing words with valuable information from OSNs data that come with large volumes of noisy and limited-content messages. However, TFIDF shows poor performance in keyword and phrase extraction when data volumes are too small. RAKE algorithm, on the other hand, has been proven fast and effective in phrase extraction from small-sized OSNs datasets for the purpose of automatically and coherently interpreting topics inferred from our TFIDF-LDA topic model. Throughout our large-scale temporal and topic-based analysis during COVID-19 pandemic, we observe that there is a correlation between sentiment and hate behavior; the presence of high hate behavior indicates a presence of negative behavior as seen in Trump and China, and wearing mask topics. The opposite is not true; the presence of high negative behavior does not guarantee the presence of hate behavior. This is seen in topics related to COVID-19 cases and death tolls. In addition, our analysis shows more sentimental positivity and less hate behavior in Canada compared to sentiment positivity and hate behavior in USA, and that Canadian tone in reporting news is less sharp than the corresponding tone of USA's news.

For future directions, dynamic topic modeling will be considered in order to track changes in pattern of topics over time. This will assist in a deeper understating of trends and concerns through their evolution over time in order to improve the understanding of corresponding online social behavior of users. Also, we are interested in recognizing the implicit online social behavior that hide behind sarcastic expressions. We are also interested in incorporating the multi-lingual aspect to the online social behavior modeling. In addition, current research and efforts in LDA-style topics interpretation and labeling are very limited and hence more attention should be given to improve the quality of topics comprehensibility. we encourage research community to address this problem and extend the efforts in this area.

REFERENCES

[1] M. S. Hossain, G. Muhammad, and N. Guizani, "Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics," *IEEE Netw.*, vol. 34, no. 4, pp. 126–132, Jul. 2020.

[2] M. Tavakoli, J. Carriere, and A. Torabi, "Robotics, smart wearable technologies, and autonomous intelligent systems for healthcare during the COVID-19 pandemic: An analysis of the state of the art and future vision," *Adv. Intell. Syst.*, vol. 2, no. 7, Jul. 2020, Art. no. 2000071.

[3] S. Singh, D. Roy, K. Sinha, S. Parveen, G. Sharma, and G. Joshi, "Impact of COVID-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations," *Psychiatry Res.*, vol. 293, Nov. 2020, Art. no. 113429.

[4] Z. Long, R. Alharthi, and A. E. Saddik, "NeedFull—A tweet analysis platform to study human needs during the COVID-19 pandemic in New York state," *IEEE Access*, vol. 8, pp. 136046–136055, 2020.

[5] F. Alzamzami, M. Hoda, and A. E. Saddik, "Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020.

[6] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words—A feature-based approach," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1046–1056.

[7] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 sensing: Negative sentiment analysis on social media in China via BERT model," *IEEE Access*, vol. 8, pp. 138162–138169, 2020.

[8] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074–181090, 2020.

[9] R. Abaalkhail, F. Alzamzami, S. Aloufi, R. Alharthi, and A. El Saddik, "Affectional ontology and multimedia dataset for sentiment analysis," in *Proc. Int. Conf. Smart Multimedia*. Cham, Switzerland: Springer, 2018, pp. 15–28.

[10] R. Cohen-Almagor, "Fighting hate and bigotry on the Internet," *Policy Internet*, vol. 3, no. 3, pp. 1–26, 2011.

[11] (1997). *Council of Europe Committee of Ministers*. Accessed: Dec. 1, 2020. [Online]. Available: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b

[12] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2019, pp. 363–370.

[13] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 490–499.

[14] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media," 2020, *arXiv:2007.02304*. [Online]. Available: http://arxiv.org/abs/2007.02304

[15] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, "Detection of hate speech in COVID-19–related tweets in the Arab region: Deep learning and topic modeling approach," *J. Med. Internet Res.*, vol. 22, no. 12, 2020, Art. no. e22609. [Online]. Available: https://www.jmir.org/2020/12/e22609, doi: 10.2196/22609.

[16] X.-L. Mao, Y.-J. Zhao, Q. Zhou, W.-Q. Yuan, L. Yang, and H.-Y. Huang, "A novel fast framework for topic labeling based on similarity-preserved hashing," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 3339–3348.

[17] M. K. Saini, F. Al-Zamzami, and A. E. Saddik, "Towards storytelling by extracting social information from OSN photo's metadata," in *Proc. 1st Int. Workshop Internet-Scale Multimedia Manage. (WISMM)*, 2014, pp. 15–20.

[18] F. Alzamzami, M. Saini, and A. El Saddik, "DST: Days spent together using soft sensory information on OSNs—A case study on Facebook," *Soft Comput.*, vol. 21, no. 15, pp. 4227–4238, 2017.

[19] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Syst.*, vol. 25, no. 5, pp. 565–575, Oct. 2019.

[20] M. S. Hossain, "Cloud-supported cyber–physical localization framework for patients monitoring," *IEEE Syst. J.*, vol. 11, no. 1, pp. 118–127, Mar. 2017.

[21] G. Barkur, Vibha, and G. B. Kamath, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India," *Asian J. Psychiatry*, vol. 51, Jun. 2020, Art. no. 102089.

[22] J. Uyheng and K. M. Carley, "Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines," *J. Comput. Social Sci.*, vol. 3, pp. 1–24, Oct. 2020.

[23] D. Li, H. Chaudhary, and Z. Zhang, "Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining," *Int. J. Environ. Res. Public Health*, vol. 17, no. 14, p. 4988, Jul. 2020.

[24] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of COVID-19 tweets using topic modeling, UMAP, and DiGraphs," 2020, *arXiv:2005.03082*. [Online]. Available: http://arxiv.org/abs/2005.03082

[25] P. Suri and N. R. Roy, "Comparison between LDA & NMF for event-detection from large text stream data," in *Proc. 3rd Int. Conf. Comput. Intell. Commun. Technol. (CICT)*, Feb. 2017, pp. 1–5.

[26] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *Proc. World Wide Web Conf.*, 2018, pp. 1105–1114.

[27] A. Schofield, M. Magnusson, and D. Mimno, "Pulling out the stops: Rethinking stopword removal for topic models," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 432–436.

[28] A. Fan, F. Doshi-Velez, and L. Miratrix, "Prior matters: Simple and general methods for evaluating and improving topic quality in topic modeling," 2017, *arXiv:1701.03227*. [Online]. Available: http://arxiv.org/abs/1701.03227

[29] F. Martin and M. Johnson, "More efficient topic modelling through a noun only approach," in *Proc. Australas. Lang. Technol. Assoc. Workshop*, 2015, pp. 111–115.

[30] W. M. Darling, M. Paul, and F. Song, "Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic Bayesian HMM," in *Proc. Workshop Semantic Anal. Social Media*, 2012, pp. 1–9.

[31] B. Jeong, J. Yoon, and J.-M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," *Int. J. Inf. Manage.*, vol. 48, pp. 280–290, Oct. 2019.

[32] A. Gourru, J. Velcin, M. Roche, C. Gravier, and P. Poncelet, "United we stand: Using multiple strategies for topic labeling," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2018, pp. 352–363.

[33] A. E. C. Basave, Y. He, and R. Xu, "Automatic labelling of topic models learned from Twitter by summarisation," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 618–624.

[34] D. He, M. Wang, A. M. Khattak, L. Zhang, and W. Gao, "Automatic labeling of topic models using graph-based ranking," *IEEE Access*, vol. 7, pp. 131593–131608, 2019.

[35] J. Fenglei and G. Cuiyun, "An online topic modeling framework with topics automatically labeled," in *Proc. Workshop Widening NLP*, 2019, pp. 73–76.

[36] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware BERT for cross-domain sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4019–4028.

[37] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification," 2019, *arXiv:1908.11860*. [Online]. Available: http://arxiv.org/abs/1908.11860

[38] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using BERT," in *Proc. Artif. Intell. Transforming Bus. Soc. (AITB)*, vol. 1, Nov. 2019, pp. 1–5.

[39] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*. [Online]. Available: http://arxiv.org/abs/1903.09588

[40] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.

[41] N. Azzouza, K. Akli-Astouati, and R. Ibrahim, "TwitterBERT: Framework for Twitter sentiment analysis based on pre-trained language model representations," in *Proc. Int. Conf. Reliable Inf. Commun. Technol.* Cham, Switzerland: Springer, 2019, pp. 428–437.

[42] C. Dyer, A. Cordova, A. Mont, and J. Lin, "Fast, easy, and cheap: Construction of statistical machine translation models with MapReduce," in *Proc. 3rd Workshop Stat. Mach. Transl.*, 2008, pp. 199–207.

[43] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.

[44] B. Myagmar, J. Li, and S. Kimura, "Transferable high-level representations of BERT for cross-domain sentiment classification," in *Proc. Int. Conf. Artif. Intell. (ICAI)*, 2019, pp. 135–141.

[45] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Proc. Int. Conf. Complex Netw. Their Appl.* Cham, Switzerland: Springer, 2019, pp. 928–940.

[46] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–22, May 2020.

[47] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0237861.

[48] A. Nikolov and V. Radivchev, "Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 691–695.

[49] N. S. Samghabadi, P. Patwa, P. Srinivas, P. Mukherjee, A. Das, and T. Solorio, "Aggression and misogyny detection using BERT: A multi-task approach," in *Proc. 2nd Workshop Trolling, Aggression Cyberbullying*, 2020, pp. 126–131.

[50] H. S. Alatawi, A. M. Alhothali, and K. M. Moria, "Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT," 2020, *arXiv:2010.00357*. [Online]. Available: http://arxiv.org/abs/2010.00357

[51] Z. Waseem, J. Thorne, and J. Bingel, "Bridging the gaps: Multi task learning for domain transfer of hate speech detection," in *Online Harassment*. Cham, Switzerland: Springer, 2018, pp. 29–55.

[52] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 1354–1364.

[53] D. Nolasco and J. Oliveira, "Subevents detection through topic modeling in social media posts," *Future Gener. Comput. Syst.*, vol. 93, pp. 290–303, Apr. 2019.

[54] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[55] 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.

[56] R. Zhao and V. Y. F. Tan, "Online nonnegative matrix factorization with outliers," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 555–570, Feb. 2017.

[57] K. S. Hasan and V. Ng, "Automatic keyphrase extraction: A survey of the state of the art," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 1262–1273.

[58] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 2, p. e1339, Mar. 2020.

[59] S. Siddiqi and A. Sharan, "Keyword and keyphrase extraction techniques: A literature review," *Int. J. Comput. Appl.*, vol. 109, no. 2, pp. 18–23, Jan. 2015.

[60] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.

[61] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge," in *Proc. AAAI*, vol. 8, 2008, pp. 855–860.

[62] S. Danesh, T. Sumner, and J. H. Martin, "SGRank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction," in *Proc. 4th Joint Conf. Lexical Comput. Semantics*, 2015, pp. 117–126.

[63] C. Florescu and C. Caragea, "PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1105–1115.

[64] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining, Appl. Theory*, vol. 1, pp. 1–20, Mar. 2010.

[65] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: An overview," *Sci. China Inf. Sci.*, vol. 63, no. 1, pp. 1–36, Jan. 2020.

[66] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461.

[67] B. Huang and K. M. Carley, "Parameterized convolutional neural networks for aspect level sentiment classification," 2019, *arXiv:1909.06276*. [Online]. Available: http://arxiv.org/abs/1909.06276

[68] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Gener. Comput. Syst.*, vol. 101, pp. 542–554, Dec. 2019.

[69] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[70] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[72] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[74] A. Fan, F. Doshi-Velez, and L. Miratrix, "Assessing topic model relevance: Evaluation and informative priors," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 12, no. 3, pp. 210–222, Jun. 2019.

[75] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 530–539.

[76] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 54–63.

[77] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, ''SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval),'' 2019, *arXiv:1903.08983*. [Online]. Available: http://arxiv.org/abs/1903.08983

[78] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, ''Large scale crowdsourcing and characterization of Twitter abusive behavior,'' 2018, *arXiv:1802.00393*. [Online]. Available: http://arxiv.org/abs/1802.00393

[79] Z. Waseem and D. Hovy, ''Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,'' in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.

[80] M.-T. Nguyen, D. V. Lai, H. T. Nguyen, and M. Le Nguyen, ''TSix: A human-involved-creation dataset for tweet summarization,'' in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–5.

[81] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, ''Audio–visual emotion-aware cloud gaming framework,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.

[82] S. Syed and M. Spruit, ''Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation,'' in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2017, pp. 165–174.

[83] C.-Y. Lin, ''Rouge: A package for automatic evaluation of summaries,'' in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[84] D. M. Blei and J. D. Lafferty, ''Topic models,'' *Text Mining, Classification, Clustering, Appl.*, vol. 10, no. 71, p. 34, 2009.

[85] Y. Chen, J.-B. Bordes, and D. Filliat, ''An experimental comparison between NMF and LDA for active cross-situational object-word learning,'' in *Proc. Joint IEEE Int. Conf. Develop. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Sep. 2016, pp. 217–222.

FATIMAH ALZAMZAMI received the M.Sc. degree in computer science from the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada, where she is currently pursuing the Ph.D. degree in computer science. She is supervised by Prof. Abdulmotaled El Saddik. Her research interests include machine learning, deep learning, big data, social multimedia analysis, and mining.

ABDULMOTALEB EL SADDIK (Fellow, IEEE) is currently a Distinguished University Professor and the University Research Chair of the School of Electrical Engineering and Computer Science, University of Ottawa. He is a fellow of the Engineering Institute of Canada and the Canadian Academy of Engineers. He has supervised more than 120 researchers. He has received research grants and contracts totaling more than \$18 M. He has authored or coauthored ten books and over 550 publications. His research focus is on the establishment of digital twins using AI, AR/VR, and tactile internet that allows people to interact in real-time with one another and with their digital representation. He has received several international awards, including the IEEE I&M Technical Achievement Award, the IEEE Canada C. C. Gotlieb (Computer) Medal, and the A. G. L. McNaughton Gold Medal for important contributions in the field of computer engineering and science. He has chaired more than 50 conferences and workshops. He is an ACM Distinguished Scientist.