

A Big Data Analytics System for Predicting Suicidal Ideation in Real-Time Based on Social Media Streaming Data

Mohamed A. Allayla^{1,2}^a and Serkan Ayvaz^{2,3} ^b

¹*Dams and Water Resources Research Center, University of Mosul, Mosul, Iraq*

²*Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey*

³*Centre for Industrial Software, University of Southern Denmark, Sonderborg, Denmark*

Keywords: Big Data, Suicidal Ideation, Apache Spark, Apache Kafka, Social Media.

Abstract: Online social media platforms have recently become integral to our society and daily routines. Every day, users worldwide spend a couple of hours on such platforms, expressing their sentiments and emotional state and contacting each other. Analyzing such huge amounts of data from these platforms can provide a clear insight into public sentiments and help detect their mental status. The early identification of these health condition risks may assist in preventing or reducing the number of suicide ideation and potentially saving people's lives. The traditional techniques have become ineffective in processing such streams and large-scale datasets. Therefore, the paper proposed a new methodology based on a big data architecture to predict suicidal ideation from social media content. The proposed approach provides a practical analysis of social media data in two phases: batch processing and real-time streaming prediction. The batch dataset was collected from the Reddit forum and used for model building and training, while streaming big data was extracted using Twitter streaming API and used for real-time prediction. After the raw data was preprocessed, the extracted features were fed to multiple Apache Spark ML classifiers: NB, LR, LinearSVC, DT, RF, and MLP. We conducted various experiments using various feature-extraction techniques with different testing scenarios. The experimental results of the batch processing phase showed that the features extracted of (Unigram + Bigram) + CV-IDF with MLP classifier provided high performance for classifying suicidal ideation, with an accuracy of 93.47%, and then applied for real-time streaming prediction phase.

1 INTRODUCTION


Suicidal ideation is a serious public health concern. The number of suicidal ideations is increasing at an alarming rate every year. According to a report issued by the World Health Organization (WHO), more than 703,000 people commit suicide annually, which means roughly one person dies every 45 seconds due to suicide. Additionally, for every suicide, 25 attempted suicides and many more had serious thoughts about suicide (Organization, 2022). Suicidal ideation has continuously been linked to emotional states such as depression and hopelessness (Gijzen et al., 2021). The early detection of suicidal ideation may help to prevent many suicide attempts and identify individuals needing psychosocial support.


Traditional methods and programs for suicide prevention are still reactive and require patients to take the initiative to seek medical help. However, many

patients are not highly motivated to receive the necessary support. Due to the anonymity on online social media platforms, it has become an alternative space where people can express their honest feelings or thoughts about their pain or health issues without fear of stigma or revealing their true identity as in face-to-face conversations (Roy et al., 2020b).

This is considered a valuable source for detecting high-risk suicidal ideations instances and uncovering these dangerous intentions before they become irreversible or the sufferers end their lives. Suicidal sufferers may show suicidal intentions online through brief ideas or detailed planning. Social media have been successfully leveraged to assist in detecting physical and mental illnesses more easily (Aldhyani et al., 2022). Therefore, researchers have begun using online postings to detect suicidal ideation manually or with the help of machine learning techniques (Baghdadi et al., 2022). Manual identification of suicidal ideation has become more challenging due to the vast amount of content on social media platforms.

Moreover, social media posts are generated as

^a  <https://orcid.org/0000-0002-6958-1208>

^b  <https://orcid.org/0000-0003-2016-4443>

streaming data in real-time. However, real-time systems require direct input and rapid processing capability to make decisions in a short time (Senthilkumar et al., 2018). Several problems must be addressed before developing a real-time analytics system. The first is to provide a reliable and efficient framework for distributing data without losing accuracy. Most big-data research in healthcare focuses on the technical aspects of big data. Another problem with streaming data is that it involves high-velocity and continuous data generation. Hence, processing such a huge data stream using a traditional system environment in real-time may result in system bottlenecks.

The presented work aimed to build an effective real-time model using a big data analytics system to predict a person's suicidal ideation at an earlier stage based on their social media posts. We focused primarily on a social media platform where people talk about different mental health issues and offer a platform to help. Some notable contributions made by this paper include the following:

- This paper proposed a scalable predictive system that can analyze large volumes and high-velocity streaming data in real-time using “big data” architecture to predict suicidal ideation cases that require special attention.
- We applied various experiments with multiple Apache Spark ML algorithms using three feature extraction: TF-IDF, N-gram, and CountVectorizer, with various combinations and testing scenarios.
- We performed optimization techniques to achieve high prediction accuracy. The proposed system achieved significant performance on both batch and real-time streaming phases of suicidal ideation prediction.

2 LITERATURE REVIEW

Sentiment Analysis has attracted the attention as a research topic in various fields such as financial (Ayvaz and Shiha, 2018), public health (Alamoodi et al., 2021), product reviews (Agarwal et al., 2024), voting behavior (Rita et al., 2023), political (Öztürk and Ayvaz, 2018) and social events (Allayla and Ayvaz, 2023). Although approaches, methods and models vary across domains, it has been observed that sentiment analysis and prediction tasks often produce useful and interesting results. From the perspective of monitoring suicidal ideation and mental state, there are some studies that analyze social media data using natural language processing (NLP) and sentiment

analysis by investigating different aspects (Jain et al., 2019; Sawhney et al., 2018; Desu et al., 2022).

In the study conducted by S. Jain et al., two datasets were used to develop a machine learning-based method for predicting suicidal behaviors depending on the depression stage (Jain et al., 2019). The first dataset was collected by creating a questionnaire from students and parents and then classifying the depression according to five severity-based stages. The XGBoost classifier reported a maximum accuracy of 83.87% in this dataset. The second dataset has been extracted from Twitter. Tweets were classified according to whether the user had depression. They found that the Logistic Regression algorithm exhibited the highest performance and achieved an accuracy of 86.45%.

N. Wang et al. proposed a deep-learning (DL) architecture as well as evaluated three more machine learning (ML) models to analyze the individual content for automatically identifying whether a person will commit suicide within 30 days to 6 months before the attempt (Wang et al., 2021). They created and extracted three handcrafted feature sets to detect suicide risk using the three-phase theory of suicide and earlier work on emotions and pronouns among people who exhibit suicidal thoughts.

Similarly, M. Chatterjee et al. analyzed Twitter platform content and identified the features that can hold signs of suicidal ideation. Multiple ML algorithms were applied, including LR, RF, SVM, and XGBoost, to evaluate the effectiveness of the suggested approach (Chatterjee et al., 2022). The study involved extracting and combining various topics, linguistic, statistical features, and temporal sentiments. The study extracted multiple features from Twitter data, including sentiment analysis, emoticons, statistics, TF-IDF, N-gram, temporal features, and topic-based features (LDA). The empirical findings showed that by employing the Logistic Regression classifier, an accuracy of 87% was registered.

A. E. Aladağ et al. used text mining implemented on post titles and bodies; they built a classification model that differentiated between postings that were suicidal and others that were not suicidal (Aladağ et al., 2018). The utilized features were extracted using various techniques, including TF-IDF, word count, linguistic inquiry, and sentiment analysis of the titles and bodies of the posts. The suicidality of posts was correctly classified using Logistic Regression (LR) and Support Vector Machine (SVM) classifiers. Accuracy and F1 score were obtained as 80% and 92% respectively.

Using data collected from electronic medical records in mental hospitals, Carson et al. built and

evaluated an NLP-based machine learning approach to detect suicidal behaviors and thoughts among young people (Carson et al., 2019).

A. Roy et al. evaluated psychological weight factors, including depression, hopelessness, loneliness, stress, anxiety, burdensomeness, and insomnia (Roy et al., 2020a). Furthermore, the sentiment polarity and Random Forest (RF) algorithm were applied with ten estimated psychological measures for predicting SI within tweets and achieved an 88% AUC score.

On the other hand, V. Desu et al. proposed an approach that utilizes various ML and DL algorithms, such as XGBoost, SVM, and ANN, implemented upon a Spark cluster with multiple nodes to detect individuals who suffer from depression and suicidal thoughts and require urgent assistance or support by analyzing their social media content (Desu et al., 2022). The proposed ANN model provided superior efficacy over all other baseline algorithms and registered the best accuracy rate of 76.80%.

M. J. Vioules et al. developed a novel method that uses Twitter data to identify suicide warning signs in users and detect postings containing suicidal behaviors (Vioules et al., 2018). The key contribution of their method is its ability to detect sudden changes in users' online behavior. To identify these changes, they employed NLP algorithms with a martingale framework to collect behavioral and textual features. The experimental results demonstrated that their text-scoring method could detect warning signs in a text more effectively than standard machine learning classifiers.

W. Jung et al. designed multiple machine learning models and analyzed suicidality using Twitter data. The models were trained using 1097 suicidal and 1097 nonsuicidal tweets (Jung et al., 2021). They explored metadata and text-feature extraction to construct efficient prediction models. They trained the classifier models using Random Forest and Gradient-boosted tree (GBT). The experiments were carried out using multiple features to construct a robust classifier. The model achieved an F1 score of 84.6%.

M. M. Tadesse et al. used NLP techniques to identify the depressive content of users generated on the Reddit social website (Tadesse et al., 2019). The study mainly focused on the deployment and evaluation of several feature extraction approaches, such as LIWC, N-grams, and topic modeling using LDA to achieve the highest performance results. The authors applied several classification algorithms, including LR, SVM, RF, adaptive boost (AB), and multilayer perceptron (MLP), to assess the risk of depression among users. The Multilayer Perceptron (MLP) model showed high effectiveness with the combina-

tion of LIWC, Bigram, and LDA features, which resulted in the best performance for identifying depression with precision 91% with an F1 score of 93%.

N. A. Baghdadi et al. presented a detailed framework for text content classification, specifically for Twitter content (Baghdadi et al., 2022). The trained model was employed to identify the tweets as "Suicide" or "Normal." The dataset contains 14,576 tweets. The dataset was annotated through multiple annotators, and the framework's effectiveness was evaluated using various assessment methods. Valuable understandings were gained through the Weighted Scoring Model (WSM). Both USE and BERT classifier models were also explored. The WSM models registered the highest-weighted sum of 80.20%.

3 PROPOSED METHODOLOGY

Real-time streaming analysis of social media content can provide helpful and up-to-date information on individuals with mental health problems. The current analytics methods that analyze social media content with massive volume offline are not robust and active for supporting real-time decision-making under essential conditions. Thus, these analysis methods must be built to provide effective stream real-time prediction. The methodology comprises two phases: batch processing and real-time streaming prediction.

Our system methodology was built based on four primary components: the input source system, where the system obtains the stream data (Apache Kafka); the stream data processing, where the stream data are processed (Apache Spark Structured Streaming); building the classification algorithms (Apache Spark ML); and the sink node, where the final results are analyzed and visualized (Power BI). We built several Apache Spark ML models using multiple feature extraction techniques. Also, we compared the classification performance of multiple models using various evaluation methods to determine the optimal architecture for predicting suicidal-related posts from real-time Twitter streaming data. Figure 1 provides a clear overview of the proposed methodology and the experimental workflow used in this work.

3.1 Big Data Architecture

This section describes the big data architecture applied in this work. Our proposed methodology was developed to efficiently analyze massive volumes of social media content with high velocity in real-time

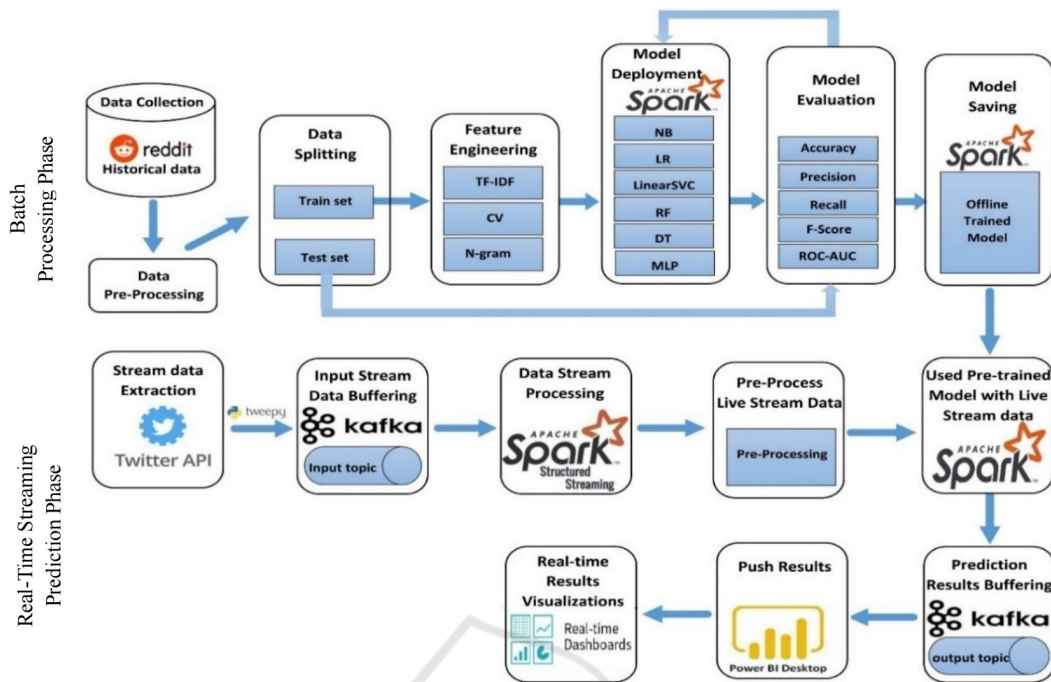


Figure 1: Proposed methodology for predicting suicidal ideation on social media content.

streaming data using a distributed big data environment.

3.1.1 Apache Spark

Apache Spark has been applied in the proposed methodology as a data processing engine. It is an analytics platform that supports batch and stream data processing (Shaikh et al., 2019). Spark is a cluster computing system designed to be open source with various scalable and distributed ML built-in libraries (Junaid et al., 2022). A key feature of Spark is its scalability, which enables building spark clusters with several nodes. It employs a master-slave design consisting of a Driver program that operates as the cluster's master node and a set of executors that act as worker nodes. The core components of Spark include Spark SQL, which is used for structured query language (SQL), and Spark Streaming, which is used to process stream data. Spark Structured Streaming is developed on top of Spark SQL. Structured Streaming manages its execution incrementally and continuously, changing the final output whenever new data streams are received.

3.1.2 Apache Kafka

Apache Kafka has been used to develop real-time prediction pipelines and stream data messaging. Kafka is an open-source and widely powerful ingestion system primarily used in big data applications (Deshpande

and Rao, 2022). It is a low-latency, high-throughput system for managing and transferring massive and high-velocity data in a streaming manner. Producer and consumer APIs are the two primary components of the Kafka architecture. The Producer API allows the system to send data to the Kafka topics. The Consumer API provides access to Kafka topics and processes the data streams in real-time at any time.

3.2 Batch Data Processing Phase

The experiments performed during the batch processing phase aimed to develop and train multiple Spark ML models with different feature extraction and testing scenarios. The model with the highest performance was then applied for real-time streaming data prediction phase. The batch processing phase consists of five primary stages: (i) Data Collection, (ii) Data Cleaning and Preprocessing, (iii) Feature Engineering, (iv) Model Development, and (v) Model Evaluation. The upcoming subsections will provide a detailed description of each phase's steps.

3.2.1 Data Collection

Datasets play an essential role in any text-data analysis. The dataset required for our experiment in the batch processing phase was gathered and acquired from Reddit social media platforms. The primary source of batch datasets is the Kaggle website, a pub-

licly accessible benchmark dataset for various applications (Komati, 2022). The obtained dataset was utilized to train and assess the classifier models during the batch processing phase. The dataset was organized in a separate CSV file format and contained posts from Reddit's platform from subreddits titled "Suicide Watch" and "Teenagers Forum," which were collected using the 'Pushshift' API. The dataset comprised approximately 232,074 posts collected between Dec. 16, 2008, and Jan. 2, 2021, of 116,037 were classified as suicidal and 116,037 as nonsuicidal. We cleaned and preprocessed the dataset to remove duplicate posts, empty rows, and unnecessary columns. After the preprocessing step, the dataset resulted in 232,042 rows, including 116,028 suicidal and 116,014 nonsuicidal instances. For our task, we used only the post content and target columns for the analysis task. Some batch data samples are presented in Table 1.

Table 1: Samples of the Batch Dataset Postings.

class type	postings
Suicide	I need help just help me im crying so hard.
	I have nothing to live for. My life is so bleak.
	Suicidal tics and intrusive anxiety...
Non-suicide	I just got a Russian Hardbass song in my Spotify...
	I wish I could change my name to Seymour...
	My life is not a joke Jokes have meaning.

3.2.2 Data Cleaning and Preprocessing

The text analysis performance can be improved by selecting the proper data preprocessing strategy since the input data collected from social media may contain many non-meaning words or characters, which can increase the complexity of the analysis. Hence, we aimed to prepare and refine the raw data into a suitable and understandable format for each classifier model. Some preprocessing methods are standard for text-analyzing tasks, while others depend on the complexity of data and affect the final result. We preprocessed and prepared the dataset using Natural Language Processing (NLP) techniques before passing it to the feature extraction and training stages.

Filtering Data: In this step, we filtered the obtained tweets to remove duplicate content, URL links ("https://", "http://"), punctuation (e.g., "?", "!"), special symbols (e.g., "\$", "%", "&") and the hashtag ("#"). The filtering step also includes case folding and expanding contractions with their corresponding complete form (i.e., "let's" into "let us", "didn't" into "did not"). This step has a significant effect on improving the effectiveness of the classifiers as it reduces the dataset complexity.

Tokenization: The tokenization step is essential

for any natural language processing (NLP) pipeline. It has a considerable influence on the remaining phases of the pipeline. It breaks down the text data into individual, more meaningful terms, including words, punctuation marks, symbols, and abbreviations, to make data exploration more accessible. The result of this process is known as a token (Vijayarani et al., 2015). These tokens were then used as input data for the processing pipeline.

Stopword Removal: Stop words are the most frequently used terms in the documents. We aimed to reduce the size and complexity of the dataset by removing stop words that do not carry emotional value. So, in this stage, we eliminated most frequently used stopwords, such as pronouns like "she" and "he" articles such as "and," "the," "a," "an," and prepositions like "on," "of," "to," "but," "for." and so on.

Lemmatization: The input data was lemmatized at this step. Lemmatizing removes inflectional ends and returns each word in the dataset to its basic or dictionary form. Lemmatizing requires a comprehensive vocabulary and morphological analysis to lemmatize the words. Among various lemmatization methods, we focused on rule-based approaches using "WordNetLemmatizer." It employs a pre-established set of morphological and syntactic rules to find the lemma of each word within the input text. The use of Lemmatization helps to reduce the dimensionality and the vocabulary size of textual data, which leads to improved performance of analytical techniques.

Dataset Splitting: To train the classification models, it is necessary to split the dataset. Therefore, we divided the entire historical Reddit data into two subsets: Out of 80% of the dataset applied for training data, the remaining 20% were unseen data and applied for testing data. The classification models were trained and optimized using the training data to determine the most accurate features. On the other hand, the testing data (unseen data) was employed to assess the effectiveness of the classification models. Table 2 provides descriptive statistics for the testing and training sets.

Table 2: Training and Testing Dataset Statistics.

Data Subset	Class Type	No. of postings
Train set	Suicide	92726
	Non-suicide	92704
Test set	Suicide	23302
	Non-suicide	23310

3.2.3 Feature Engineering

Once a clean data corpus was generated, the data corpus was processed by the different feature engineer-

ing methods. Our goal was to find the optimal features that provide the highest classification performance, reduce the complexity, and speed up the data transformation. In this task, we used three feature engineering techniques to obtain and extract the dataset's essential features, including N-gram, TF-IDF, and CountVectorizer (CV) with multiple combinations.

N-gram is a feature extraction method identifying N successive word groups within a text (Haviana and Poetro, 2022). This method is widely used as a feature extraction and analysis tool in NLP and text mining. It involves converting the input data into a series of n separate tokens. In our work, the most important features are represented using Unigrams (single words) and Bi-grams (two words have different meanings when combined) with the help of the PySpark library. Also, we assigned high importance to N-grams that appear more than four times in the document.

TF-IDF is a statistical method to extract relevant features from textual data input. TF-IDF builds a vector matrix to demonstrate a word's importance in the document. A word with fewer occurrences in a document is more appropriate for classification. TF-IDF provides a lower score for the most frequent terms and a higher score for lower-frequency terms in a document (Shang and Underwood, 2021) (Vijaya Prakash, 2022). The Spark ML API provides two methods for calculating term frequencies: HashingTF and CountVectorizer (CV). TF-IDF is calculated using the equations 1, 2 and 3 as below.

$$TF(t) = \frac{\text{No. of times term } t \text{ appears in a document}}{\text{Total No. of terms in a document}} \quad (1)$$

$$IDF(t) = \log \left(\frac{\text{Total documents}}{\text{No. of documents containing the term } t} \right) \quad (2)$$

$$TF_IDF(t) = TF(t) \times IDF(t) \quad (3)$$

CountVectorizer (CV) is a basic method for tokenizing data and generating a numerically-representative wordlist (Brownlee, 2017). It builds several columns depending on the occurrence of a unique word in the vocabulary. These columns represent each row by replacing words with their frequencies. CV can be employed when a prior dictionary is unavailable to extract the vocabulary and build the required dictionary (Mehmood et al., 2018). As part of this study, we conducted the experiments using the following combinations of feature extraction methods: Unigram + TF-IDF, Unigram + CV-IDF, Bigram + CV-IDF, (Unigram + Bigram) + CV-IDF

3.2.4 Models Development

In our proposed methodology, we built the classification models using multiple Spark ML algorithms, namely Naïve Bayes (NB), Logistic Regression (LR), Linear Support Vector Classifier (LinearSVC), Decision Tree (DT), Random Forest (RF), and Multilayer Perceptron (MLP) classifiers. The classifier models were trained and tested with various parameter and feature extraction combinations until the best performance values were achieved.

Naïve Bayes Classifier (NB) is a well-known machine learning classification algorithm based on supervised learning. The NB classifier implies that the attributes are independent of each other and that the presence or absence of one attribute does not affect the other attributes. The Naïve Bayes algorithm builds based on Bayes' theorem (Reddy et al., 2022). The NB classifier is often used and ideal for text classification challenges due to its simplicity and speed (Goel et al., 2016).

Logistic Regression Classifier (LR) algorithm is commonly employed for classifying problems and belongs to the generalized linear model category. LR can help calculate and predict the likelihood of allocating a new sample to a particular category for binary or multiclass classification tasks. The algorithm performs well on linearly separable datasets and can be applied to determine the correlations within dataset attributes.

Linear Support Vector Classifier (LinearSVC) is a standard algorithm often used for large-scale classification tasks. Linear SVC is a non-probabilistic classification model that needs an extensive training set. It uses a hyperplane that optimally splits the classes represented in a high-dimensional field space. LinearSVC is widely known for its practical abilities, mainly in dealing with real-world data, which include a solid theoretical basis and insensitivity to high-dimensional data.

Decision Tree Classifier (DT) is a common machine-learning method categorized as a non-parametric supervised algorithm (Jena et al., 2022). It is a hierarchical model designed as a tree structure. Every interior node holds at least one child, representing the evaluation of an input feature or variable. Based on the results of a decision test, the branching procedure will repeat itself, directing the corresponding child node along the suitable path, and this process continues until the last leaf node. The optimal tree is the shortest tree that can correctly categorize all data points and has the fewest splits.

Random Forest Classifier (RF) is a popular and widely applied ML method that may be utilized or adopted for both classification and regression pur-

poses. It was introduced by L. Breiman (Breiman, 2001). RF algorithm decreases the prediction variance a decision tree generates and improves its performance. For this purpose, many decision trees were merged using a bagging aggregation technique. RF learns in parallel from numerous decision trees made at random, trained on different data sets, and uses various features to get at its individual decisions.

Multilayer Perceptron Classifier (MLP) is a form of feedforward neural network. MLP employs backpropagation, a supervised learning approach. MLP includes three sets of nodes: the first set is input-layer neurons, the second set is hidden-layer neurons, and the last set is called the output-layer neurons, which represent the final results of the system. Neurons in a perceptron require an activation function that applies a threshold, such as a sigmoid or ReLU.

3.2.5 Models Evaluation and Metrics

The performance of the proposed architecture was evaluated using various assessment methods, including Accuracy (ACC.), Precision (PRE.), Recall (REC.), F1-scores (F1), and the ROC-AUC. Furthermore, the k-fold Cross-Validation approach was employed to ensure the models fit properly without overfitting and underfitting issues. Each classifier was evaluated by calculating the average accuracy of the 10-fold cross-validation to achieve a better model performance.

3.3 Real-Time Streaming Prediction Phase

Our primary aim of the real-time Streaming prediction phase is to build a framework methodology to analyze the high velocity of streaming data arriving each second in real-time. Our methodology has four main components: data collection, data ingestion system, stream processing, and results visualization, as shown in Figure 1. To check the proposed architecture's ability to identify suicidal ideation in real-time scenarios. We used Twitter API to retrieve real-time streaming tweets from Twitter. Twitter Streaming API ¹ is the basic method for accessing Twitter data. Twitter API allows access to real-time with a limited set of approximately 1% of all tweets. Furthermore, Tweepy ² allows us to search tweets using hashtags, keywords, trends, geolocation, or timelines. Our methodology used keyword searches for retrieval of the tweets. We employed rules to

retrieve only English tweets and filtered all duplicate tweets created by retweets. A total stream of 764 tweets was retrieved using multiple keywords related to suicidal ideation, including “feel,” “want to die,” and “kill myself”. The retrieved tweets included multiple columns, including tweet content, retweet counts, and usernames. Only the “tweet” column was used for our work, while the other columns were not utilized and were removed from the collected data. Apache Kafka was utilized to develop real-time pipelines and stream data ingestion. The key benefit of Kafka is its ability to handle huge amounts of real-time data within low latency, and it is fault-tolerant and scalable to ingest large data streams. We created an input topic, “Source-tweets,” on the Kafka system.

The collected tweets were then ingested as data streams into the Apache Kafka input topic. Spark Structured Streaming consumes stream tweets from the Kafka topic in real-time into the unbounded table. We implemented several preprocessing steps to refine the tweets' stream effectively. These steps involve removing irrelevant information, reducing the noise, and extracting appropriate stream data. After preprocessing and cleaning the streaming tweets, we generated a feature vector and fed it into the highest accurate model previously developed and trained in the batch processing phase to predict suicidal ideation in real time. The prediction results were then pushed and buffered in a Kafka output “Predicted-tweets” topic before being consumed by the Power BI application to visualize the final prediction results in real time.

4 EXPERIMENTAL SETUP AND PERFORMANCE ANALYSIS

4.1 Experimental Setup

The proposed ApacheSpark-based architecture was implemented using the “PySpark” library to build the classification algorithms: NB, LR, LinearSVC, DT, RF, and MLP algorithms. Apache Spark Cluster was installed on a laptop with 64 GB of RAM, a 1 TB SSD disk drive, and an Intel Core i7 CPU (14 cores, 20 logical processors). In addition, we integrated multiple API libraries for implementation. ML library of Apache Spark was used to develop classification algorithms. Apache Kafka version of “2.0.2” was deployed as an input system for ingesting data streams from Twitter. Tweepy version of “4.10.0” for connecting to the Twitter API. Spark Structured Streaming was applied for receiving and processing stream tweets from Kafka topics—Power BI application for

¹<https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data>

²<https://docs.tweepy.org/en/latest/index.html>

Visualizing the real-time streaming prediction results.

4.2 Exploratory Data Analysis

We checked the most frequently used terms in suicide and nonsuicide posts. It was observed most suicidal postings contained the words “want,” “friend,” and “think.” On the other hand, the most frequently non-suicidal posts of repeated words, included the words “though,” “feel” and “die.”

4.3 Evaluation of Batch Processing Phase

This section presents and discusses the experimental applied in the batch processing phase for identifying suicidal ideation in individuals based on their social media posts. Our primary objective was to determine the most efficient model with the highest performance to adopt for real-time streaming prediction phase.

We used multiple Apache Spark ML algorithms in this work, including Naïve Bayes (NB), Logistic Regression (LR), Linear Support Vector classifier (LinearSVC), Decision Tree (DT), Random Forest (RF), and Multilayer Perceptron (MLP). The algorithms were trained and evaluated using data from the Reddit forum, using three different strategies for feature extraction: TF-IDF, N-gram, and the CountVectorizer technique. Multiple combinations of these feature extraction methods were implemented to extract the essential features.

A hyperparameter tuning strategy was adopted to detect the optimal parameter tune for each model configuration. Two methods are commonly employed for Hyperparameter tuning: Random search and Grid search. In this work, we utilized the Grid search as a hyperparameter technique in the experiments. The Grid search hyperparameter tuning process aims to find the optimal parameters and most suitable values for each classifier to enhance the overall performance.

Furthermore, we made use of 10-fold Cross-validation, which is a widespread technique and reliable method for minimizing overfitting, enhancing the validity and reliability of the classification models, and balancing the bias and variance values. With the 10-fold cross-validation strategy, the given data were subdivided randomly into ten subsets of the same size; one subset was used for testing purposes, while the other nine subsets were used for the training process. Cross-validation was executed ten times, with each of the ten subsets used as validation only once. To get a final estimate, the data were averaged across ten folds. Table 3 and Figures 2, 3, 4, and 5 illustrate the experimental results and comparative

performance assessment of multiple Spark ML classifiers using a binary classification evaluator.

Table 3: Performance Comparison of Classification Algorithms on testing dataset.

Model	Feature Extraction Combination	ACC.	PRE.	REC.	F1.	AUC.
NB	Unigram+TF-IDF	88.02	88.66	88.02	87.97	95.41
	Unigram+CV-IDF	89.49	90.21	89.49	89.44	96.41
	Bigram+CV-IDF	75.86	81.07	75.86	74.81	94.60
	(Unigram + Bigram) + CV-IDF	90.36	91.09	90.36	90.32	96.97
LR	Unigram+TF-IDF	91.40	91.64	91.40	91.38	97.17
	Unigram+CV-IDF	91.98	92.20	91.98	91.96	97.55
	Bigram+CV-IDF	87.56	88.50	87.56	87.48	94.54
	(Unigram + Bigram) + CV-IDF	92.14	92.36	92.13	92.12	97.67
LinearSVC	Unigram+TF-IDF	90.58	91.01	90.58	90.56	96.69
	Unigram+CV-IDF	91.59	92.01	91.59	91.57	97.45
	Bigram+CV-IDF	86.36	88.05	86.36	86.21	94.62
	(Unigram + Bigram) + CV-IDF	90.90	91.54	90.89	90.86	97.59
DT	Unigram+TF-IDF	86.05	86.02	86.05	86.03	87.70
	Unigram+CV-IDF	86.46	86.60	86.46	86.44	87.81
	Bigram+CV-IDF	72.92	77.87	72.92	71.66	73.82
	(Unigram + Bigram) + CV-IDF	86.46	86.60	86.45	86.44	87.81
RF	Unigram+TF-IDF	86.25	86.22	86.25	86.22	93.71
	Unigram+CV-IDF	86.47	86.80	86.47	86.44	93.96
	Bigram+CV-IDF	79.77	82.31	79.77	79.37	88.03
	(Unigram + Bigram) + CV-IDF	85.86	86.27	85.86	85.82	93.52
MLP	Unigram+TF-IDF	92.66	92.66	92.66	92.66	97.70
	Unigram+CV-IDF	93.33	93.33	93.33	93.33	97.99
	Bigram+CV-IDF	88.84	88.93	88.84	88.84	94.48
	(Unigram + Bigram) + CV-IDF	93.47	93.47	93.47	93.47	98.12

From all experimental results, we found that the Multilayer Perceptron (MLP) classifier outperformed the other classification algorithms and achieved a greater accuracy rate of 93.47% and an AUC score of 98.12%. The logistic Regression (LR) classifier also performed well but somewhat less than the Multilayer Perceptron (MLP) classifier and achieved the second-greatest performance, with an accuracy rate of 92.14%.

In addition, the results showed no significant performance difference between the Linear Support Vector classifier (LinearSVC) and Naïve Bayes (NB). Unexpectedly, from the experimental results, we found that Decision Tree (DT) and Random Forest (RF) underperformed other classifiers utilized in this work despite their efficacy in numerous machine-learning scenarios.

Based on all experimental results, we observed that most classifier models that used N-gram + CV-IDF as their feature extraction approach performed better than those that used the N-gram +TF-IDF feature approach. The classifier algorithms were also evaluated using another metric known as the Area-Under-Curve (AUC). The metric provides a value ranging from 0 to 1. A value closer to 1 indicated better classification results. Figures 6, 7, 8 and 9 display the AUC comparison of all the classification methods.

4.4 Evaluation of Real-Time Streaming Prediction Phase

The real-time streaming prediction phase used the classifier models already developed and pre-trained

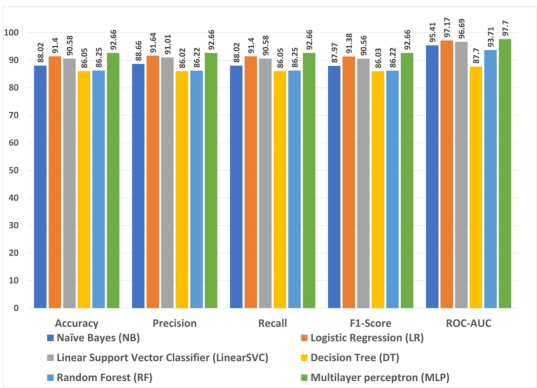


Figure 2: Comparison of performance results of all classification algorithms with Unigram +TF-IDF features.

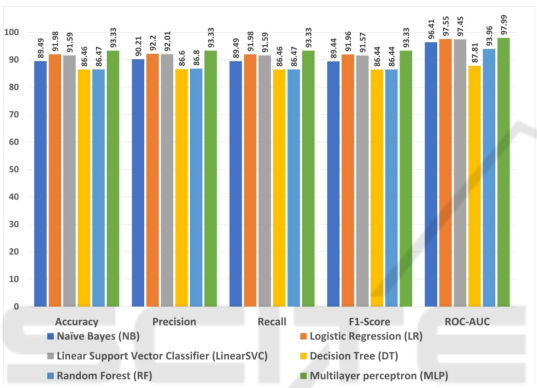


Figure 3: Comparison of performance results of all classification algorithms with Unigram + CV-IDF features.

during the batch processing phase to evaluate their ability to predict suicidal ideation from Twitter streaming data. After designing and assessing the classifier models in the batch processing phase, the classifier with the greatest performance, as in our experiment, MLP with (Unigram + Bigram) + CV-IDF feature extraction combination, was applied for predicting Twitter suicidal ideation-related content in real-time.

We collected streaming tweets using Twitter API with multiple keywords, including “feel,” “want to die,” and “kill myself”, which were then pushed into the Apache Kafka input topic. These streams of tweets were consumed by Apache Spark Structure Streaming from the Kafka input topic, which was then preprocessed as a data stream and used to generate a feature vector. The best pre-trained model developed in the batch processing phase was deployed in the framework for real-time prediction. This model was then used to analyze the pre-processed stream of tweets and predict whether these tweets were suicidal content or normal content in real time.

The prediction results were then pushed to a Kafka

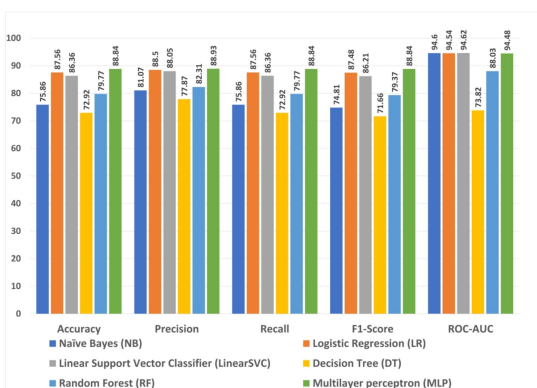


Figure 4: Comparison of performance results of all classification algorithms with Bigram + CV-IDF features.

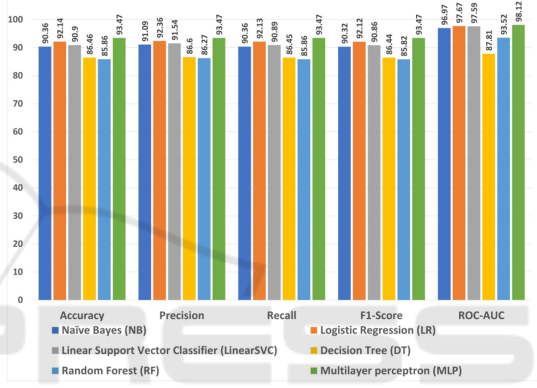


Figure 5: Comparison of performance results of all classification algorithms with (Unigram + Bigram) + CV-IDF features.

output topic for buffering and then consumed from the Power BI application to visualize the prediction results in real-time. In our work, a total of 764 tweets as a data stream were collected to examine the prediction ability in the real-time streaming prediction phase. The real-time streaming prediction phase results indicated that (9.29%) of the tweets were predicted as suicide, whereas (90.71%) were non-suicide.

5 DISCUSSIONS

In this study, we proposed a big data approach to predict suicidal ideation based on data collected from social media platforms. The proposed methodology comprised two phases on batch processing and streaming predictions in real-time. The systems utilized six Spark ML algorithms to build the classification model and compared the performances of the models. In the streaming data pipeline, live streams of a tweet are collected from Twitter using the keywords “feel”, “want to die” and “kill myself” and then sent

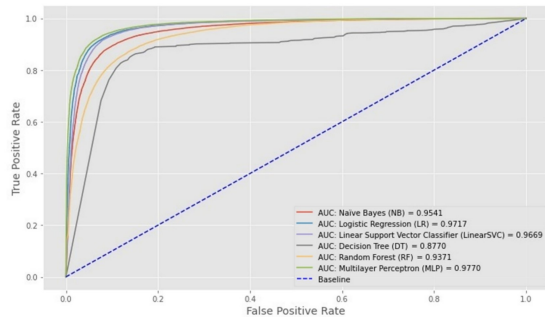


Figure 6: Comparison of ROC-AUC of all classification algorithms with Unigram + TF-IDF features method.

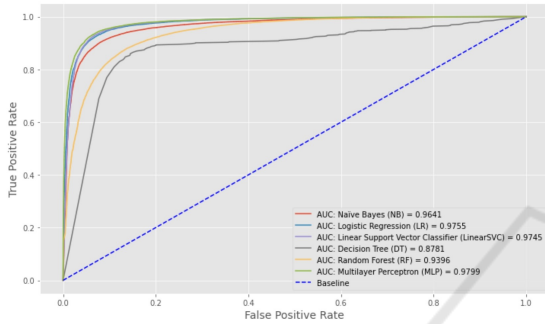


Figure 7: Comparison of ROC-AUC of all classification algorithms with Unigram + CV-IDF features.

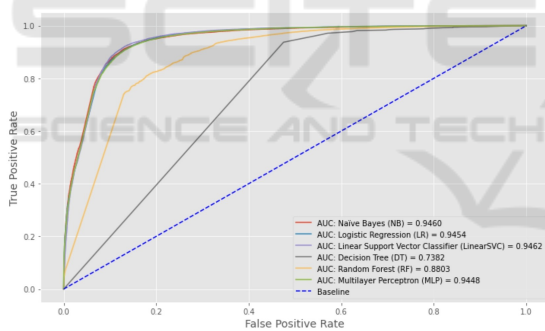


Figure 8: Comparison of ROC-AUC of all classification algorithms with Bigram + CV-IDF features.

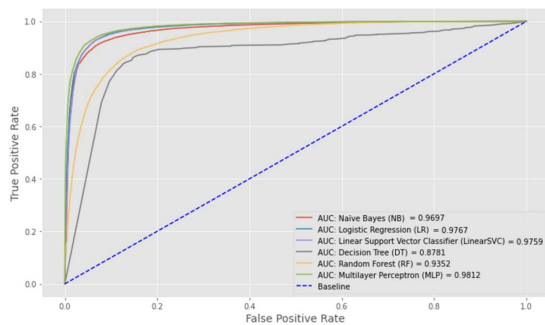


Figure 9: Comparison of ROC-AUC of all classification algorithms with (Unigram + Bigram) + CV-IDF feature.

the collected data to the Kafka topic. Spark Structured Streaming receives the stream data from the Kafka topic, extracts the optimal feature, and then sends batches of preprocessed data to the real-time streaming prediction model to predict whether the tweet contains indications of suicidal ideation.

This work used three feature extraction methods, including TF-IDF, N-gram, and Count Vectorizer, with different combination scenarios to extract the optimal features from the input data. The experimental results of six classification models showed that the MLP classifier had the highest accuracy value of 93.47% with the features extracted using (Unigram + Bigram) + CV-IDF feature extraction scenario. At the same time, a high accuracy of 93.33% was obtained from the MLP classifier with features extracted using (Unigram + CV-IDF). In addition, MLP provided the best accuracy of 92.66% using (Unigram + TF-IDF).

Comparing our experimental results with related work, we noticed that the highest accuracy obtained from the MLP classifier was higher than the accuracies of XGBoost and logistic regression of 83.87% and 86.45%, respectively, obtained by Jain et al. (Jain et al., 2019). Moreover, our methodology outperformed the best performing models obtained by (Aladağ et al., 2018) and (Desu et al., 2022). In the study conducted by Aladağ et al., the accuracy and F1 score rates were reported as 80% and 92%, respectively, and the accuracy rate of the model by Desu et al. was found to be 76.80%. In addition, our proposed approach outperformed the Naïve Bayes model developed by Birjali et al., which achieved 87.50% Precision value, 78.8% Recall value and 82.9% F1. value (Birjali et al., 2017). Therefore, we adopted the MLP classifier with (Unigram + Bigram) + CV-IDF feature combination scenario to predict suicidal ideation in the second phase of real-time streaming prediction using Twitter streaming data.

That being said, further improvements can be made to extend this study. The first improvement can be achieved by increasing the number of features of the textual data using additional data such as emoticons, special characters, and symbols to extract optimal features and reduce the misclassification results. Moreover, the dataset can be expanded by gathering additional textual data from other social media platforms to make our data more representative and varied.

6 CONCLUSION AND FUTURE WORK

In conclusion, this paper proposed a real-time streaming prediction system for suicidal ideation prediction of users' posts on social networks using a big data analytics environment—the work methodology analysis of social media content with two-phase batch processing and real time streaming prediction. Our system applied two types of datasets. Reddit's historical big data are used for model building, while Twitter streams big data have been used for real-time streaming prediction.

Our proposed methodology for building binary classification models was evaluated using various assessment metrics and showed high levels of accuracy and AUC scores with stable Recall and Precision. The experimental results of the batch processing phase revealed that the MLP classifier achieved the highest classification accuracy of 93.47% on an unseen dataset and was used for the real-time streaming prediction phase.

According to the results of various testing scenarios, we can conclude that the features retrieved from stream data could accurately determine the suicidal ideation of users in real time. The developed system might also assist public health professionals with limited resources in determining and controlling suicidal ideation and preparing preventative steps to save lives. Multiple languages, such as Turkish and Arabic, can be added for future work. To deal with such datasets, which require sequential information and local feature engineering, we may use Ensemble LSTM and CNN models for better performance. We also plan to develop a web or mobile interface as a text-analysis tool to detect the individual's health status.

REFERENCES

- Agarwal, G., Dinkar, S. K., and Agarwal, A. (2024). Binarized spiking neural networks optimized with nomadic people optimization-based sentiment analysis for social product recommendation. *Knowledge and Information Systems*, 66(2):933–958.
- Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O., and Bingol, H. O. (2018). Detecting suicidal ideation on forums: proof-of-concept study. *Journal of medical Internet research*, 20(6):e9840.
- Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., et al. (2021). Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review. *Expert systems with applications*, 167:114155.
- Aldhyani, T. H., Alsubari, S. N., Alshebami, A. S., Alkah-tani, H., and Ahmed, Z. A. (2022). Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International journal of environmental research and public health*, 19(19):12635.
- Allayla, M. A. and Ayvaz, S. (2023). A Hybrid and Scalable Sentiment Analysis Framework: Case of Russo-Ukrainian War. In *2023 3rd International Scientific Conference of Engineering Sciences (ISCES)*, pages 13–18. IEEE.
- Ayvaz, S. and Shiha, M. O. (2018). A scalable streaming big data architecture for real-time sentiment analysis. In *Proceedings of the 2018 2nd international conference on cloud and big data computing*, pages 47–51.
- Baghdadi, N. A., Malki, A., Balaha, H. M., AbdulAzeem, Y., Badawy, M., and Elhosseini, M. (2022). An optimized deep learning approach for suicide detection through Arabic tweets. *PeerJ Computer Science*, 8:e1070.
- Birjali, M., Beni-Hssane, A., and Erritali, M. (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, 113:65–72.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.
- Brownlee, J. (2017). *Deep learning for natural language processing: develop deep learning models for your natural language problems*. Machine Learning Mastery.
- Carson, N. J., Mullin, B., Sanchez, M. J., Lu, F., Yang, K., Menezes, M., and Cook, B. L. (2019). Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PloS one*, 14(2):e0211116.
- Chatterjee, M., Kumar, P., Samanta, P., and Sarkar, D. (2022). Suicide ideation detection from online social media: A multi-modal feature based technique. *International Journal of Information Management Data Insights*, 2(2):100103.
- Deshpande, K. and Rao, M. (2022). An Open-Source Framework Unifying Stream and Batch Processing. In *Inventive Computation and Information Technologies*, pages 607–630. Springer.
- Desu, V., Komati, N., Lingamaneni, S., and Shaik, F. (2022). Suicide and Depression Detection in Social Media Forums. In *Smart Intelligent Computing and Applications, Volume 2: Proceedings of Fifth International Conference on Smart Computing and Informatics (SCI 2021)*, pages 263–270. Springer.
- Gijzen, M. W., Rasing, S. P., Creemers, D. H., Smit, F., Engels, R. C., and De Beurs, D. (2021). Suicide ideation as a symptom of adolescent depression. a network analysis. *Journal of Affective Disorders*, 278:68–77.
- Goel, A., Gautam, J., and Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 257–261. IEEE.

- Haviana, S. F. C. and Poetro, B. S. W. (2022). Deep learning model for sentiment analysis on short informal texts. *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, 10(1):82–89.
- Jain, S., Narayan, S. P., Dewang, R. K., Bhartiya, U., Meena, N., and Kumar, V. (2019). A machine learning based depression analysis and suicidal ideation detection system using questionnaires and twitter. In *2019 IEEE Students Conference on Engineering and Systems (SCES)*, pages 1–6. IEEE.
- Jena, M., Behera, R. K., and Dehuri, S. (2022). Hybrid Decision Tree for Machine Learning: A Big Data Perspective. In *Advances in Machine Learning for Big Data Analysis*, pages 223–239. Springer.
- Junaid, M., Ali, S., Siddiqui, I. F., Nam, C., Qureshi, N. M. F., Kim, J., and Shin, D. R. (2022). Performance Evaluation of Data-driven Intelligent Algorithms for Big data Ecosystem. *Wireless Personal Communications*, 126(3):2403–2423.
- Jung, W., Kim, D., Nam, S., and Zhu, Y. (2021). Suicidal-ity detection on social media using metadata and text feature extraction and machine learning. *Archives of suicide research*, pages 1–16.
- Komati, N. (2022). Suicide and depression detection. <https://www.kaggle.com/datasets/nikhileswarkomati-suicide-watch>.
- Mehmood, R., Bhaduri, B., Katib, I., and Chlamtac, I. (2018). *Smart Societies, Infrastructure, Technologies and Applications: First International Conference, SCITA 2017, Jeddah, Saudi Arabia, November 27–29, 2017, Proceedings*, volume 224. Springer.
- Organization, W. H. (2022). World Suicide Prevention Day 2022. <https://www.who.int/news-room/events/detail/2022/09/10/default-calendar/world-suicide-prevention-day-2022>.
- Öztürk, N. and Ayvaz, S. (2018). Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147.
- Reddy, E. M. K., Gurralla, A., Hasitha, V. B., and Kumar, K. V. R. (2022). Introduction to Naive Bayes and a Review on Its Subtypes with Applications. *Bayesian Reasoning and Gaussian Processes for Machine Learning Applications*, pages 1–14.
- Rita, P., António, N., and Afonso, A. P. (2023). Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period. *Social Network Analysis and Mining*, 13(1):46.
- Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., and Kaminsky, Z. A. (2020a). A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*, 3(1):1–12.
- Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., and Kaminsky, Z. A. (2020b). A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*, 3(1):1–12.
- Sawhney, R., Manchanda, P., Singh, R., and Aggarwal, S. (2018). A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98.
- Senthilkumar, S. A., Rai, B. K., Meshram, A. A., Gunasekaran, A., and Chandrakumarmangalam, S. (2018). Big data in healthcare management: a review of literature. *American Journal of Theoretical and Applied Business*, 4(2):57–69.
- Shaikh, E., Mohiuddin, I., Alufaisan, Y., and Nahvi, I. (2019). Apache spark: A big data processing engine. In *2019 2nd IEEE Middle East and North Africa COMMUNICATIONS Conference (MENACOMM)*, pages 1–6. IEEE.
- Shang, W. and Underwood, T. (2021). Improving Measures of Text Reuse in English Poetry: A TF-IDF Based Method. In *International Conference on Information*, pages 469–477. Springer.
- Tadesse, M. M., Lin, H., Xu, B., and Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.
- Vijaya Prakash, R. (2022). Machine Learning Approach To Forecast the Word in Social Media. *Social Network Analysis: Theory and Applications*, pages 133–147.
- Vijayarani, S., Ilamathi, M. J., and Nithya, M. (2015). Pre-processing techniques for text mining-an overview. *International Journal of Computer Science and Communication Networks*, 5(1):7–16.
- Vioules, M. J., Moulahi, B., Azé, J., and Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, 62(1):1–7.
- Wang, N., Luo, F., Shvtare, Y., Badal, V. D., Subbalakshmi, K. P., Chandramouli, R., and Lee, E. (2021). Learning models for suicide prediction from social media posts. *arXiv preprint arXiv:2105.03315*.