IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 August 2022, accepted 10 September 2022, date of publication 26 September 2022,
date of current version 30 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3208470

APPLIED RESEARCH

Predicting Depression in Canada by Automatic Filling of Beck's Depression Inventory Questionnaire

RUBA S. SKAIK[1] **AND DIANA INKPEN**[2]
[1]Quality Assurance, Environment and Climate Change Canada, Gatineau, QC J8Y 3Z5, Canada
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Diana Inkpen (diana.inkpen@uottawa.ca)

This work was supported in part by the Natural Science and Engineering Research Council of Canada (NSERC).

ABSTRACT The risk for depression and anxiety increased as people adjusted to a new normal after the COVID-19 pandemic. Early detection and appropriate onset treatment and support can reduce the consequences of depression. Automatic detection of depression in social media has recently become an important area of investigation. However, because of the lack of extensive annotated data, we propose a method for using a model that learns to answer a depression questionnaire and apply it to make population-level predictions. We used the eRisk 2021 Task 3 training dataset to build an automated model to fill the Beck's Depression Inventory (BDI) questionnaire. We selected the best performing model for each group of questions based on predefined metrics and consolidated those models into one model (called the *BDI_Multi_Model*). The *BDI_Multi_Model* achieved better performance than the state-of-the-art for this challenging task. Then, we used this model for inference on a Canadian population dataset and compared its predictions with the statistics of the most recent mental health survey conducted by Statistics Canada. The correlation between the inference of the answered questionnaire based on our *BDI_Multi_Model* and the official statistics showed a strong Pearson correlation of 0.90.

INDEX TERMS Deep learning, beck depression inventory, text classification, mental health, depression detection.

I. INTRODUCTION

Depression is a severe public health issue and one of the world's most recognized mental disorders, with an estimated 3.8% of the population worldwide being impacted.[1] It causes numerous disability-adjusted years worldwide for the workforce. Depression in Canada has received national attention, in general and also during the COVID-19 pandemic. Recognizing persons suffering from depression and assisting those in need is a critical step toward building a better living environment. However, the process of identifying those who have a mental illness is a difficult task.

Various psychiatric scales are used to assess individuals' mental health. For example, researchers may use PHQ-9 (Patient Health Questionnaire) to quantify depressed symptoms. It is a commonly used tool for diagnosing and measuring the severity of depression, and it assesses behavioral characteristics, self-harm, and suicidal thoughts. Another option is the Beck Depression Inventory (BDI) questionnaire, developed by [1]. It is one of the most commonly used tools for estimating the severity of depression. It is a self-reported inventory of 21 questions with multiple choices that are grounded in the patient's thoughts rather than psychodynamic perspectives. Even though these measures are well-established psychiatric instruments, choosing the scale that will be most accurate for a given demographic sample and using it is a complicated task.

Meanwhile, the use of social media has significantly increased over the past few years. As a result, it has attracted many researchers to analyze its contents in different fields and attempt to predict mental health problems within a particular

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

[1]https://www.who.int/news-room/fact-sheets/detail/depression

population. In this article, we aim to build a risk-mitigation tool that allows professionals and decision makers such as the Public Health Agency of Canada (PHAC) to detect the level of depression among social media users without having to ask them to spend time manually filling out a depression questionnaire. Governments can target each group with appropriate monitoring programs, plan necessary medical assistance to the concerned parties in the early stages, and allocate the necessary resources to reduce the burden of mental illness in their region by identifying suffering clusters in demographic information.

Our methodology is summarized in Figure 1. Using the dataset of Task 3 of the eRisk 2021 shared task at CLEF 2021, which contains the user's history of posts on Reddit, together with the answers of these users to the BDI questionnaire, we trained multiple machine learning models that resulted in a multi-model that can answer the questions with higher accuracy than the state-of-the-art models. This model can now be applied to new users who do not provide answers to the questions. To evaluate the performance of our multi-model approach at the population level, we applied it to a dataset of social media users representative of the Canadian population (tweets). We will show that our multi-model approach can detect depression indicators in tweets.

Finally, we use Pearson correlations scores to study associations between age and sex population statistics and our multi-model inference scores.

We used deep learning techniques based on selective (filtered) posts to classify the data extracted from the Reddit postings. Next, we utilized pre-trained models to classify the related posts for each category and assign a category-based level. Then we fine-tuned the deep learning models using different parameters based on the classifier's topic, which enhanced the general depression prediction score from a previous maximum score of 83.59% to 84.38% and improved the depression category score from 41.25% to 48.75% (see section IV for details about the evaluation measures).

The main contributions of this article are as follows:

1) Further test the possibility of using different social media platforms to train multiple machine learning models based on linguistic features to generalize on a population level.
2) Enhance the accuracy of the automated BDI answering system using a multi-model category-based architecture.
3) Apply the model on a population level to automatically answer the BDI-Questionnaire and estimate the depression level accordingly.

The rest of this article is organized as follows: Section II summarizes relevant works of machine learning research in predicting depression with a special focus on population-level analysis. In addition, the related auto-fill systems in the eRisk shared task are reviewed. Section III introduces the datasets used in this research. In Section V, our *BDI_Multi_Model* methodology is explained in detail, and the user-level results for the eRisk dataset are presented. In Section VI we present our population-level experiments and discuss the results. Finally, conclusions and future work are discussed in Section VII.

II. RELATED WORK

Depression has gained significant interest from researchers due to its effect on human beings and society. Current research shows essential associations between an individual's mental health and the linguistic content they share on social media. Recent advances in applying Natural Language Processing and other Machine Learning techniques to social media to address mental health are found in the following surveys [2], [3], [4], [5], [6], [7].

Analysis of social media for predicting mental disorders can be done on a post-level basis using explicit or implicit attributes of the post [8], [9], [10], at the user-level by aggregating multiple posts as a single document or analyzing behavioral changes over time [9], [11], [12], [13], [14], [15], [16], or finally at population-level. For example, [17] developed a probabilistic model to detect the behavioral changes associated with the onset of depression. Whereas, [18] achieved 0.88 AUC score by training a random forest model using an estimated weight of psychological factors such as stress, depression, anxiety, hopelessness, loneliness, burdensomeness, insomnia, and sentiment polarity to predict suicide ideation within the tweets.

On the population level, [19] represented the US counties as graph interactions between Linguistic Inquiry, and Word Count (LIWC) features, then trained several graphs neural networks: graph convolutional network, graph attention network, a hybrid network, and graph isomorphism network to learn the population health representation, and finally, used logistic regression (LR) to estimate the health indices of 3,221 counties. A significant correlation was observed with six health measures, and models with linguistically analyzed Twitter data improved predictive accuracy for 20 community health measures. Note that this work is not about mental health.

Based on data from Reddit users who changed from a mental health condition to suicide ideation, [20] developed a statistical approach relying on three cognitive psychological integrative theories of suicide, including thinking, ambivalence, and decision making; they identified markers to detect this changeover episode.

From the Chinese Longitudinal Healthy Longevity Study (CLHLS) survey, [21] chose 1,538 senior persons. Six machine learning models, including deep neural network (DNN), gradient boosted decision tree (GBDT), SVM, and LR with lasso regularisation were used along with multivariate long short-term memory (LSTM). Different depression risk indicators and the risk of depression in the older population have been studied using this LSTM.

The most related to our research is the eRisk 2021 task 3, which is concerned with measuring the severity of depression signs. The dataset is explained in detail in section III. The task is a continuation of Task 3 at eRisk 2019 and Task 2 at eRisk 2020, and its objective is to automatically estimate

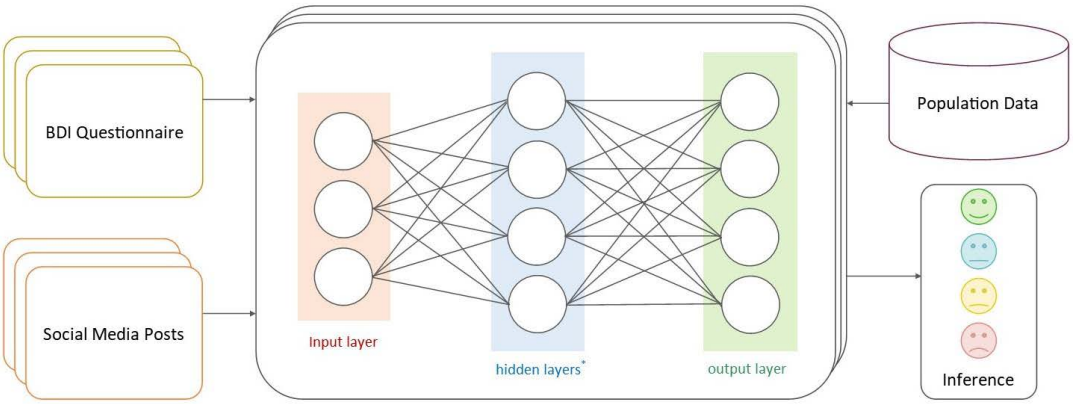

FIGURE 1. Methodology overview.

a user's degree of depression based on their social media postings. It was difficult to achieve good performance on the task metrics in the three editions of the shared task. Most of the submitted runs were barely above the baseline of choosing the most frequent answer to each question. The best performing run [22] in 2021 achieved 73.17% for the depression score, exceeding the values from the previous two years. However, only 41.25% of the social media users were correctly classified into the right level of depression. These findings support the task's potential to extract specific depression-related data from social media behavior automatically. However, there is still room for improvement in the generalization process to advance toward a more comprehensive and adequate depression screening tool. The multi-model we will present in section V is able to achieve higher scores, advancing state of the art.

III. DATASETS

The datasets used in this research are listed in Table (1). We utilize eRisk Dataset (R1) explained in section (III-A) to train a machine learning model to answer the BDI depression questionnaire automatically. Then, the model is used to automatically answer the BDI questionnaire for the users from the P1 dataset, in order to estimate the depression for a representative Canadian population sample. Finally, to compare our model's predictions with official statics, we use the 2015-2016 population-based Canadian Community Health Survey (CCHS) on Mental Health and Well-being conducted by Statistics Canada (P_CCHS). P_CCHS estimates the depression in the Canadian provinces and territories during the year under study; it was a telephone-based survey.

TABLE 1. Summary of the datasets used in this study.

Topic	Dataset	Platform	No. of Users	No. of Posts
eRisk	R1	Reddit	170	63,317
Population	P1	Twitter	15,982	2,582,912
Population	P_CCHS	Statistics Canada	52,996	-

A. eRisk DATASET

eRisk is an initiative to explore issues of evaluation methodologies, performance metrics, and other aspects related to building test collections and defining challenges for early risk detection related to health and safety.[2] The dataset used in this research is based on the eRisk 2021 Task 3 (Measuring the severity of the signs of depression). The task is a continuation of Task 3 at eRisk 2019 and Task 2 at eRisk 2020, and it is the last in this series. Its objective is to automatically estimate a user's degree of depression by building machine learning models to answer a standard depression questionnaire (BDI) using the users' social media postings.

The dataset includes 170 social media users who have filled the BDI questionnaire and voluntarily provided the reference to their Reddit forum posts, the history of their writings was extracted right after the user filled the questionnaire.

The questionnaire contains 21 questions (see Appendix VII for the complete BDI questionnaire [1]) that assess the existence of depression signs such as sadness, pessimism, fatigue, and so on. Each question has four possible responses (0, 1, 2, 3) except for question 16 (about sleep patterns) and question 18 (about appetite) that have seven possible answers namely: 0, 1a, 1b, 2a, 2b, 3a, and 3b. The training set is composed of 43, 514 Reddit posts and comments authored by 90 users who answered the BDI questionnaire throughout the last two years. The test dataset contains 19, 803 posts and comments submitted by 80 people. More information regarding the dataset can be found at [23].

Figure 3 displays the distribution of the answers for BDI questions in the training data. When making these counts, branches (a) and (b) for questions 16 and 18 were grouped together because they all contribute the same number of points when determining a user's depression level.

The following preprocessing rules are applied to R1 Reddit posts:

- Concatenate the title of the post and the post's text.
- Expand contractions.

[2]https://erisk.irlab.org/

FIGURE 2. Class distribution for the users' answers in the training data.

- Remove the words/sentences between square brackets.
- Clean punctuation, special characters and extra white spaces.
- Convert text to lowercase.
- Remove administrative posts, for example: ''Your post was removed for breaking [**rule ....]''.
- All posts with less than four characters are ignored.

B. POPULATION-LEVEL TWITTER DATASET

We used a population dataset from Advanced Symbolics Inc.[3] (ASI), a market research company in Canada. ASI is continuously collecting tweets posted by Twitter users using Conditional Independence Coupler (CIC) sampling algorithm that is based on Coupling from the Past (CFTP) [24]. The stopping condition is enhanced by measuring the distance between the new node and the seed node, then adjusting the weights of sampling using post-stratification to compensate for the underrepresented groups of the population. The algorithm's ability to produce a representative population sample has been mathematically proved, and the author confirmed the sample's representativeness by comparing 3,032 Toronto Twitter user profiles with census data from the same year. The author found that the demographics of Twitter users closely matched those of the 2011 census [25]. By 2018, they have collected millions of tweets from 278,627 users. Thus, P1 dataset (a subset of ASI data) is statistically representative data of Canada's population for 2015 tweets.

The P1 dataset contains spatial, demographic, and textural information as follows:

1) SPATIAL INFORMATION

The location of Twitter users can be inferred either by using the GPS coordinates of the tweets or by using Microsoft's Bing Maps to determine the coordinates of the self-declared location.

2) DEMOGRAPHIC INFORMATION

ASI estimates the age and sex probability distribution by analyzing the Twitter profile photo using Face++.[4] Then the probability distribution is adjusted by comparing the first name with Canada's birth records and the life tables[5] that contain life expectancy and associated age and sex projections for Canada [26], [27].

The age and sex probability distribution is deduced for each user in 12 fields as follows:

$$\mathbb{A} = \{A_i : 1 \leq i \leq 1\}$$
$$\mathbb{S} = \{``M'', ``F''\}$$
$$Sex_{age} = \{P(s_a) : \forall s \in \mathbb{S} \wedge a \in \mathbb{A}\}$$

where

$$\sum_{(\forall s \in \mathbb{S})} P(s) = 1 \text{ and } \sum_{(\forall a \in \mathbb{A})} P(a) = 1$$
$$A_1 : `` \leq 25'', A_2 : ``25 - 34''$$
$$A_3 : ``35 - 44'', A_4 : ``45 - 54''$$
$$A_5 : ``55 - 64'', A_6 : `` > 65''$$

The differences between the probabilities of each category vary. Thus, we decided to keep users with high confidence for both *Age* and *Sex* prediction based on the following rules:

- Sex: We assign to the user the sex of the maximum sex probability of all age groups with a probability more than 92.5%, using the following equation:

$$P(\mathcal{S}) = \{\max(P_{Male}, P_{Female}) : |P_{Male} - P_{Female}| > \delta; \delta = 0.85\}$$

- Age: We assign to the user the age group of the maximum age group probability ($P(\alpha)$), given that the difference between the largest and the second largest is greater than ϵ, where $\epsilon = 2 * P(\alpha_j)$, using the following steps:

$$P(\alpha) \leftarrow \max\{\sum_{(\forall \zeta \in \mathbb{S})} P(\alpha) : \alpha \in \mathbb{A}\}$$
$$\mathbb{B} \leftarrow \mathbb{A} - Age(P(\alpha))$$
$$P(\beta) \leftarrow \max\{\sum_{(\forall \zeta \in \mathbb{S})} P(\beta) : \beta \in \mathbb{B}\}$$
$$\text{and} \quad |P(\alpha) - P(\beta)| > \epsilon$$

3) TEXTUAL INFORMATION

A tweet is a short status update posted by the user with a limit of 140 characters, which doubled to 280 in 2017.

The P1 dataset is a subset of the ASI dataset with the following conditions:

- Each user must have a location mapped to a Canadian province/territory.
- Each user must have age or sex prediction with the minimum defined confidence.
- Each user must have a minimum of 5 posts.
- Each post must be at least 32 characters in length.
- The posts need to have timestamps during 2015.

After applying the above conditions, the number of posts decreased from 9,304,441 to 2,582,912 tweets and the number of users from 278,627 to 15,982 users.

[3] https://advancedsymbolics.com/

[4] A deep learning system developed by Megvii Technology to obtain face attributes.

[5] https://www150.statcan.gc.ca/n1/en/catalogue/84-537-X

C. CCHS-MH SURVEY

The P_CCHS data source is the output of the 2015-2016 Canadian Community Health Survey (CCHS). CCHS is a cross-sectional survey that provides information on health at federal and provincial levels conducted by Statistics Canada. Details about the survey methodology are described in [28]. The depression module was optional in 2015; therefore, only seven provinces and one territory participated *(the number of respondents was 52,996)*, from which 28,738 were female, and the remaining 24,258 were male. Depressive symptoms were assessed using the Patient Health Questionnaire-9 (PHQ-9) to assess the severity of depression symptoms during the previous two weeks. The PHQ-9 is a commonly used screening tool, with a score of 10 or higher suggesting more severe depression symptoms. Considering the depression scale for PHQ-9 $>$ 9, the 12-month prevalence rate for depression in Canada for 2015 was estimated to be 7.1% distributed as illustrated in Table (2).

TABLE 2. Depression prevalence among males and females in Canada's provinces based on CCHS 2015-2016 statistics MN: Manitoba, NB: New Brunswick, NF: Newfoundland and Labrador, NS: Nova Scotia, NT: Northwest Territories, ON: Ontario, PE: Prince Edward Island, SK: Saskatchewan.

Province	Census			P_CCHS Dataset		
	Female	Male	Both	Female	Male	Both
MN	2,708	2,327	5,035	5.11	842	676
NB	1,746	1,354	3,100	3.29	543	393
NL	1,660	1,396	3,056	3.13	516	405
NS	2,473	2,031	4,504	4.67	769	590
NT	468	477	945	0.88	146	138
ON	16,370	13,933	30,303	30.89	5092	4045
PE	1,013	720	1,733	1.91	315	209
SK	2,300	2,020	4,320	4.34	715	586
Canada	28738	24258	52996	54.23	8940	7042

IV. EVALUATION METRICS

We used four evaluation metrics on the R1 dataset level: the ones from the shared task explained in the task overview paper [29]. On the P1 and P_CCHS datasets, we used Pearson correlation as the evaluation measure. The metrics on the R1 dataset were: Average Hit Rate (AHR), the Average Closeness. Rate (ACR), the Depression Category Hit Rate (DCHR), and the Average Difference between Overall Depression Levels (ADODL). We briefly explained them as follows:

AVERAGE HIT RATE (AHR)

AHR is a strict metric that computes the proportion of occurrences in which the automatically completed questionnaire has the exact same response as the filled questionnaire. The Hit Rate is a stringent measure that computes the ratio of cases where the automatically answered questionnaire has the same answers as the actual users' answers to the same question.

THE AVERAGE CLOSENESS RATE (ACR)

ACR is the closeness rate averaged over all users. Taking in the account that the multi-choices implemented in the questionnaire represent the well-established depression categories in psychology, the Closeness Rate CR computes the standard deviation called absolute difference (AD) between the real and the automated answers. The absolute difference is transformed into an effectiveness score as (IV), where MAD is the maximum absolute difference, which is equal to the number of possible answers minus one:

$$CR = \frac{(MAD - AD)}{MAD} \tag{1}$$

THE DEPRESSION CATEGORY HIT RATE (DCHR)

Measures the correctness of the estimation achieved over all users according to the well-established depression categories in psychology. The previous measures assess the systems' ability to answer each question in the form. DCHR instead does not look at question-level hits or differences but computes the overall depression results. The DCHR consists of computing the fraction of cases where the automated questionnaire led to a depression category that is equivalent to the depression category obtained from the real questionnaire level with well-established four categories of depression, as shown in Table 3.

TABLE 3. Depression severity scale.

Depression degree	Scale
Minimal depression	0 - 9
Mild depression	10 - 18
Moderate depression	19 - 29
Severe depression	30 - 63

The depression class distribution in the training dataset is displayed in Figure 3.

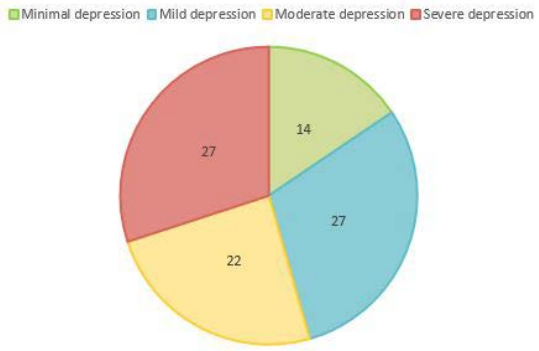

FIGURE 3. Class distribution of depression level based on BDI questionnaire in the training data.

AVERAGE DODL (ADODL)

Measures the difference between overall depression levels (DODL) averaged over all users. Like the DCHR, the DODL computes the overall depression level (sum of all the answers) for the actual and automated answers then the absolute difference (AD overall) between the actual and the automated score is computed. Depression levels are integers

between 0 and 63 and, thus, DODL is normalized into [0, 1] as in (2):

$$\begin{aligned} ADODL &= \frac{(63 - AD\_Overall)}{63 * \upsilon} \\ AD\_Overall &= \sum_{i=1}^{\upsilon} |AS[i] - US[i]| \\ where\ AS &= actual\_score, US = automated\_score \end{aligned} \tag{2}$$

PEARSON CORRELATION

The association between the prediction output (PD) and the CCHS data is calculated using the Pearson correlation coefficient (ρ) based on the following equation:

$$Correl(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The correlation coefficient ranges from -1 to 1. The higher the correlation coefficient would indicate a significant and positive relationship between the two sets of variables.

V. USER-LEVEL CLASSIFICATION METHOD AND RESULTS

We employed deep learning techniques to classify information extracted from the postings into four classes except for questions 16 and 18, which have seven choices (seven classes).

For answering the 21 questions of the BDI questionnaire, we trained $n * 21$ classifiers and formed the *BDI_Multi_Model* based on the best performing classifier on a specific question, where n is the total number of models using k different parameters, as illustrated in Figure 4. The performance is measured based on the evaluation criteria of the validation set described in Section IV. The n classifiers were built based on the filtering methods, pretrained models, deep learning architectures including different hyperparameters, and custom parameters as described below.

A. FILTERING METHODS

For each question, we focused on selecting a subset of posts for each user. The goal is to keep the most relevant posts to each question to increase the probability of finding an answer to the topic of the question. We call this process filtering of the posts. We also experimented with keeping all posts, with the caveat that training deep learning models on long texts (the concatenation of all posts) is slow or sometimes problematic for BERT-like models.

We employed three methods for filtering posts: topic-based, similarity-based, and a hybrid approach, as follows:

1) TOPIC-BASED FILTERING

Topic models can be valuable tools for discovering latent topics in collections of documents. As explained in our eRisk shared task paper [30],[6] we leveraged topic modeling using `top2vec` [31] to help identify relevant posts for each question. We used `top2vec` to divide all posts into topics that were then used to find relevant posts for each question. The `top2vec` algorithm automatically finds the number of topics in a corpus. It is an unsupervised learning algorithm that finds topic vectors in a semantic space from a jointly embedded document and word vectors of a corpus. The algorithm assumes that the dense area of document vectors represents an area of highly similar documents representative of a topic. Hence, a topic vector is calculated from each dense area of documents as the centroid of those document vectors. The topics are then described with the nearest word vectors to the topic vector. Then, each document is assigned to its nearest topic vector, allowing for the size of each topic to be calculated. Using the default Doc2Vec to generate the common word and document embeddings, the number of topics generated was 1, 328, whereas the topic identification got enhanced using pretrained models. Employing the *"distiluse-base-multilingual-cased"* pretrained model produced 332 topics, while using *"universal-sentence-encoder"*, `top2vec` defined 83 topics in the R1 dataset. To select posts based on `top2vec` models, we applied the algorithm 1:

Algorithm 1 Top2Vec Post Selection Algorithm

Input: p_ψ:clean posts, υ: list of users, θ: Threshold, $\mathbb{C}$: BDI_Categories
Output: $\vec{p}$: selected posts,τ:BDI topics
Begin
  $\forall \psi \in \upsilon$
    top2vec (p_ψ)
    $\forall c \in \mathbb{C}$
      $\tau \leftarrow$topics(p_ψ)
      if $\tau \geq \theta$
        $\vec{p} \leftarrow p_\psi$
End

For instance: `top2vec` had identified the top topics shown in Figure 5 for the "suicidal thoughts or wishes" category, whereas for "changes in appetite" related to question-18 the identified topics are illustrated in Figure and 6.

2) SIMILARITY-BASED METHOD

The similarity-based method utilizes pre-trained sentence transformer models based on *BERT*, *RoBERTa*, or *all-mpnet-base-v2* to embed each post and all the BDI answers instead of randomly initializing the embedding vectors. Then, the relatedness of each post to each answer of BDI answers is measured by computing the cosine distance between the post_embedding (p) and the answer_embedding (a),

[6]This article uses the models for our shared task paper and builds a new multi-model that achieves better performance than all the state-of-the-art models. In addition, it applies and tests this new model to population-level social media data.

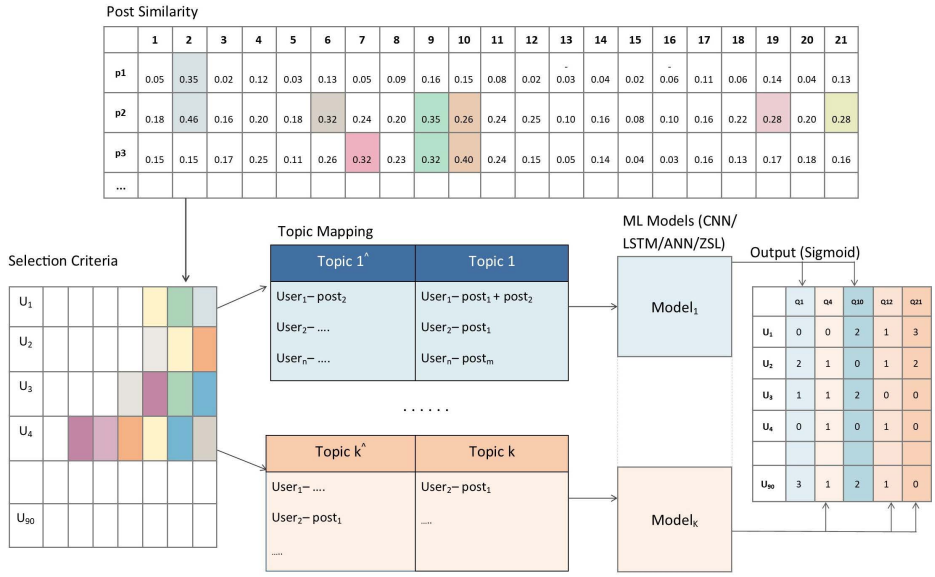

FIGURE 4. Representation of *BDI_Multi_Model* architecture for $Model_1$ and $Model_4$.

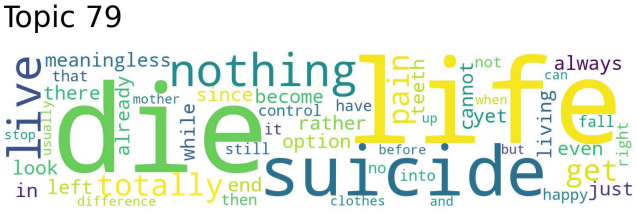

FIGURE 5. The word cloud of the top topic (Topic 79) related to Q9 "Suicide thoughts" using Top2Vec.

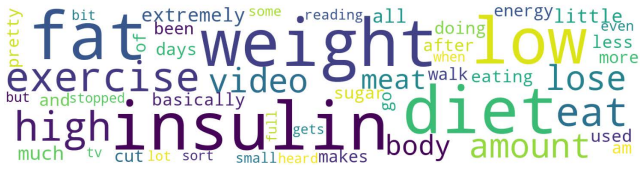

FIGURE 6. The word cloud of the top topic (Topic 15) related to Q18 "Changes in Appetite" using Top2Vec.

as shown in the Equation (3).[7]

$$1 - \frac{p \cdot a}{\|p\|_2 \cdot \|a\|_2} \text{ where } \|x\|_2 \text{ is the 2-norm of x} \qquad (3)$$

If the similarity value of a post with all questionnaires' answers is less than θ_1, then the post is excluded since it means that the post is not related to any of the BDI questionnaire questions. Additionally, we exclude general posts that may not assist in answering any of the BDI questionnaire questions using the coefficient of variation (CV) as a measure of relative variability as shown in Equation (4).

$$CV = \frac{1}{21}\sum_{i=1}^{21}(x_i - \overline{x})^{21} - \overline{\mu} \qquad (4)$$

where x is the similarity score of question i and μ is the mean

Furthermore, it should be noted that not all the categories are discussed in the posts and some categories appear more often than others. If there are a limited number of posts for a specific user, we would consider all the posts for the learning process (all the posts are included in the top n posts). Table 7 shows the number of posts for each BDI question as per *RoBERTa* similarity with$\theta_1 = 0.6$. It shows that the posts related to eating and sleeping habits are significantly less common, and posts relating to guilt and punishment feelings are the most frequent.

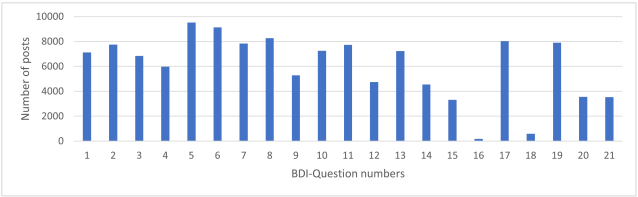

FIGURE 7. Number of posts per BDI-question based on RoBERTa ($\theta_1 = 0.6$).

3) HYBRID APPROACH

The Hybrid approach uses a combination of topic and similarity-based approaches based on different sentence transformer models and topic modeling. We used the *all-mpnet-base-v2* sentence transformer, which was developed by HuggingFace.[8] *all-mpnet-base-v2* is trained on more

[7]https://docs.scipy.org/doc/scipy/reference/generated/

[8]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

than 1 billion training sentence pairs, maps posts to a 768-dimensional dense vector space that is used for computing the semantic similarity between the questionnaire's questions and answers on one side and the users' posts on the other side.

The similarity is calculated based on different factors:

- Similarity to the question header.
- Similarity to any of the questions' answers.
- The topic of the post is classified using zero-shot learning as one of the main fields of the questionnaire with a probability of more than 0.5, specifically the following topics: {sad, encourage, fail, pleasure, guilty, trouble, confident, blame, suicide, cry, upset, activity, decisions, useful, make, sleep, bother, eat, focus, tired, sex}.

Figure 8 shows the distribution of the posts for every question in general based on the hybrid method.

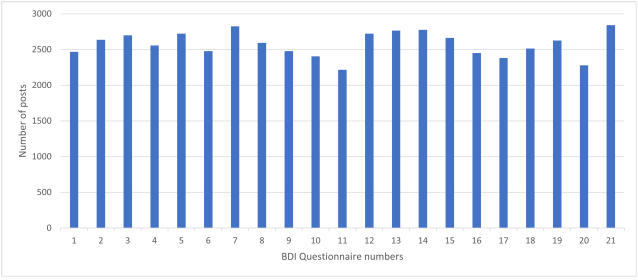

FIGURE 8. Number of posts per BDI-question based on topic and "all-mpnet-base-v2" model for similarity ($\theta_1 = 0.5$).

B. PRETRAINED MODELS

We used language models based on Sentence Transformers for deep contextual post representations: All-mpnet-base-v2, Sentence-BERT (SBERT) and Sentence-RoBERTa (SRoBERTa) [32]. All-mpnet-base-v2 used the pretrained microsoft/mpnet-base model and fine-tuned it on more than 1 billion sentence pairs. Whereas, SBERT/SRoBERTa employs siamese and triplet network architectures [33] as illustrated in Figure 9.

C. DEEP LEARNING ARCHITECTURES

In the following sections, we explain each deep learning model that contributed to our *BDI_Multi_Model*.

1) HIERARCHICAL ATTENTION NETWORK

One component of the *BDI_Multi_Model* is a hierarchical attention network (HAN) for document classification inspired by [34]. We used the Hierarchical Attention Network (HAN) for the multi-classification task for each category of the BDI questionnaire. We trained 21*k HAN classifiers using the top-related posts for each category- based on k parameter settings, detailed in Table 4. HAN employs bi-directional GRU on the word level, followed by an attention model, to extract the most informative words, which are then aggregated to generate a sentence vector. Similarly, bi-directional LSTM

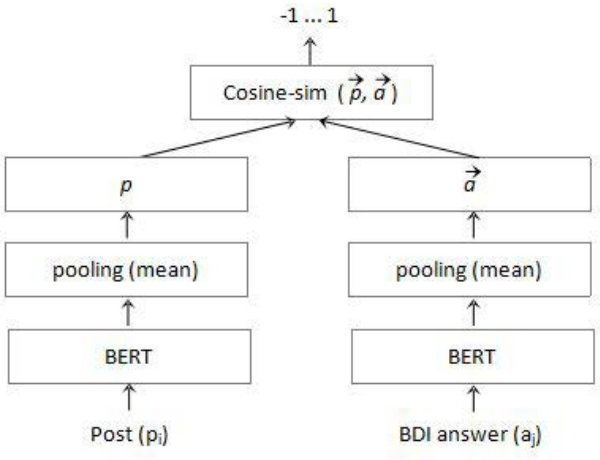

FIGURE 9. SBERT architecture (https://www.sbert.net/).

on the sentence level is used with an attention mechanism to aggregate the most significant sentences to form the user-category vector, which is then passed on to a dense layer for text classification using softmax activation as shown in Figure 10.

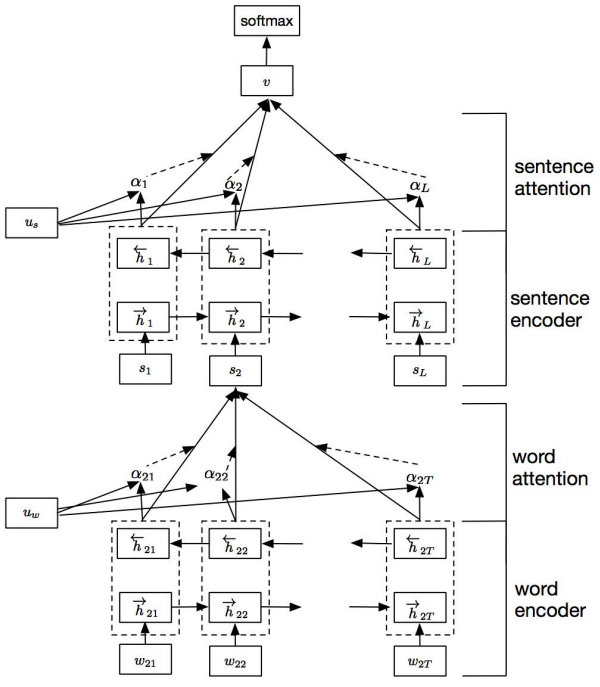

FIGURE 10. Hierarchical attention network [34].

HAN employs two levels of attention mechanisms at the word and sentence levels. First, a word attention mechanism is utilized to identify keywords and then aggregate them to create a sentence vector. Then a sentence attention mechanism is used to emphasize the importance of a sentence.

2) LONG SHORT-TERM MEMORY

In addition, we trained Bi-LSTM models. Starting with posts tokenization, followed by placing words in an indexed dictionary, then a sequence of indices for the words was fed to the embedding layer in an LSTM network using pre-trained word embeddings. The 1D CNN model inspired by Kim2014 is trained on top of pre-trained word embeddings on a sentence level. We applied the max-overtime pooling operation to capture essential features. Different hyperparameters were applied to tune the model as explained in Table 4.

3) TRANSFORMERS

Transformer attention models introduced in [35], use Scaled Dot-Product attention. Transformer models use the Multi-Head Attention layer, which runs in parallel, and has multiple scaled dot-product attention and multiple linear transformations (learnable parameters) of the input queries, keys, and values. It contains a transformer encoder starting with a Multi-head Attention module that performs self-attention (each word in the input attends to all other words in the input). Self-attention gives a representation of the meaning of each word within the sentence. Followed by a residual connection that helps prevent the vanishing gradient problem and keep the original 'state' information. The encoder provides 'Context' for each item in the input sequence. Masked Self-Attention is when each position only attends to previous positions (not every single word).

TABLE 4. Hyperparameters settings for *H: HAN, L:LSTM, T: Transfomer* models.

Parameters	Ranges	ML
Pretrained_models	{all-mpnet-base-v2, SBERT, SRoBERTa }	H,L,T
Word embeddings	{fastText, GloVe, GoogleNews-vectors}	H,L
Similarity method	{Topic, Similarity, Hybrid}	H,L,T
Max_features	{2000,5000,10000}	H,L
No_of_posts	{10,30,50,100}	H,L,T
Batch_size	{64, 128, 256}	H,L
Max_len	{128, 256, 512}	H,L,T
Embed_size	{50, 300}	H,L
Filter_sizes	{[1, 3, 5], [2, 3, 4]}	L
Num_filters	{32, 64, 128}	L
Dropout	{0.2, 0.3, 0.5}	N,L

4) BDI_MULTI_MODEL

As explained earlier, we adapted different deep learning models to learn the answers to each question in the BDI questionnaire based on the related posts. Figure 4 shows an example of two components of *BDI_Multi_Model*. First, models were built based on topic-based filtering of the related posts. Then, several iterations for model selection were done to end up assigning Model_1 to answer questions 1 and 10, and likewise promoting Model_4 to answer questions 4, 12, 21.

Table 5 shows the results of the selected models $Model_i$: $1 \leq i \leq 10$ to all questions. Finally, we designed an algorithm that selects the best model for each question and groups the models once applicable.

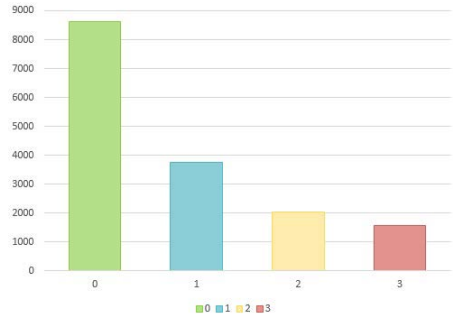

FIGURE 11. Distribution of depression levels based on $\hat{P}1$, where 0: no depression signs, 1: mild signs, 2: moderate signs, 3:server signs.

The *BDI_Multi_Model* is formed after several iterations of the above-mentioned deep learning models ending up with five hierarchical attention networks (HAN), three LSTM models, and finally, two transformers. The models and the parameters were set - based on the accuracy for each question - ending up with ten different versions. The ten models' parameters are included in Appendix VII. Although there is not much improvement in the ACR metric, the performance of *BDI_Multi_Model* exceeded the latest best model [22] using the same training and test dataset in the following metrics: AHR, ADL, and DCHR with a difference of more than 7% of the latest. In addition, the ADL metric used to predict the depression level exceeded 84%, considering that ADL is the most critical metric for measuring depression at the population level.

The model's performance is enhanced in small steps due to the lack of an adequate dataset for deep learning model training. The training dataset contains only 90 users, and the total number of training posts is less than 50, 000, which is relatively small for a deep learning model. The quality of the posts can be enhanced by filtering the indicative posts by experts. This labeling would help train a model to filter the data based on the extracted features, which may help enhance the posts filtering process and the classifiers' accuracies.

TABLE 5. The evaluation metric of *BDI_Multi_Model* subset models ($Model_i$: $1 \leq i \leq 10$) to all BDI_questions.

Model Name	AHR	ACR	ADL	DCHR
Best-2021	35.36	**73.17**	83.59	**41.25**
$Model_1$	30.47	65.61	83.67	38.75
$Model_2$	29.82	65.75	84.20	46.25
$Model_3$	28.45	65.17	82.48	42.50
$Model_4$	31.96	66.94	83.81	37.50
$Model_5$	24.11	60.14	80.14	43.75
$Model_6$	34.10	66.79	76.47	46.25
$Model_7$	34.52	66.65	71.59	47.50
$Model_8$	38.33	69.15	82.88	36.25
$Model_9$	33.81	66.37	83.79	46.25
$Model_{10}$	35.42	66.96	**84.38**	41.25
BDI_Multi_Model	**41.25**	70.40	83.79	**48.75**

VI. POPULATION LEVEL

In this research, we use a bottom-up technique for population-level detection [27], starting with individual models, then

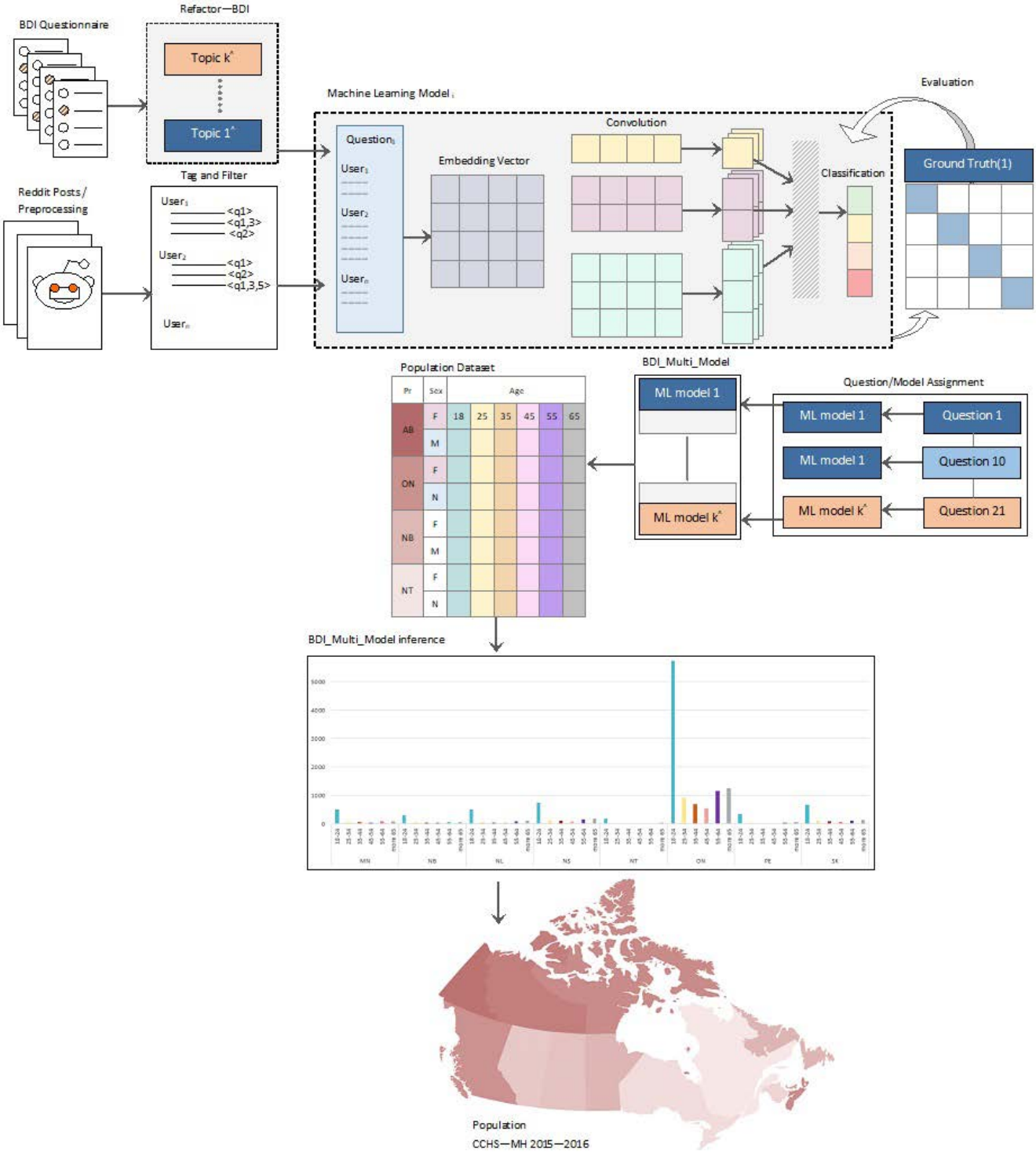

FIGURE 12. From user-level to population-level methodology.

generalizing to make an inference about the population. Finally, comparing the model's predicted results with official statistics (in our case CCHS) published in the same year as the dataset under study.

The overall process is illustrated in Figure 12, and it can be summarized as follows: Starting with Reddit posts pre-processing and BDI questionnaire responses, each post is examined against the BDI questionnaire's topics and tagged

TABLE 6. Pearson linear correlation between P_CCHS and $\hat{P}1$ among males and females for the provinces included in 2015-2016 survey where, P_CCHS is CCHS-MH for 2015-2016, $\hat{P}1$:the prediction on the P1 based on *BDI_Multi_Model* model.

Province/Sex	Tweets	Users	P_CCHS	$\hat{P}1$
MN				
Female	117,496	833	55	72
Male	935	10	77	10
NB				
Female	88,624	531	46	26
Male	3,003	13	60	13
NL				
Female	121,926	519	36	55
Male	56,361	341	31	26
NS				
Female	161,987	786	154	78
Male	109,227	612	142	44
NT				
Female	43,772	283	8	17
Male	668	5	9	5
ON				
Female	936,660	5,481	750	579
Male	697,574	4,853	833	455
PE				
Female	53,137	325	18	47
Male	30,225	203	19	18
SK				
Female	33,861	258	80	61
Male	127,456	929	123	53
Total	2,582,912	15,982	2,441	1,559

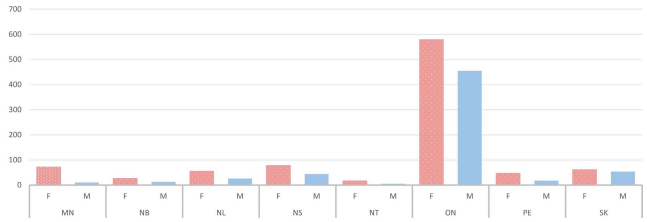

FIGURE 13. Predicted depressed users using the *BDI_Multi_Model* on the P1 dataset.

appropriately. Afterward, a classifier is trained using a subset of the user postings to respond to the BDI query. The classifier is then assessed using various models and hyperparameters until it reaches the settings that maximize the performance according to the evaluation set - which is 20% of the training dataset in our case. Accordingly, the question is assigned to the best-performing model. This process is repeated for each question, resulting in a model mapped to each question in the BDI questionnaire represented as $(Machine - Learning - Model_i)$. Finally, those models are grouped to form the BDI_Multi_Model that is used to predict the answers of Twitter users representative of the Canadian population (P1). As mentioned, P1 contains 15, 982 Twitter users who posted 2, 582, 912 tweets during 2015. Accordingly, the depression level is calculated for each user, and the number of users per depression category is calculated for each province. The levels of depression in P1 is presented in Figure 11. The users included in the third level of depression, i.e., categorized with severe depression, are referred to as $\hat{P}1$.

TABLE 7. Correlation between P_CCHS:2015-2016 and $\hat{P}1$:the prediction on the P1 dataset based on the *BDI_Multi_Model* model among different age_groups for the provinces included in 2015-2016 survey.

Province/Age	P_CCHS (CCHS)	# $\hat{P}1$ with level 3
MN		
18-24	30,021	500
25-34	38,763	66
35-44	31,041	60
45-54	31,414	46
55-64	30,180	83
>65	15,170	88
NB		
18-24	17,309	307
25-34	18,590	56
35-44	22,113	34
45-54	28,316	29
55-64	21,845	52
>65	12,728	66
NL		
18-24	10,310	500
25-34	13,506	66
35-44	16,532	47
45-54	14,702	51
55-64	11,984	95
>65	8,600	101
NS		
18-24	21,015	740
25-34	33,233	134
35-44	23,527	110
45-54	36,694	75
55-64	28,659	160
>65	14,294	179
NT		
18-24	991	175
25-34	1,888	22
35-44	1,080	26
45-54	967	13
55-64	807	22
>65	257	30
ON		
18-24	387,535	5,745
25-34	425,760	924
35-44	353,871	693
45-54	416,661	550
55-64	327,139	1,167
>65	169,536	1,255
PE		
18-24	2,961	332
25-34	3,462	47
35-44	3,441	23
45-54	2,809	23
55-64	2,790	48
>65	1,829	55
SK		
18-24	20,443	678
25-34	42,576	112
35-44	29,197	78
45-54	30,913	73
55-64	25,203	120
>65	11,750	126
Grand Total	15,982	160,082

The inference of *BDI_Multi_Model* estimated that 9.75% of the sample is classified as depressed ($|\hat{P}1|$), whereas 54% are classified with non depressed or minimal depression, as opposed to the 7.1% estimated official prevalence rate for depression in 2015.

Table 6 demonstrates a consistent association between the number of users predicted as severe depression and the ones reported as depressed based on CCHS-MH for the same year with $\rho = 0.97$ that indicates a strong positive relationship.

Figure 13 and 14 show that the correlation between the estimated depressed males versus females among the population complies with Canadian population studies indicating that women have a higher prevalence of major depressive episodes than men [36]. $\hat{P}1$ estimated 56% prevalence of depression among females. At the same time, Figure 13 shows the males/females prevalence within the P_CCHS survey and the *BDI_Multi_Model* estimates.

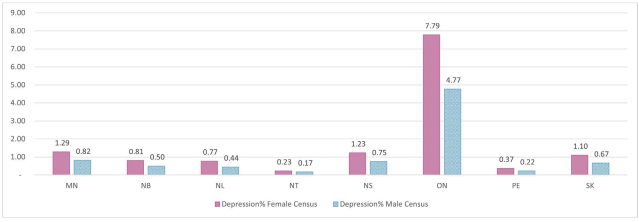

FIGURE 14. CCHS 2015-2016 Survey Results (P_CCHS).

Similarly, Table 6 shows the distribution of age demographics within 7 of the Canadian provinces and the NT territory, and the estimated depression on the P1 dataset. The age distribution shows bias towards younger age in the population, mainly for the age group $18 - 24$. This is due to the demographics of social media users. Thirty percent of internet users under the age of 50 use Twitter, compared to eleven percent of online users aged fifty and more [37]. Nevertheless, our estimated $\hat{P}1$ users showed a good correlation with the CCHS statistics data. The results showed a positive correlation between the predicted depressed users and the available statistics for 2015 at the province/age level ($\rho = 0.6$).

VII. CONCLUSION

Analysis of social media posts is a helpful tool for quickly seeing patterns and diagnosing psychiatric illnesses in a defined population. Furthermore, the BDI questionnaire is a valuable tool for evaluating the level of depression. Different filtering approaches have been applied and fed to different machine learning architectures that were later on evaluated to check their ability to answer each BDI question. Based on the defined evaluation metric, a multi-model has been trained to select the best model/parameters that yield better accuracy for answering the BDI questions. As a result, the automatic answering model can be applied at the population level on a large scale and can help monitor trends and risks by authorities.

However, the use of this tool differs in the psychiatric environment from social media settings. We suggest that the BDI questionnaire could be revised in future work to be adapted to social media characteristics. In addition, the eRisk dataset could be enhanced by experts to label the questionnaire-related posts so that a reliable automated questionnaire model can be trained to fill up the questionnaire for a better estimation of the depression level of a defined population.

APPENDIX A BDI INSTRUCTIONS

This questionnaire (BDI-II) consists of 21 groups of statements. Please read each group of statements carefully. And then pick out the one statement in each group that best describes the way you have been feeling during the past two weeks, including today. Circle the number beside the statement you have picked. If several statements in the group seem to apply equally well, circle the highest number for that group. Be sure that you do not choose more than one statement for any group, including Item 16 (Changes in Sleeping Pattern) or Item 18 (Changes in Appetite).

1) Sadness
   0 I do not feel sad.
   1 I feel sad much of the time.
   2 I am sad all the time.
   3 I am so sad or unhappy that I can't stand it.
2) Pessimism
   0 I am not discouraged about my future.
   1 I feel more discouraged about my future than I used to.
   2 I do not expect things to work out for me.
   3 I feel my future is hopeless and will only get worse.
3) Past Failure
   0 I do not feel like a failure.
   1 I have failed more than I should have.
   2 As I look back, I see a lot of failures.
   3 I feel I am a total failure as a person.
4) Loss of Pleasure
   0 I get as much pleasure as I ever did from the things I enjoy.
   1 I don't enjoy things as much as I used to.
   2 I get very little pleasure from the things I used to enjoy.
   3 I can't get any pleasure from the things I used to enjoy.
5) Guilty Feelings
   0 I don't feel particularly guilty.
   1 I feel guilty over many things I have done or should have done.
   2 I feel quite guilty most of the time.
   3 I feel guilty all of the time.
6) Punishment Feelings
   0 I don't feel I am being punished.
   1 I feel I may be punished.
   2 I expect to be punished.
   3 I feel I am being punished.
7) Self-Dislike
   0 I feel the same about myself as ever.
   1 I have lost confidence in myself.
   2 I am disappointed in myself.
   3 I dislike myself.
8) Self-Criticalness

0 I don't criticize or blame myself more than usual.
1 I am more critical of myself than I used to be.
2 I criticize myself for all of my faults.
3 I blame myself for everything bad that happens.

9) Suicidal Thoughts or Wishes
0 I don't have any thoughts of killing myself.
1 I have thoughts of killing myself, but I would not carry them out.
2 I would like to kill myself.
3 I would kill myself if I had the chance.

10) Crying
0 I don't cry anymore than I used to.
1 I cry more than I used to.
2 I cry over every little thing.
3 I feel like crying, but I can't.

11) Agitation
0 I am no more restless or wound up than usual.
1 I feel more restless or wound up than usual.
2 I am so restless or agitated, it's hard to stay still.
3 I am so restless or agitated that I have to keep moving or doing something.

12) Loss of Interest
0 I have not lost interest in other people or activities.
1 I am less interested in other people or things than before.
2 I have lost most of my interest in other people or things.
3 It's hard to get interested in anything.

13) Indecisiveness
0 I make decisions about as well as ever.
1 I find it more difficult to make decisions than usual.
2 I have much greater difficulty in making decisions than I used to.
3 I have trouble making any decisions.

14) Worthlessness
0 I do not feel I am worthless.
1 I don't consider myself as worthwhile and useful as I used to.
2 I feel more worthless as compared to others.
3 I feel utterly worthless.

15) Loss of Energy
0 I have as much energy as ever.
1 I have less energy than I used to have.
2 I don't have enough energy to do very much.
3 I don't have enough energy to do anything.

16) Changes in Sleeping Pattern 0 I have not experienced any change in my sleeping.
1a I sleep somewhat more than usual.
1b I sleep somewhat less than usual.
2a I sleep a lot more than usual.
2b I sleep a lot less than usual.
3a I sleep most of the day.
3b I wake up 1-2 hours early and can't get back to sleep.

17) Irritability
0 I am not more irritable than usual.
1 I am more irritable than usual.
2 I am much more irritable than usual.
3 I am irritable all the time.

18) Changes in Appetite 0. I have not experienced any change in my appetite.
1a My appetite is somewhat less than usual.
1b My appetite is somewhat greater than usual.
2a My appetite is much less than before.
2b My appetite is much greater than usual.
3a I have no appetite at all.
3b I crave food all the time.

19) Concentration Difficulty
0 I can concentrate as well as ever.
1 I can't concentrate as well as usual.
2 It's hard to keep my mind on anything for very long.
3 I find I can't concentrate on anything.

20) Tiredness or Fatigue
0 I am no more tired or fatigued than usual.
1 I get more tired or fatigued more easily than usual.
2 I am too tired or fatigued to do a lot of the things I used to do.
3 I am too tired or fatigued to do most of the things I used to do.

21) Loss of Interest in Sex
0 I have not noticed any recent change in my interest in sex.
1 I am less interested in sex than I used to be.
2 I am much less interested in sex now.
3 I have lost interest in sex completely.

TABLE 8. Parameters for Posts Filtering and BDI_MultiModel hyperparameters.

Model	Parameters
$Model_1$	
Filtering	Similarity, RoBerta, $p = 10$, $\theta = 0.6$
Architecture	HAN, $Glove$,M=50000, D=0.3, G=20
$Model_2$	
Filtering	Similarity, RoBerta, $p = 10$, $\theta = 0.6$
Architecture	LSTM, $Glove$,M=10000, D=0.3
$Model_3$	
Filtering	Similarity, bert_base_nli_tokens, $p = 50$, $\theta = 0.5$
Architecture	HAN,$GoogleNews$, M=5000, D=0.3, G=100
$Model_4$	
Filtering	Similarity, , $p = 50$, $\theta = 0.5$
Architecture	Transformers, SBERT, L=256
$Model_5$	
Filtering	Similarity, , $p = 100$, $\theta = 0.25$
Architecture	LSTM, $Twitter$, M = 3000, D=0.2
$Model_6$	
Filtering	Hybrid, all-mpnet-base-v2, $p = 100$, $\theta = 0.5$
Architecture	HAN,$GoogleNews$, M=10000, D=0.3, G=100
$Model_7$	
Filtering	Hybrid, all-mpnet-base-v2, $p = 50$, $\theta 1 = 0.25$
Architecture	HAN,$Twitter$, M=5000, D=0.3, G=100
$Model_8$	
Filtering	Similarity, RoBerta, $p = 50$, $\theta = 0.5$
Architecture	HAN,$GoogleNews$, M=5000, D=0.3, G=100
$Model_9$	
Filtering	Hybrid, all-mpnet-base-v2, $p = 50$, $\theta 1 = 0.25$
Architecture	LSTM, $GoogleNews$,M=5000, D=0.3
$Model_{10}$	
Filtering	Hybrid, , $p = 100$, L=512 $\theta = 0.5$
Architecture	Transformers, SRoBERTa

APPENDIX B BDI_MULTIMODEL PARAMETERS

Table 8 specifies the main parameters changed per model in terms of posts filtering and deep learning model architecture and hyperparameters, whereas Table 9 shows the grouping of the questions and the corresponding model.

p: top_similarity
M: Max_Features
D: Dropout
G: GRU Units
L: Maximum length
GoogleNews:Pretrained word vectors based on Google News dataset (100B tokens, 3M vocab, 300d vectors)[9]
Glove:Pretrained word vectors based on Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors[10]
Twitter: Pretrained word vectors based on Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 200d vectors)[11]

TABLE 9. Models and BDI questions mapping.

Model	Question	Title
1	Q1	Sadness
	Q10	Crying
2	Q2	Pessimism
	Q9	Suicidal thoughts or wishes
3	Q3	Past failure
	Q8	Self-criticalness
4	Q4	Loss of pleasure
	Q12	Loss of interest
	Q21	Loss of interest in sex
5	Q5	Guilty feelings
	Q6	Punishment feelings
6	Q7	Self-dislike
	Q14	Worthlessness
7	Q11	Agitation
	Q17	Irritability
8	Q13	Indecisiveness
	Q19	Concentration difficulty
9	Q15	Loss of energy
	Q20	Tiredness or fatigue
10	Q16	Changes in sleeping pattern
	Q18	Changes in appetite

REFERENCES

[1] J. Upton, *Beck Depression Inventory (BDI)*. New York, NY, USA: Springer, 2013, pp. 178–179, doi: 10.1007/978-1-4419-1005-9_441.
[2] S. J. Teague, A. B. R. Shatte, E. Weller, M. Fuller-Tyszkiewicz, and D. M. Hutchinson, "Methods and applications of social media monitoring of mental health during disasters: Scoping review," *JMIR Mental Health*, vol. 9, no. 2, Feb. 2022, Art. no. e33058. [Online]. Available: https://mental.jmir.org/2022/2/e33058/
[3] J. Kim, D. Lee, and E. Park, "Machine learning for mental health in social media: Bibliometric study," *J. Med. Internet Res.*, vol. 23, no. 3, Mar. 2021, Art. no. e24870. [Online]. Available: https://www.jmir.org/2021/3/e24870/
[4] A. L. Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, and C. Lemey, "Machine learning and natural language processing in mental health: Systematic review," *J. Med. Internet Res.*, vol. 23, no. 5, May 2021, Art. no. e15708. [Online]. Available: https://www.jmir.org/2021/5/e15708/
[5] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–11, Dec. 2020. [Online]. Available: https://www.nature.com/articles/s41746-020-0233-7
[6] R. Skaik and D. Inkpen, "Using social media for mental health surveillance: A review," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–31, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3422824
[7] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," *J. Med. Internet Res.*, vol. 19, no. 6, p. e228, Jun. 2017. [Online]. Available: https://www.jmir.org/2017/6/e228/
[8] D. Howard, M. M. Maslej, J. Lee, J. Ritchie, G. Woollard, and L. French, "Transfer learning for risk classification of social media posts: Model evaluation study," *J. Med. Internet Res.*, vol. 22, no. 5, May 2020, Art. no. e15371. [Online]. Available: https://www.jmir.org/2020/5/e15371/
[9] Biradar and S. G. Totad, "Detecting depression in social media posts using machine learning," in *Proc. Int. Conf. Recent Trends Image Process. Pattern Recognit.* Singapore: Springer, 2018, pp. 716–725.
[10] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–12, Dec. 2018. [Online]. Available: https://link.springer.com/article/10.1007/s13755-018-0046-0
[11] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104499. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0010482521002936
[12] K. A. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on Twitter data," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2021, pp. 960–966. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9432203
[13] P. Verma, K. Sharma, and G. S. Walia, "Depression detection among social media users using machine learning," in *Proc. Int. Conf. Innov. Comput. Commun.*, D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, and A. Jaiswal, Eds. Singapore: Springer, 2021, pp. 865–874.
[14] N. A. Asad, M. A. Mahmud Pranto, S. Afreen, and M. M. Islam, "Depression detection by analyzing social media posts of user," in *Proc. IEEE Int. Conf. Signal Process., Inf., Commun. Syst. (SPICSCON)*, Nov. 2019, pp. 13–17. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9065101
[15] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., Keyboard Clinic*, Jan. 2018, pp. 88–97. [Online]. Available: https://aclanthology.org/W18-0609/
[16] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," Assoc. Comput. Linguistics, Vancouver, BC, Canada, Aug. 2017, pp. 32–40. [Online]. Available: https://aclanthology.org/W17-3104/
[17] M. D. Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proc. 5th Annu. ACM Web Sci. Conf. (WebSci)*, 2013, pp. 47–56. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2464464.2464480
[18] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, "A machine learning approach predicts future risk to suicidal ideation from social media data," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–12, Dec. 2020. [Online]. Available: https://www.nature.com/articles/s41746-020-0287-6
[19] T. Nguyen, D. T. Nguyen, M. E. Larsen, B. O'Dea, J. Yearwood, D. Phung, S. Venkatesh, and H. Christensen, "Prediction of population health indices from social media using kernel-based textual and temporal features," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 99–107. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3041021.3054136
[20] M. D. Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 2098–2110. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2858036.2858207
[21] D. Su, X. Zhang, K. He, and Y. Chen, "Use of machine learning approach to predict depression in the elderly in China: A longitudinal study," *J. Affect. Disorders*, vol. 282, pp. 289–298, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016503272033250X
[22] (2021). *A RoBERTa-Based Model on Measuring the Severity of the Signs of Depression*. [Online]. Available: http://ceur-ws.org/Vol-2936/paper-86.pdf

[9]https://code.google.com/archive/p/word2vec/
[10]hhttps://nlp.stanford.edu/data/glove.840B.300d.zip
[11]https://nlp.stanford.edu/data/glove.twitter.27B.zip

[23] D. E. Losada and F. Crestani, *A Test Collection for Research on Depression and Language Use*. Cham, Switzerland: Springer, 2016, pp. 28–39.
[24] K. White, G. Li, and N. Japkowicz, "Sampling online social networks using coupling from the past," in *Proc. IEEE 12th Int. Conf. Data Mining Workshops*, J. Vreeken, C. Ling, M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, Eds. Brussels, Belgium: IEEE Computer Society, Dec. 2012, pp. 266–272. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6406450
[25] K. White, "Forecasting Canadian elections using Twitter," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9673. Cham, Switzerland: Springer, 2016, pp. 186–191.
[26] S. Daneshvar, "User modeling in social media: Gender and age detection user modeling in social media," M.S. thesis, School Elect. Eng. Comput. Sci., Univ. Ottawa, 2019. [Online]. Available: https://ruor.uottawa.ca/handle/10393/39535
[27] R. Skaik, "Predicting depression and suicide ideation in the Canadian population using social media data," Ph.D. dissertation, School Elect. Eng. Comput. Sci., Univ. Ottawa, 2021. [Online]. Available: https://ruor.uottawa.ca/handle/10393/42346
[28] *Canadian Community Health Survey 2015–2016: Annual Component*, Statistics Canada, Ottawa, ON, Canada, 2017. [Online]. Available: https://www.odesi.ca
[29] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of eRisk 2021: Early risk prediction on the internet," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Cham, Switzerland: Springer, 2021, pp. 324–344.
[30] D. Inkpen, R. Skaik, P. Buddhitha, D. Angelov, and M. T. Fredenburgh, "uOttawa at eRisk 2021: Automatic filling of the beck's depression inventory questionnaire using deep learning," in *Proc. CLEF*, 2021, pp. 966–980. [Online]. Available: http://ceur-ws.org/Vol-2936/paper-79.pdf
[31] D. Angelov, "Top2 Vec: Distributed representations of topics," 2020, *arXiv:2008.09470*.
[32] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2019, pp. 1–11. [Online]. Available: https://arxiv.org/abs/1908.10084
[33] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[34] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, San Diego, CA, USA, Jun. 2016, pp. 1480–1489. [Online]. Available: https://aclanthology.org/N16-1174.pdf
[35] 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.*, vol. 30, 2017, pp. 1–11.
[36] C. Pearson, T. Janz, and J. Ali, "Mental and substance use disorders in Canada," Health Glance, Statist. Canada, Tech. Rep. 82-6Sep.24-X, 2018. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/82-624-x/2013001/article/11855-eng.htm
[37] M. Duggan, *The Demographics of Social Media Users*. Washington, DC, USA: PEW Research Center, Aug. 2015. [Online]. Available: https://www.pewresearch.org/internet/2015/08/19/the-demographics-of-social-media-users/

RUBA S. SKAIK received the B.Sc. and M.Sc. degrees in computer science from Kuwait University, and the Ph.D. degree from the School of Electrical Engineering and Computer Science, University of Ottawa. She is currently the Automation Team Lead at Quality Assurance, Environment and Climate Change Canada. Her research interests include machine learning, natural language processing, and distributed database. She is an Oracle Certified Professional with 20 years of experience in development, system analysis, and project management.

DIANA INKPEN received the B.Eng. and M.Sc. degrees in computer science and engineering from the Technical University of Cluj-Napoca, Romania, and the Ph.D. degree from the Department of Computer Science, University of Toronto. She is currently a Professor at the School of Electrical Engineering and Computer Science, University of Ottawa. Her research is in applications of natural language processing and text mining. She organized seven international workshops and she was a Program Co-Chair for the 25th Canadian Conference on Artificial Intelligence (AI 2012, Toronto, ON, Canada, in May 2012) conference. She is the Editor-in-Chief of the *Computational Intelligence* journal and an Associate Editor of the *Natural Language Engineering* journal. She published a book on *Natural Language Processing for Social Media* (Morgan and Claypool Publishers, Synthesis Lectures on Human Language Technologies, the third edition appeared in 2020), ten book chapters, more than 35 journal articles, and more than 120 conference papers. She received many research grants, from which the majority include intensive industrial collaborations.