# Real-time Depression Risk Detection From Social Media Using Big Data

Minh Huu-Tuan Nguyen[1,2], Khang Nguyen-Trong Huynh[1,2], Hop Trong Do[1,2]

[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam.
[2]Vietnam National University, Ho Chi Minh City, Vietnam.

Contributing authors: 21520348@gm.uit.edu.vn; 21520958@gm.uit.edu.vn; hoptd@gm.uit.edu.vn;

**Abstract**

In recent years, mental health disorders, particularly depression, have become a critical public health issue worldwide. With the proliferation of social media platforms, people increasingly share their thoughts, feelings, and experiences online, creating a valuable repository of data that can be harnessed for mental health monitoring. We utilized post from specific subreddits on Reddit platform as the training dataset and the streaming data to the system. We trained three machine learning models from Pyspark's MLLib which are Logistic Regression, Support Vector Machine and Decision Tree and achived an accuracy of 91.82% in the test set. Then, an online streaming system is built to flag posts that exhibit indicators of depression. The findings of this research contribute to the growing field of digital mental health, demonstrating the feasibility and effectiveness of big data technologies in enhancing traditional mental health care.

**Keywords:** Depression Detection, Big Data, Machine Learning

## 1 Introduction

In recent years, mental health disorders, particularly depression, have become a critical public health issue worldwide. Depression affects millions of individuals, impacting their daily lives, productivity, and overall well-being. Traditional methods of diagnosing depression often rely on self-reported questionnaires and clinical interviews,

which, although effective, can be time-consuming and limited by accessibility constraints. With the proliferation of social media platforms, people increasingly share their thoughts, feelings, and experiences online, creating a valuable repository of data that can be harnessed for mental health monitoring.

This paper explores the potential of using big data analytics to detect signs of depression in real-time from social media activity. By leveraging machine learning and text pre-processing techniques, we aim to identify linguistic and behavioral patterns indicative of depression. The real-time aspect of this approach offers the possibility of timely intervention, providing support and resources to individuals when they need it most. Our study focuses on analyzing large volumes of social media posts from platform Reddit. We employ various data preprocessing methods to handle the noise and diversity of user-generated content, ensuring the reliability of our detection system. Additionally, we address the ethical considerations of using personal data for mental health monitoring, emphasizing the importance of privacy and informed consent.

The findings of this research contribute to the growing field of digital mental health, demonstrating the feasibility and effectiveness of big data technologies in enhancing traditional mental health care. By integrating real-time depression detection systems into social media platforms, we envision a future where mental health support is more accessible, proactive, and personalized.

Our paper is structured as follows. Section 2 shows previous works and attempts on the depression risk detection task. Next, Section 3 illustrates our proposed method, including the dataset, data-preprocessing techniques and the machine learning models, to build the classification model for the depression risk detection task. Section 4 describes the execution evaluation and practical results of the trained model in the online streaming system. Then, Section 5 shows our empirical results to find the advantages and disadvantages of the current models. Finally, Section 6 concludes our works and proposes further research.

## 2 Related Work

Determining the necessity for treatment in a mental disorder is a complex clinical decision, influenced by various factors. These include the intensity of symptoms, the distress experienced by patients due to these symptoms, the benefits and drawbacks of specific treatments, the disabilities caused by the symptoms, and the potential for symptoms to adversely affect other conditions. Assessing the severity of a disorder is also a challenging task that requires the expertise of a highly trained professional. This assessment involves various techniques, including text descriptions, clinical interviews, and professional judgment [1]. Given the complexity of the processes and the level of expertise required to diagnose mental disorders and determine appropriate treatments, using web mining and sentiment analysis techniques to detect mental illness on social media can be seen as an initial step to raise awareness.

Machine learning and deep learning architectures have been used in several researches in detecting signs of depression. Orabi [2] recommended the Continuous Bag of Words (CBOW) embedding technique to detect depression from Twitter data.

Syms and JS Raj in [3] proposed using N-gram language modeling to classify anxiety levels based on generated emotional characteristics, integrating these with topic analysis. Wonkoblap in [4] introduced a deep neural network method to analyze depression on social media, utilizing Twitter data for their study. Kim Jin in [5] explored the use of supervised machine learning algorithms to identify predictive factors for post-traumatic stress disorder, building models based on various language styles from Twitter users. Q un Nisa in [6] employed convolutional neural networks to compare different models for predicting emotions, relying on linguistic metadata. Their proposed strategy achieved state-of-the-art performance in several tasks.

Big data analysis systems have shown efficiency in real-time detection tasks, which needs low latency. Vo et al. [7] and Doan et al. [8] developed a real-time streaming system for social platforms in Vietnam, such as Facebook and YouTube, to identify hateful online comments. Khanh et al. [9] introduced an online streaming system designed to detect hate speech comments, utilizing efficient data-processing techniques to enhance the performance of classification models based on the ViHSD dataset [10].

To the best of our knowledge, there has not been any big data streaming system that assists real-time depression detection. Hence, in this project, we propose a new system that can treat the online streaming comments and detect post that indicates risk about depression automatically.

# 3 Methodology

In this section, we cover the models developed to address the tasks and the training process used to create the offline processor in our system. We discuss the dataset, models, pre-processing techniques, experimental setup, and results.

## 3.1 Dataset

We utilize a dataset from Kaggle comprising social media posts from the "SuicideWatch" and "depression" subreddits on Reddit. These posts were gathered using the Pushshift API. Posts from "SuicideWatch" were collected from its inception on December 16, 2008, until January 2, 2021, while posts from the "depression" subreddit were gathered from January 1, 2009, to January 2, 2021. Posts from "SuicideWatch" are designated as suicide-related, whereas those from the "depression" subreddit are classified as depression-related. Non-suicide posts were sourced from the "r/teenagers" subreddit. This dataset contains 232074 posts and the number of depression related posts is equal to the number of non-related ones.

## 3.2 Data Preprocessing

To account for noisy data present in social platform input, especially streaming video, we implement a data preprocessing process to enhance the dataset's quality and extract valuable features. Our pre-processing process includes these stages:

- Tokenize the text into words using RegexTokenizer which is more advanced tokenization based on regular expression matching

- Removal of Stopwords using StopWordsRemover. Default english stopwords are used to remove unwanted words from the data.
- Converting text to numerical. Word2Vec is an Estimator which takes sequences of words representing documents, computes distributed vector representation of words and trains a Word2VecModel. The model maps each word to a unique fixed-size vector. The Word2VecModel transforms each document into a vector using the average of all words in the document. The advantage of this representations is that similar words are close in the vector space, which makes generalization to novel patterns easier and model estimation more robust.
- Convert target categorical values into intergers using StringIndexer. StringIndexer encodes a string column of labels to a column of label indices.

## 3.3 Models

### 3.3.1 Decision tree

Decision tree is a tree-like model used for decision-making and classification, where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. The topmost node is called the root node, while the terminal nodes are referred to as leaf nodes. Decision trees are transparent machine learning algorithms capable of handling high-dimensional data with commendable accuracy.

In decision trees, the best feature for splitting at the root node is selected based on Attribute Selection Measures (ASMs) such as the Gini Index or Entropy. These measures are calculated at every step before a node is split until the desired number of nodes is achieved. In this context, the Gini Index is the ASM employed. The attribute with the highest Gini Index is chosen for the split. The Gini Index evaluates a binary split for each attribute using the following formula:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} P_i^2$$

where $P_i$ is the probability that a tuple in $D$ belongs to class $C_i$.

The decision parameters are as follows:

- labelCol: Name of the label column
- featureCol: Name of the features column
- impurity: Gini Index, with a default value of "gini"

### 3.3.2 Logistic Regression

Logistic regression is a widely used predictive modeling technique that belongs to the family of generalized linear models (GLMs). It is primarily used for binary classification problems where the outcome variable takes on two possible values, often represented as 0 and 1. Its popularity stems from its simplicity, interpretability, and effectiveness in various fields, including medicine, social sciences, and machine learning.

The logistic regression model estimates the probability that a given instance belongs to a particular category. The model is specified as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

where $P(Y = 1|X)$ is the probability of the outcome being 1 given the predictors $X_1, X_2, \ldots, X_k$, and $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients to be estimated.

A 3-fold cross-validation is performed to tune the parameters of the Linear Regression (LR), which are defined using *ParamGridBuilder*. The parameters tuned are:

- regParam: Controls the regularization strength in logistic regression models.
- elasticNetParam: determines the balance between L1 and L2 regularization. It takes a value between 0 and 1: 0 corresponds to pure L2 regularization. 1 corresponds to pure L1 regularization. Values between 0 and 1 represent a combination of L1 and L2 regularization.

### 3.3.3 Support Vector Classifier

LinearSVC aims to find the optimal hyperplane that separates data points of different classes with a maximum margin. The decision boundary is defined as:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input feature vector, and $b$ is the bias term.

The objective function to be minimized is given by:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))$$

where $C$ is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error.

A 3-fold cross-validation is performed to tune the parameters of the Linear Support Vector Classifier (LinearSVC), which are defined using *ParamGridBuilder*. The parameters tuned are:

- regParam: Controls the regularization strength in logistic regression models.

# 4 Online Streaming System

Motivating from previous efforts in developing real-time data streaming systems, we present a system designed to identify depression risk in real-time streaming comments from social media platforms. This system operates by analyzing comments as they are posted, flagging those that exhibit indicators of depression. Its architecture can be visualized in Figure 1

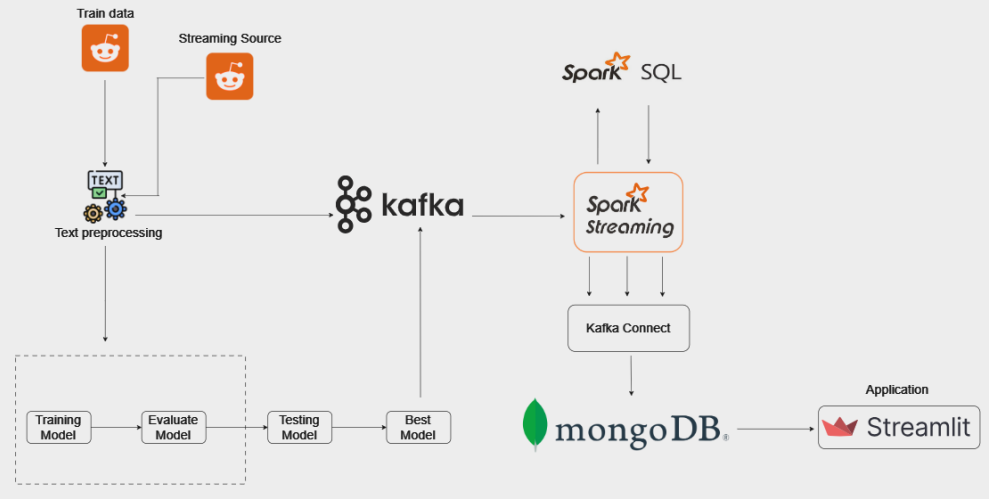


**Fig. 1** Our system's architecture for the detection of depression risk from social media text

**Streaming Processing System:** The continuous handling of vast quantities of live data necessitates a streaming processing system, which gathers real-time data from various sources without interruption. Popular social networking platforms like Facebook, YouTube, and TikTok are prime examples of such data sources. To manage this real-time data flow, technologies such as Apache Kafka and Spark Streaming are frequently employed. In this context, our implementation leverages Spark Streaming for data streaming.

**Input Data:** Our primary focus is on processing comments from external sources, in this project we use Reddit. Since this platform takes charge of using the API, we use a dataset on Kaggle which contains Reddit posts that have been crawled.

**Main Model**: The core components of our model involve data pre-processing and detecting indicators of depression in the collected data.

**Output Data:** The output will be a single value indicating whether the post shows potential depression risk.

**Main Process**: Initially, data is pre-processed and normalized to ensure a consistent format. This processed data is then published to Kafka topics, a distributed messaging system that facilitates real-time data streaming. The data within these Kafka topics is subsequently consumed by a trained model, which generates predictions based on the input. These predictions are then organized and presented by Spark Streaming, a real-time processing framework. SparkSQL queries and displays the data, while Kafka Connect integrates the predictions into MongoDB, a document database. This setup allows for the predictions to be queried and utilized in building applications, thereby enabling real-time analytics and data mining.

# 5 Empirical Result

## 5.1 Experimental Setup

We trained our models using Google Colab environment connecting to a T4-GPU runtime. We utilized the models using Pyspark MLLib built-in library. For the settings, we split the dataset randomly with a ratio of 20% for the test set. The metrics used in this project are accuracy, AUC, precision, recall and F1-score.

## 5.2 Results

After preprocessing the data and building the model, we get the experimental results on Logistic Regression, Support Vector Machine models. The evaluation metrics include recall, precision, and F1-score, which are crucial for assessing the model's ability to correctly identify depression-related posts.

| **Class** | **Model** | **Recall** | **Precision** | **F1 score** |
|---|---|---|---|---|
| Depression | Logistic Regression | 0.8962 | 0.9342 | 0.9148 |
| | SVM | **0.9008** | **0.9378** | **0.9190** |
| | Decision Tree | 0.8794 | 0.8714 | 0.8753 |
| Non-depression | Logistic Regression | 0.9359 | 0.8988 | 0.9170 |
| | SVM | **0.9407** | **0.9052** | **0.9226** |
| | Decision Tree | 0.8725 | 0.8804 | 0.8764 |

**Table 1** Evaluation on test set with respect to each label.

| **Model** | **Accuracy** | **AUC** |
|---|---|---|
| Logistic Regression | 91.71% | 0.92 |
| SVM | **91.82**% | **0.92** |
| Decision Tree | 87.34% | 0.88 |

**Table 2** Overall performance metrics

The results table indicates that the Support Vector Machine (SVM) model outperforms Logistic Regression and Decision Tree for classifying both Depression and Non-depression. SVM achieves the highest recall, precision, and F1 scores for both classes. In terms of overall performance, SVM also leads with the highest accuracy and AUC, closely followed by Logistic Regression, while the Decision Tree shows the lowest performance. These results highlight SVM as the most effective model for this classification task.

# 6 Conclusion and Future Work

In this study, we proposed a framework for real-time depression risk detection leveraging big data analytics from social media platforms. By harnessing the vast amount

of user-generated content, our approach aimed to provide timely insights into individuals' mental health statuses. Through a combination of data preprocessing, feature engineering, and machine learning techniques, we successfully developed a predictive model capable of identifying potential signs of depression based on textual cues extracted from social media posts. Our experimental results demonstrated promising performance in detecting depression-related indicators. This suggests that our method can effectively complement traditional diagnostic approaches by offering a scalable and non-invasive means of early detection. Moreover, our study contributes to the growing field of digital mental health by showcasing the feasibility and potential benefits of utilizing big data analytics in real-time mental health monitoring. By automating the detection process, our framework has the potential to assist healthcare professionals in proactively identifying at-risk individuals and providing timely interventions.

While our study lays a solid foundation, future research could explore several avenues for improvement and extension. Enhancing feature engineering with techniques such as sentiment analysis, topic modeling, and context-aware embeddings could capture richer semantic information. Investigating advanced machine learning models like LSTM and Transformer architectures could further enhance prediction accuracy. Integrating multimodal data sources such as images and videos could provide a more holistic view of users' mental states. Developing a scalable real-time monitoring system for continuous data processing would ensure timely detection and response. Addressing ethical concerns regarding data privacy and compliance with regulations is essential for maintaining user trust. Conducting validation studies across diverse populations and collaborating with mental health professionals would validate and refine the clinical applicability of our approach.

# References


[1] Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders (5th Ed.), 5th edn. American Psychiatric Publishing, Washington (2013)

[2] Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: from Keyboard to Clinic, pp. 88–97 (2018)

[3] Smys, S., Raj, J.S.: Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. Journal of trends in Computer Science and Smart technology (TCSST) **3**(01), 24–39 (2021)

[4] Wongkoblap, A., Vadillo, M., Curcin, V.: Depression detection of twitter posters using deep learning with anaphora resolution: Algorithm development and validation. JMIR Mental Health (2021)

[5] Kim, J., Lee, D., Park, E., *et al.*: Machine learning for mental health in social media: bibliometric study. Journal of Medical Internet Research **23**(3), 24870 (2021)

[6] Un Nisa, Q., Muhammad, R.: Towards transfer learning using bert for early detection of self-harm of social media users. Proceedings of the Working Notes of CLEF, 21–4 (2021)

[7] Hong-Phuc Vo, H., Nguyen, H.H., Do, T.-H.: Online hate speech detection on vietnamese social media texts in streaming data. In: Dang, N.H.T., Zhang, Y.-D., Tavares, J.M.R.S., Chen, B.-H. (eds.) Artificial Intelligence in Data and Big Data Processing, pp. 315–325. Springer, Cham (2022)

[8] Doan, L.-A., Nguyen, P.-T., Phan, T.-O., Do, T.-H.: An implementation of large scale hate speech detection system for streaming social media data. In: 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), pp. 155–159 (2022). https://doi.org/10.1109/COMNETSAT56033.2022.9994299

[9] Tran, K.Q., Nguyen, A.T., Hoang, P.G., Luu, C.D., Do, T.-H., Nguyen, K.V.: Vietnamese hate and offensive detection using phobert-cnn and social media streaming data. Neural Computing and Applications **35**, 573–594 (2022)

[10] Luu, S.T., Nguyen, K.V., Nguyen, N.L.-T.: A large-scale dataset for hate speech detection on vietnamese social media texts. In: Fujita, H., Selamat, A., Lin, J.C.-W., Ali, M. (eds.) Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices, pp. 415–426. Springer, Cham (2021)