Arabian Journal for Science and Engineering (2024) 49:12867–12885
https://doi.org/10.1007/s13369-024-08845-6

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Robust Drug Use Detection on X: Ensemble Method with a Transformer Approach

Reem Al-Ghannam[1] · Mourad Ykhlef[1] · Hmood Al-Dossari[1]

Received: 10 October 2023 / Accepted: 5 February 2024 / Published online: 14 March 2024

Abstract

There is a growing trend for groups associated with drug use to exploit social media platforms to propagate content that poses a risk to the population, especially those susceptible to drug use and addiction. Detecting drug-related social media content has become important for governments, technology companies, and those responsible for enforcing laws against proscribed drugs. Their efforts have led to the development of various techniques for identifying and efficiently removing drug-related content, as well as for blocking network access for those who create it. This study introduces a manually annotated Twitter dataset consisting of 112,057 tweets from 2008 to 2022, compiled for use in detecting associations connected with drug use. Working in groups, expert annotators classified tweets as either related or unrelated to drug use. The dataset was subjected to exploratory data analysis to identify its defining features. Several classification algorithms, including support vector machines, XGBoost, random forest, Naive Bayes, LSTM, and BERT, were used in experiments with this dataset. Among the baseline models, BERT with textual features achieved the highest $F1$-score, at 0.9044. However, this performance was surpassed when the BERT base model and its textual features were concatenated with a deep neural network model, incorporating numerical and categorical features in the ensemble method, achieving an $F1$-score of 0.9112. The Twitter dataset used in this study was made publicly available to promote further research and enhance the accuracy of the online classification of English-language drug-related content.

Keywords Deep learning · Transformer · Ensemble method · Drug use detection · Exploratory data analysis · Natural language processing

1 Introduction

Social media platforms and online social networks (OSNs) have allowed remote and instantaneous communication and social interaction globally. Unfortunately, these platforms also have disadvantages [1]. In early 2023, it was estimated that social media platforms had approximately 4.8 billion users, of whom 150 million were new users who had joined them within the previous 12 months. The principal social media platforms are X (formerly Twitter), Facebook, and YouTube, which are used by nearly 60% of the world's population, on average, for 2 h and 24 min every day [2]. These networks have become integral to people's lives, providing a variety of benefits. Nevertheless, the large size of the user bases now supported by social media networks has made the monitoring of user-generated content challenging, while networks simultaneously need to maintain a competitive edge and innovate to keep their users engaged.

It is well documented that social media networks create opportunities for undesirable or illegal activities. For example, some media users employ the features of these platforms (e.g., direct messaging functions) to facilitate criminal acts, including coordinating the supply of drugs. A strong positive correlation has also been widely reported between excessive use of social media and adolescent drug addiction [3]. Other findings suggest when alcohol-related content is available in social media, especially when it is shared among young people, it can increase the general propensity for alcohol use, as well as drug use and consequent addiction. Reflecting these

✉ Reem Al-Ghannam
r.g.alghannam@gmail.com

Mourad Ykhlef
ykhlef@ksu.edu.sa

Hmood Al-Dossari
hzaldossari@ksu.edu.sa

[1] College of Computer and Information Sciences, King Saud University, 11584 Riyadh, Saudi Arabia

trends, it has been reported that eight social media users have been arrested and charged by the General Directorate of Narcotics Control (GDNC) in Saudi Arabia after being found using OSNs to facilitate drug smuggling. It was notable that these individuals were all men aged 30–39 years, living in different parts of Saudi Arabia, and together, they were running seven Twitter accounts under female profiles to attract younger men and ensnare them in their drug network [4].

The World Health Organization (WHO) and the United Nations Office on Drugs and Crime (UNODC) both operate under constitutional mandates to address issues related to drug use and addiction. At the heart of the WHO's attempts to reduce the demand for drugs are its initiatives in prevention of drug use and management of drug use disorders. They achieve these objectives by collecting, analyzing, and disseminating evidence-based policies on the issues of preventing access to drugs, early intervention when addiction is suspected, and the treatment and rehabilitation of addicts. Additionally, the WHO lends support to monitoring endeavors and provides pertinent information and technical assistance to alleviate the impact of drug use across all levels [5]. The operations conducted by Saudi Arabia's GDNC extend beyond the physical realm, encompassing illicit narcotic trade among real-world entities and efforts to counter drug-related activities online. With the rapid emergence of social media platforms, those involved in illegal narcotic markets have promptly seized the opportunity to appeal to and interact with young individuals [6].

This background reflects the significance of the proposed research, which aims to explore the application of machine learning (ML) and natural language processing (NLP) techniques to accurately detect drug-related content on social media platforms [7]. The literature review in Sect. 2 shows that a suitable application based on artificial intelligence (AI) was designed using machine learning (ML) techniques. However, it is widely known that this problem is complex, and several researchers have attempted to advance knowledge in this area.

In January 2023, Twitter had 556 million active users [8], providing researchers with a significant source of data. Twitter (rebranded as X in September 2023) serves as a real-time platform that is publicly accessible and whose users span a range of social backgrounds, including international organizations, celebrities, and ordinary individuals. Starting in November 2017, Twitter has allowed users to post concise text messages, limited to 280 characters, known as tweets, with which other users can interact. Detecting references to drug use on Twitter is challenging because these text messages are concise, making it difficult to establish a contextual meaning.

A review of existing scientific literature [9–17] revealed that few publicly available datasets have been specifically designed for studying the use of Twitter in disseminating information about drugs. Most datasets are derived from repositories [9–11], open government sources [12, 13], research institutions [14–16], and digital communities [17] that serve a diverse range of interested people. To date, only a few datasets related to Twitter have been published. Two datasets were made available in 2017 [18, 19], one incorporating 267,215 posts and the other 688,757 posts, both with a focus on drug-related activities. In addition, the dataset published in 2021 by Lokala et al. [20] contained 9888 annotated tweets associated with a lexicon linking cannabis to depression. In 2020, a dataset [21] comprising tweets featuring 157 keywords associated with drug use was published. However, these keywords are extensively used; therefore, the dataset includes significant quantities of distracting and irrelevant content such as joy, sarcasm, and news items. The 3,696,150 rows of the dataset included only 2661 tweets that were manually labeled. The authors attempted to improve the dataset and make the convolutional neural network (CNN)-based classifier more accurate through synthetic data generation. However, it must be noted that synthetic data might fail to reproduce natural language patterns and varied expressions occurring in real tweets. Our analysis revealed a shortage of datasets well-suited for the detection of online drug references. Moreover, some existing datasets are incomplete and biased, stemming from ambiguous or subpar annotation processes. Consequently, it is essential that additional data are gathered and annotated to ensure that research on online drug-related content continues to progress.

Machine learning techniques have demonstrated high effectiveness in distinguishing between drug- and non-drug-related content. Several methods have been widely used for this purpose, including random forest (RF) [22], support vector machines (SVM) [23], and long short-term memory (LSTM) networks [22]. These methods have proven to be moderately successful, achieving 85% accuracy in identifying drug-related tweets; however, their margin of error still limits their practical usefulness. Thus, there are clear opportunities for further innovation, and the proposed research aims to refine the data quality and add more features to maximize the capability of the framework to improve the performance achieved by machine learning models. Specifically, this study proposes a novel ensemble method that incorporates a transformer model using multilevel features to bridge the existing gaps.

A new dataset has been created for detecting online drug-related content in English. This dataset consists of 112,057 labeled tweets that were expertly annotated and validated using a range of techniques to guarantee data quality. An exploratory data analysis was conducted to gain insights into the proposed dataset for the recognition of online drug-related content, and a range of classification models were introduced for this purpose. For greater accuracy, N-gram features were assessed together with various feature sets.

This study presents an innovative approach for the enhanced detection of drug-related content on social media. Unlike previous studies, our method focuses on an ensemble approach that combines the outputs from two unrelated prediction models employing different sets of features. To accomplish this objective, experiments were conducted to validate and compare the proposed approach with baseline models. This comparison highlights the efficacy of our method, particularly in terms of improving the detection of drug-related content.

The remainder of this paper is organized as follows: Sect. 2 reviews related work, including datasets used in previous studies. Section 3 introduces the proposed methodology for detecting drug-related content. In Sect. 4, we elaborate on the procedures used to evaluate the proposed method. Section 5 presents and analyzes the evaluation results. Finally, Sect. 6 concludes the paper by offering insights and suggestions for potential avenues for future research.

2 Related Works

Over the past decade, experts from various disciplines, including psychology, computer science, and social sciences, have collaborated to address the problem of illicit online behavior using artificial intelligence methods. There has been a growing focus on developing methods for detecting drug-related content and associated issues, especially in the context of social media. This section presents a literature review that focuses on automated approaches for detecting and classifying drug-related content on social media platforms.

Interdisciplinary research in the field of online detection has mainly focused on examining the processes of online drug-related interactions [24, 25], studying online evidence of an increase in addiction [26], and developing methods that can automatically detect drug-related content and associated events [27]. Researchers have utilized both deep learning and traditional machine learning techniques to identify and detect patterns in social media networks. SVM, LSTM, and RF are the most commonly employed algorithms. Studies such as [23] and [28] have reported accuracy rates of over 90% with SVM, although SVM was outperformed by RF on the F-measure criterion [22, 29]. There have also been promising results from deep learning techniques, particularly CNNs, which were part of the method utilized in [27], and recurrent neural networks (RNNs), which were proposed in [26]. LSTM networks have achieved a precision of 85% in detecting drug-related content on social media, together with techniques such as SVM and RF, as well as bidirectional encoder representations from transformers (BERT) [22].

Machine learning techniques, especially those that use textual features, have recently gained popularity. Textual feature incorporating techniques such as term frequency-inverse document frequency (TF-IDF), bag-of-words (BOW), N-gram, word-to-vector (Word2Vec), and part-of-speech (POS) have been primarily used in classification tasks. Moreover, the BERT and extension of the transformer-XL (XLNET) models are state-of-the-art pretrained language models designed for use in a wide range of NLP tasks [30]. BERT is based on a transformer architecture and is trained using a masked language modeling (MLM) objective, whereas XLNET uses a permutation-based approach instead of MLM. Most previous studies on detecting drug-related content online have examined only a limited number of feature types, impacting their accuracy rates. Textual features considered include Word2Vec [7, 23], TF-IDF [24], both of them [31, 32], Doc2Vec [27], NER, POS [28], BERT, and XLNET [22, 33], whereas contextual and categorical features were examined in [34–36]. These studies are summarized in Table 1 (along with their associated performance metrics). Al-Garadi et al. [33] demonstrated that models using BERT outperform both traditional machine learning and deep learning models (which included XLNET). However, according to [22], XLNET demonstrated greater precision than BERT, although identical recall and $F1$-score values were achieved by both models in the non-drug-related class, even though the drug-related class included two other subtypes.

Furthermore, Nasralah et al.'s research [21] included various drug-related terms associated with opioids as textual features in their evaluation matrix. These terms were drawn from an opioid ontology linked to substantial quality data, which led to accuracy rates of up to 96% in detecting online drug-related content (the most accurate system we are aware of to date), although it had a high level of false positives. Furthermore, this study faced limitations involving the manual work inherent in the ontology employed, making it non-automated and unable to accept new rules. In addition, the authors of [29] used personal attributes as well as social network analysis (SNA) measures, such as degree centrality, betweenness centrality, and clustering coefficients, to describe co-offenders and individual networks, but did not provide either quantitative or qualitative evaluations of the reliability or validity of these measures.

Independently, both BERT and XLNet are powerful language models capable of a wide range of NLP tasks, including analyzing tweets. However, there are some potential advantages to using BERT over XLNet. BERT has been extensively used and evaluated in a wide range of NLP tasks, including text classification, which is a common approach for analyzing tweets [37]. This means that there are many pretrained BERT models available that have been fine-tuned specifically for text classification [38, 39], which could potentially provide better performance on this task than

Table 1 Studies on machine learning and deep learning for online drug-related content detection

References	Year	Algorithm	Feature selection	SN	Dataset size	Performance metric
[18]	2017	Data and language models	Word representation, N-gram	Twitter	267,215 Twitter posts	
[19]	2017	Text classifier and analytical approach	Sentiment scores, substance use variables, and underage variables	Twitter	79,848,992 tweets	
[27]	2017	CNN	Image feature learning with CNN, textual feature learning with Doc2Vec	Instagram	100,500 posts	Acc $= 0.9$ $F = 0.75$
[31]	2017	DT, RF SVMs, NB	String2WordVector, TF-IDF	Twitter	300 tweets	$P = 0.748$ $R = 0.757$ $F = 0.746$
[41]	2017	Biterm Topic Model (BTM)	Text	Twitter	28,711 tweets	
[42]	2017	BTM	URL	Twitter	619,937 tweets	
[34]	2017	LLGC	BOW, users' profiles	Twitter	19,722 tweets, 2,312 users	Acc $= 0.8336$ $F1 = 0.8215$
[36]	2017	SVD, LDA, D-DM, D-DBOW	User feature embedding	Facebook	22M posts	AUC $= 0.86$ for predicting tobacco use, AUC $= 0.81$ for alcohol use, and AUC $= 0.84$ for illicit drug use
[7]	2018	NB, RF, Simple Logistic	Brown clustering, Word2Vec	Doctissimo website Forum	119,562 messages	$P = 0.778$ $R = 0.772$ $F = 0.773$
[43]	2018	SVM, CNN	Word2Vec	Twitter	3M tweets	Acc $= 0.865$ $R = 0.886$ $F1 = 0.866$
[35]	2018	NA	Content features, sentiment analysis, user profile,	Twitter	10% of random tweets	
[28]	2019	J48, LR, Libsvm for SVM, and NB	Named entity recognition (NER), POS, semantic links (SL), and lexical features (LF)	Twitter	1M tweets	$P = 0.95$
[23]	2019	CNN, SVM, RF, NB	Word2Vec, Glove	Twitter	NA	Acc $= 0.857$ (ML) $P = 0.846$ (CNN) $R = 0.891$ (ML) $F1 = 0.862$ (ML)
[29]	2019	DT, NB, LR, SVM, RF, k-NN	Personal feature set and social feature set	Criminal Warehouse	5,780 records with 4,561 unique individuals	$F1 = 0.622$
[32]	2019	SVM, Naive Bayes, CNN, LSTM	TF, TF-IDF, Word2Vec	Twitter	1,794 tweets	Acc $= 0.865$ $R = 0.886$ $F1 = 0.866$
[44]	2019	DT, RF, SVM, RNN-LSTM	Text	Instagram	12,857 posts	$F1 = 0.95$
[24]	2020	Text mining	TF-IDF	Twitter	10,000 tweets	$P = 0.941$ $R = 0.966$ $F = 0.953$ Acc $= 0.928$

Table 1 (continued)

References	Year	Algorithm	Feature selection	SN	Dataset size	Performance metric
[22]	2020	RF, SVM, BiLSTM, BERT,XLNET	Word2Vec	Twitter	5,523,588 tweets	$F1 = 0.71$ for the Pain-misuse class, and 0.79 for the Recreational-misuse class
[21]	2020	SVM, XGBoost, and CNN-based classifier	Word2vec embedding	Twitter	3,696,150 tweets	Acc. $= 0.823$ $P = 0.893$ Recall $= 0.784$ $F1 = 0.835$ AUC $= 0.91$
[45]	2020	RF, KNN, SVM, and L1-regularized LR	NLP features	Free-text narratives, impressions, list of medications	54,359 trip reports	AUC $= 0.94$
[26]	2021	Similarity Network-based Deep Learning (SINDEL)	Word embedding and a network of words	Drugs-Forum	27,154 posts	$F1 = 0.767$
[33]	2021	SVM, RF, Gaussian NB, Shallow NN, KNN, CNN, BiLSTM, BERT, XLNET	BERT	Twitter	16,443 tweets	$F1 = 0.95$
[46]	2022	AdaBoost, LR, SVM, XGB, RF, LSTM, ANN, and CNN	Dataset attributes, tabular data	Database	37,127 distinct cases	

*Precision (P), Recall (R), F-measure ($F1$), Accuracy (Acc), Area Under Cover (AUC), Singular Value Decomposition (SVD), Latent Dirichlet Allocation (LDA), Document Embedding with Distributed Memory (D-DM), Document Embedding with Distributed BOW (D-DBOW), Learning with Local and Global Consistency Algorithm (LLGC), Biterm Topic Model (BTM), Adaptive Boosting (AdaBoost)

XLNet. In addition, BERT is a simpler model than XLNet, which means that it may be easier to train and deploy for analyzing a tweet dataset. In fact, the XLNET model suffers from the limitation of contextualized embedding, which struggles to capture the overall purpose or message of tweets owing to the 280-character limit. The brevity of the tweets hampers the ability of the model to extract the colloquial meanings contained within them [22].

Various combinations of features play a key role in the development of a successful system for detecting drug references in textual content. However, identifying the optimal features can present serious difficulties in classification problems [40]. One reason for this is the existence of different types of features, as well as the potential for varying levels of complexity. Consequently, there is a clear scope for further development. The aim of this research is to present refinements related to data quality and other types of features to maximize the applicability of the framework and the performance of a model based on machine learning. Specifically, a novel machine learning model was proposed using multilevel features (text/NLP, categorical, and numerical features) to bridge the current gap, whereas textual features have hitherto been exclusively considered.

Most state-of-the-art models focus primarily on textual features to classify drug-related content on Twitter. Although these models yield satisfactory results, they have certain limitations. A significant drawback is that relying solely on textual or keyword-based features leads to the misclassification of tweets dealing with news, awareness, or health information related to drug use. Therefore, to enhance performance and overcome these limitations, it is essential to incorporate additional features into the model. These include both categorical and numerical features in conjunction with existing textual features. Such a comprehensive approach can potentially address the issues of precision and recall, thereby improving the overall performance of the model.

3 Methodology

The architecture of the proposed drug-related content-detection module consists of four parts, as shown in Fig. 1. This will be explained in the following section. In the first step, data were collected from Twitter using the Twitter API. Subsequently, standard preprocessing methods for NLP were

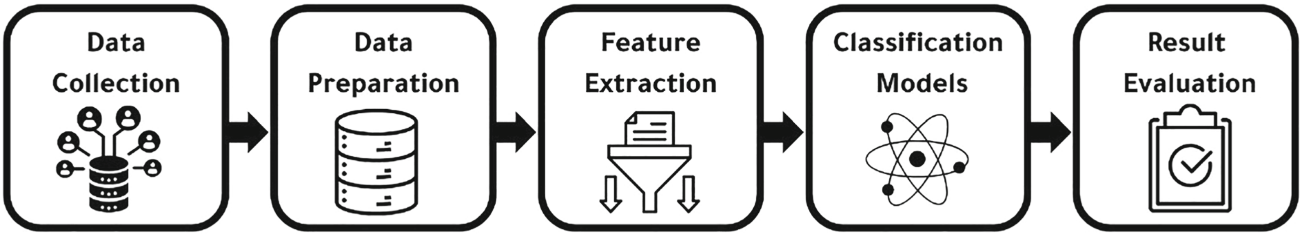


Fig. 1 Proposed architecture

applied, and the tweets were labeled manually as either drug-related or non-drug-related.

To understand the dataset more fully, we conducted and reported an exploratory data analysis (EDA). A range of traditional machine learning models were used to assess the dataset by employing different NLP features in each case. In addition, we assessed the performance of the dataset using a deep learning model, BERT. The metrics used to evaluate the effectiveness of the model included accuracy, $F1$-score, precision, recall, and area under the receiver operating characteristic curve (AUC).

The following subsections detail the essential phases of the module: data collection, preparation, feature extraction, EDA, and formation of predictive models.

3.1 Data Collection and Preparation

The process employed to collect and construct the corpus is illustrated in Fig. 2. This methodology encompasses three primary stages: data collection, corpus cleaning, and data annotation, which are discussed in detail in the following subsections.

3.1.1 Data Collection

In the first three months of 2023, Twitter averaged 436 million active users per month, who collectively contributed approximately 500 million tweets per day, each limited to 280 characters [47]. Although Twitter's public data are accessible to researchers through the Twitter API, certain constraints must be considered. Because datasets of online data-related content are scarce, we began collecting new Twitter data using the API. The collected data included tweet text and user information, including usernames, locations, friend and follower counts, likes, and user descriptions. To facilitate the data collection, we devised search criteria based on popular Twitter topics.

3.1.2 Corpus Cleaning

Before annotation, a cleaning step was performed to prepare the corpus for preprocessing. For this step, duplicate and empty tweets were eliminated, along with non-English tweets and tweets less than 10 words in length. Consequently, the overall tweet count decreased to approximately 150,000.

3.1.3 Data Preprocessing

Text preprocessing is an essential part of text processing, providing a character sequence with a structure and format that supports further analysis in the form of words, sentences, or paragraphs. Several preprocessing techniques can improve data quality in preparation for text data analysis using machine learning algorithms. Python's NLTK library [48] incorporates a comprehensive set of techniques for text preprocessing, including conversion to lower case, removal of emojis and mentions, elimination of stop words, and cleaning punctuation and white spaces. Tokenization was performed using regular expressions following the Penn Treebank tag set. Part-of-speech tags were also applied according to the Penn Treebank tagset, with tweets tokenized into sentences and words. In addition, words were reduced to their base form through lemmatization. Through these preprocessing steps, noise was removed from the text data, making it more appropriate for analysis using NLP algorithms.

3.1.4 Manual Annotation

Annotation is necessary because it has a direct influence on model accuracy when labels or tags are assigned to text data to enable uses such as information retrieval, text classification, and sentiment analysis. Manual annotation takes time and requires considerable human effort; however, it increases the quality of data, leading to improved accuracy of machine learning models in comparison with automated annotation. Moreover, manual annotation allows human expertise and context to be incorporated, which automated methods rarely match. In the context of drug-related content classification, annotation is highly subjective; however, it plays a key role in the identification and categorization of tweets according to specific criteria. For a tweet to be classified as drug-related or non-drug-related in relation to illicit drug use, it should reference certain practices, such as using, selling, smuggling, buying, promoting, or encouraging some kind of

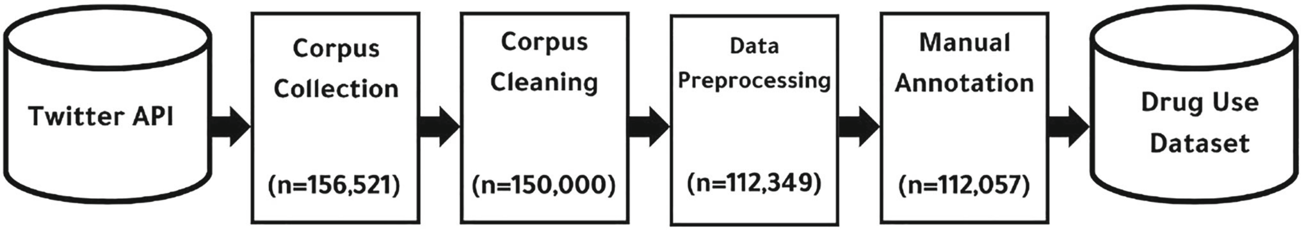


Fig. 2 Data collection methodology

proscribed drug. A tweet that does not mention any of these aspects should be classified as non-drug-related, regardless of whether it refers to themes of religion, ethnicity, politics, or social issues or if it is on a completely unrelated theme, such as sports or fashion. The use of annotations to classify tweets enables researchers and practitioners to better understand the patterns of drug-related content and to formulate effective strategies for prevention and intervention.

In this study, two quality filters were used to ensure the accuracy of the results. In the first filter, the annotator works with the validator. The evaluator then provides the second filter. Great care was taken to minimize potential bias by setting forth explicit directives and affirming our discoveries through a range of methodologies.

3.1.5 Feature Extraction

Before classification, our proposed method converts tweets into vector form, allowing classification models to perform statistical operations. The first step implements NLP preprocessing, as described in Sect. 3.1.3, to derive unigrams. In this study, various feature extraction techniques were employed to create feature vectors.

- N-grams: These are essential components in detection issues. A sequence of characters comprising *n* words can be treated as a single word (unigram), two words (bigram), three words (trigram), etc., depending on the value of *n*.
- BERT: This method generates high-quality representations of words and sentences as vectors. To achieve this, BERT processes a text input to generate a fixed-size vector representation of the input, which can be used for subsequent tasks such as sentiment analysis or text classification. Embedding using BERT has been demonstrated to outperform traditional approaches such as TF-IDF and Word2Vec in various NLP tasks.
- Sentiment analysis: Flair sentiment analysis is a highly capable tool for analyzing sentiments in text data for a variety of purposes, from monitoring social media to analyzing customer feedback. Our study employed a well-known Flair model, known for its high level of accuracy in general and specific contexts.
- Named entity recognition (NER): NER is a task within NLP that involves identifying and classifying named entities in a text. Such named entities typically comprise proper nouns referring to specific individuals, locations, or organizations, as well as dates, times, or entities of various other types.
- POS: POS refers to a word's grammatical category based on its function and how it relates to other words in a sentence. English comprises eight main parts: nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections. Each of these is governed by rules and characteristics that define its usage in a sentence. A good understanding of different parts of speech is necessary for effective and accurate communication in speech and writing.

3.2 Exploratory Data Analysis (EDA)

EDA is an important technique for examining datasets to extract useful information. It is used to identify relationships among explanatory variables, detect errors, and make preliminary selections for a model. Using descriptive statistics and graphical tools, the EDA enables a full understanding of the data to be developed [49]. The principal goal is to gain maximum insight into the dataset, identify outlying and anomalous items, and validate the assumptions of the basic characteristics [50]. EDA was performed using graphical methods to summarize the data in visual and diagrammatic forms.

For univariate analysis, we examined variables individually using techniques such as histograms, box plots, and pie charts for categorical data. Our multivariate analysis involved graphical methods to examine the relationships among combinations of variables. Specifically, we used correlation analysis to measure the total correlation among multiple numerical variables. Section 4 provides a comprehensive explanation of how exploration and visualization are performed for both categorical and numerical features. Python provides several powerful libraries, including Matplotlib, Seaborn, and Stat, along with Plotly Python, NumPy, and Pandas. Matplotlib is a popular Python library used to create static, animated, and interactive visualizations in various formats. It provides a wide range of plotting and customization features. Seaborn is a data-visualization library built on top of Matplotlib in Python. This provides a high-level interface for creating attractive and informative statistical graphics. Seaborn simplifies the process of creating common statistical plots and adds additional functionality for visualizing complex datasets, while the Stat library contains a large number of probability distributions, summary and frequency statistics, correlation functions, and statistical tests. This EDA provided in-depth insights into the dataset, revealing valuable information about the characteristics and attributes inherent in the data.

3.3 Classification Models

3.3.1 Baseline Models

Numerous classification models are available, and their efficacy depends on the specific problem domain. Therefore, it is important to select an appropriate model to construct a robust detection system. In this study, we evaluated the following algorithms:

- SVM (Support Vector Machines)
  An SVM is a supervised learning model that distinguishes two classes along multiple dimensions. This model provides several benefits such as scalability, rapid processing, and real-time ability to detect intrusions while dynamically updating the training patterns. The main function of the SVM model for classification is to determine a hyperplane that maximizes the margin between different classes in the feature space. The equation for the hyperplane is as follows:

$$f(x) = w.x + b \tag{1}$$

  where w is the weight vector, x is the input data vector, and b is the bias term.
- XGBoost (eXtreme Gradient Boosting)
  XGBoost is a popular and powerful machine learning algorithm commonly applied to supervised learning tasks. This is particularly useful for tasks involving regression, classification, and ranking. It is an implementation of the gradient-boosting decision tree algorithm and is especially effective for large datasets, where it can handle missing values and noisy data. The objective function of this model for classification is the weighted sum of the decision tree outputs:

$$f(k) = \sum_{\mathrm{k=1}}^{\mathrm{K}} w_k \cdot T_k(x) \tag{2}$$

  where w_k is the weight assigned to individual trees, $T_k(x)$ is the output of the k-th decision tree, and K is the total number of trees in the ensemble.
- RF (Random Forest)
  RF is an ensemble learning method that amalgamates numerous decision trees and makes predictions based on their combined outputs. It is "random" in that each decision tree created by the algorithm is built from a random subset of the training data, together with a random subset of the features. The primary function of the RF model for classification is to create an ensemble of decision trees with majority votes.

$$\begin{aligned}\mathrm{Output} = \mathrm{Mode}(&\mathrm{Decission\,Tree_1\,(x)}\,,\\ &\mathrm{Decission\,Tree_2\,(x)}\,,\,\ldots,\,\mathrm{Decission\,Tree_N\,(x)})\end{aligned} \tag{3}$$

  Here, each decision tree independently classifies input x, and the mode is considered the final predicted class.
- NB (Naive Bayes)
  NB is a classification algorithm based on Bayes theory. Although simplified, Naïve Bayes is powerful and popular, particularly for NLP tasks and text classification. This is a relatively fast algorithm that can easily handle high-dimensional data. The objective function of the NB model for classification involves maximizing the posterior probability of the class given the input features, which is expressed as:

$$\mathrm{P(y|x) \propto P(x|y)P(y)} \tag{4}$$

  using the Bayes' theorem, where $P(y|x)$ is the posterior probability, $P(x|y)$ is the likelihood, and $P(y)$ is the prior probability. Finally, the model predicts the class with the highest posterior probability.
- LSTM (Long Short-Term Memory)
  LSTM is a type of RNN architecture developed to overcome the issue of vanishing gradients in traditional RNNs. LSTM networks have been extensively applied to sequence modeling tasks, including NLP, speech recognition, and time-series prediction. The objective function of this classification model involves the sequential processing of input sequences through LSTM layers, followed by the application of softmax activation for classification.

$$\mathrm{Output = softmax(LSTM(Input))} \tag{5}$$

- BERT (Bidirectional Encoder Representations from Transformers)
  The BERT model is an implementation of deep learning for NLP tasks. It is based on the transformer architecture model and uses a technique known as self-attention to understand the context of a certain word or phrase in the text. The pretraining of BERT presents it with large

amounts of text data by employing a masked language modeling (MLM) task and a next-sentence prediction (NSP) task, through which it learns the relationships between words and phrases in a text. BERT has achieved state-of-the-art performance in numerous NLP tasks and has been established as a standard model for NLP. Because of its ability to create highly informative text embeddings, it has become a popular choice in a variety of industrial and academic applications. The main function of the BERT model for classification involves using the output of the [CLS] token and applying softmax activation for classification.

$$\text{Output} = \text{softmax}(\text{BERT}\left(\text{Input}\right)_{[\text{CLS}]}) \tag{6}$$

BERT processes the input sequence, and the output corresponding to the *[CLS]* token is used for classification using a softmax function.

3.3.2 Ensemble Method

We propose a method for classifying samples of drug-related texts consisting of combined models, structured as follows:

- BERT base model: The model uses a pretrained BERT model (self-bert) as a feature extractor for text inputs. It also initializes a BERT tokenizer (self-tokenizer) to tokenize the input texts.
- Dropout layer: A dropout layer (self-dropout) with a dropout rate of 0.1 is used after obtaining the BERT output.
- Deep neural network (DNN): This type of artificial neural network comprises multiple layers of interconnected nodes, often called neurons. Each layer in a DNN processes the input data, which includes categorical and numerical features, and progressively extracts more abstract and meaningful information as it moves through the network. These layers are illustrated as follows:

1. Fully connected feature (FC) layer: This layer consists of a simple feed-forward neural network (self-feature_fc).
2. Linear layer with an input size of feature_size and an output size of 128.
3. Rectified linear unit (ReLU) activation function.
4. Dropout layer with a dropout rate of 0.1.
5. Linear layer with an input size of 128 and an output size of 1.

- Ensemble fully connected (FC) Layer: Another linear layer (self-ensemble_fc) combines the outputs from the BERT model and the feature FC layers. It requires an input size of 768 + 1 (BERT output size + feature FC output size) and has an output size of 1.

In the forward method, the model tokenizes the input texts, passes them through the BERT model, applies a dropout, and processes additional features through the feature FC layer. The model concatenates the outputs from the BERT model, incorporates the FC layer, and passes them through the ensemble FC layer. Finally, a sigmoid activation function is applied to the ensemble output, yielding a probability value between zero and one. The model was designed to classify drug-related texts and used two inputs: a set of text samples (texts) and a tensor of features (features). Figure 3 shows the structure of the proposed model.

3.4 Performance Evaluation

Various performance metrics were used to assess the classifier's performance. Accuracy is the most straightforward and commonly used metric.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{7}$$

It considers the number of tweets that are correctly classified as either drug-related (TP; true positive) or non-drug-related (TN; true negative) and the number of incorrectly classified drug-related (FN; false negative) and non-drug-related (FP; false positive) tweets.

Another performance measure is precision, which is calculated by dividing the number of true positives by the total number of tweets classified as positive.

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

Recall is defined as the proportion of correctly classified positive instances to the total number of positive instances. In this study, recall served as a metric to evaluate the effectiveness of detecting drug-related tweets.

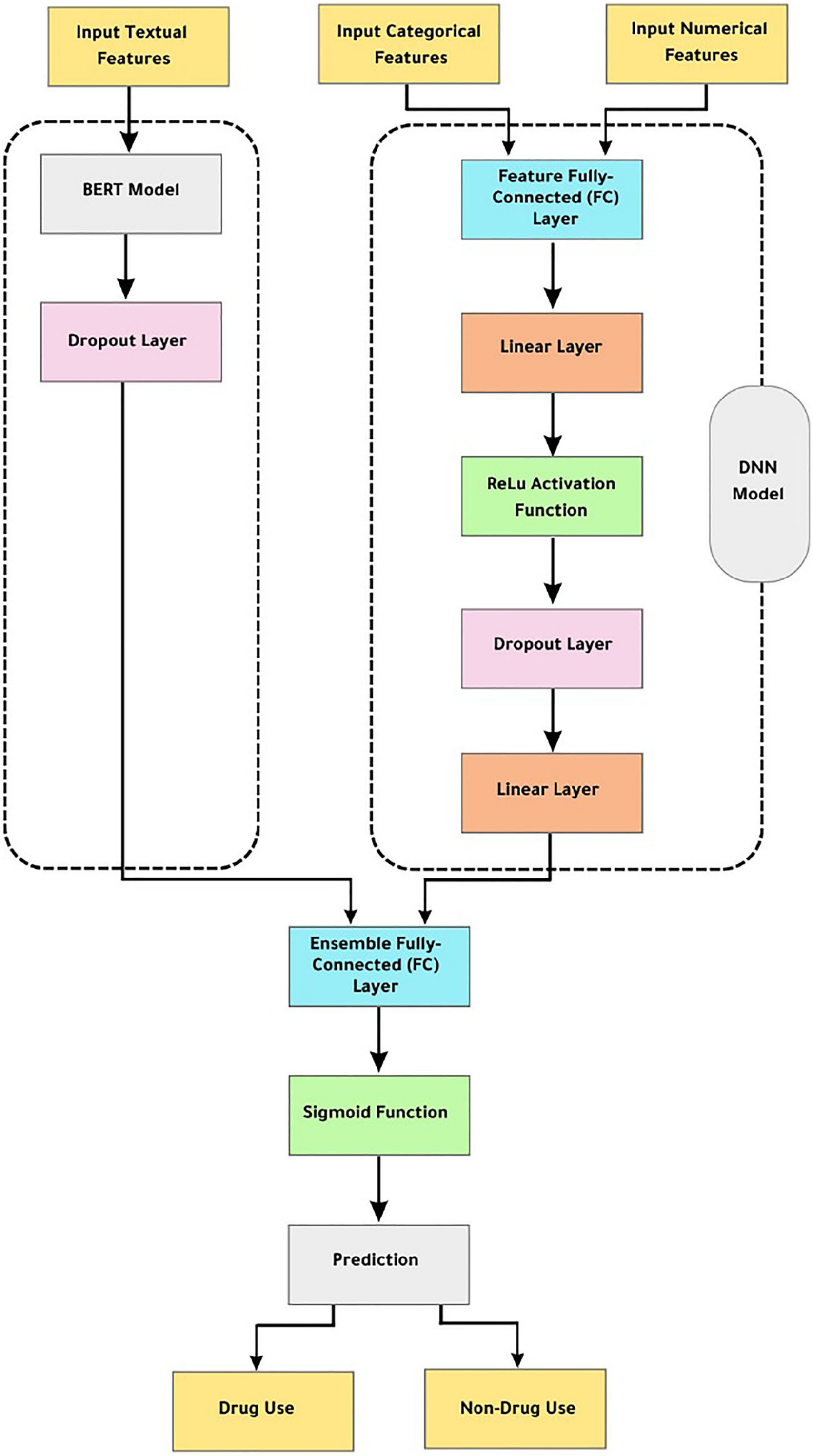
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

To strike a balance between recall (false negatives) and precision (false positives), we employed the $F1$-score, which represents the weighted average of recall and precision. This metric is commonly used to evaluate classification performance because it derives a single value from a combination of precision and recall.

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

Additionally, we generated a receiver operating characteristic (ROC) curve by plotting the true-positive rate against the false-positive rate. The AUC, which falls within the range of zero to one, typically exceeds 0.5.

Fig. 3 The structure of the proposed model



Finally, a confusion matrix, which is an evaluation metric frequently used in classification tasks, was derived. This allows the performance of a classification model to be visualized by providing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

4 Experiments

The framework was implemented in three stages. First, data were collected and prepared. Second, the data were annotated manually. Finally, our new dataset was used to develop and evaluate the classification models.

4.1 Data Collection

Upon successful registration as developers on Twitter's developer platform, we were able to access user status data by incorporating more than 28 attributes. To gather real-time tweet data, we used both the Twitter streaming and search APIs. The dataset collected in this manner spans from October 2008 to May 2022 and comprises 156,521 tweets, along with their associated metadata, including date, user, hashtags, place, retweet count, like count, reply count, and mentioned users. To ensure that the tweets collected from the public Twitter profiles were relevant, we performed searches using English search terms. Our search terms were carefully selected to cover a wide range of formal, slang, and deceptive drug names. Our query included specific terms such as cocaine, arak, codeine, drink, weed, and opium, allowing us to target and collect the desired tweets.

4.1.1 Data Cleaning

Before labeling the dataset, we cleaned the data to ensure quality and filtered out tweets with no text and those with fewer than ten words. Tweets written in languages other than English were excluded. After cleaning, the refined dataset contained 150,000 tweets.

4.1.2 Data Preprocessing

- Following the initial inspection of the dataset, we identified several crucial steps in the preliminary phase.
- Labeling the tweets as "drug-related" and "non-drug-related" classes.
- Introducing a "size" feature to capture each tweet's length and word count.
- Removing unnecessary columns, such as ConversationId, Coordinates, and Tcooutlinks.
- Checking for missing values in the dataset.
- Removing tweets shorter than 10 words.
- Eliminating emojis and mentions.
- Removing white spaces and cleaning punctuation.
- Separating attached words: Following the removal of punctuation and white spaces, the words may become conjoined. This commonly occurs when periods are deleted at the end of a sentence. For instance, the corpus might appear as: "I need another drugdealer show." In such cases, the term "drugdealer" must be divided into two distinct words.
- Converting text to lowercase and removing stop words: Stop words are a set of frequently used words. By excluding these very common words, the focus shifts to more important words, potentially enhancing text-processing accuracy.
- Lemmatizing all words to reduce inflectional word forms into linguistically valid lemmas.
- Eliminating short words, defined as words with fewer than three characters in length.
- Tokenizing text and extracting only verbs, nouns, and adjectives using POS tagging (POS_tag) via the Python Natural Language Toolkit (NLTK) library.
- Stemming—reducing words to their root form.

The implementation of these steps ensures the appropriate formatting of the dataset, extraction of relevant features, and resolution of any inconsistencies or redundancies.

4.1.3 Data Annotation

As discussed previously, manual annotation is considered more dependable and precise than annotation using automatic methods. For this reason, our study follows Wosom's data annotation technique [51], which is designed specifically to make data annotation straightforward and supports the annotation of a range of data types, such as text, audio, video, and images. This technique ensures that data are annotated accurately and reliably.

The annotation process was meticulously conducted by a team of expert annotators from Wosom, comprising approximately 30 members. These annotators, with diverse multidisciplinary backgrounds and expertise, dedicated over six months to accurately labeling each tweet. To ensure that the annotations were consistent and of high quality, a validation system was implemented at predetermined intervals. Through this system, a second layer of validation followed every 10,000 tweets, which enhanced the overall quality control throughout the annotation process.

This approach, featuring collaboration between annotators and validators, resulted in a dataset that is comprehensive and accurately labeled. This dataset has the potential to be a valuable resource in support of research and analysis of the relationship between social media and drug use, facilitating insights and promoting a deeper understanding of these issues. We publicly share our dataset with the IEEE Dataport [52].

4.1.4 Feature Extraction

Preparing the text data for extended analysis involved a set of NLP preprocessing steps, including lemmatization, stopword removal, and tokenization. In addition, to generate word vectors, several feature extraction techniques were applied, including unigrams, bigrams, and trigrams, in conjunction with the BERT and Sentiment methodologies. By applying these techniques, we made it possible to effectively represent text data in vector form, thereby facilitating further analysis and modeling.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Metadata Analysis

To reveal the key features, the tweet metadata were thoroughly analyzed. The drug-related content dataset contained 112,057 tweets from 90,621 unique users. Figure 4 shows an overview of the dataset, indicating that the ratio of drug use to non-drug use tweets categorizes the overall distribution of the two classes of tweets throughout the dataset. Of the total tweets, 48,080 (43%) were classified as drug related, whereas 63,977 (57%) were identified as non-drug related.

Shannon's entropy measure was applied to evaluate the balance within the dataset, yielding a value of 0.985. This measure of entropy showed a well-balanced dataset, representing a relatively even division between drug- and non-drug-related tweets. These statistics and Shannon's entropy provide valuable insights into the content and distribution of the dataset, laying a foundation for additional analysis and development of the model.

Our analysis included a thorough examination of the correlations among the numerical variables to determine any relationships. Figure 5 shows the strong correlation between the number of retweets and likes, demonstrating the close interconnection between these variables. Moreover, a relationship was observed between the number of quotes and likes, suggesting a possible association between these variables.

Based on these correlations, new features can be constructed with the potential to enhance the analysis. For example, features such as the distributions of retweets/likes and quotes/likes can be developed, which may provide important insights into the distribution patterns and relationships among variables.

4.2.2 NLP Analysis

To comprehensively analyze the English language online drug-related content dataset, an in-depth investigation was conducted on the leading 10 unigrams and bigrams through TF-IDF without stop words. TF-IDF assigns numerical weights to words based on their importance in a specific document compared with a corpus.

Figures 6 and 7 illustrate the dominant unigrams identified using TF-IDF for both drug-related and non-drug-related tweets. It is noteworthy that the word "Overdose" appeared in both categories of tweets, although it was more frequent in drug-related tweets. Drug-related tweets mainly featured words related to drug names and methods of drug use, distinguishing them from non-drug-related tweets.

To further explore the dataset, Figs. 8 and 9 illustrate the ranking of bigrams by TF-IDF values in the drug- and non-drug-related tweets, respectively.

The analysis of these linguistic patterns offers deeper insights into the typical features of drug- and non-drug-related tweets in the English language online drug-related content datasets. These findings highlight the high frequency of certain words and the characteristic language used in drug-related discourse compared to non-drug-related content.

4.3 Classification Models

4.3.1 Baselines Models

Classifying tweets as either drug- or non-drug-related involves various classification algorithms. TF-IDF features were employed in the dataset. Every tweet in the dataset was prelabeled with an appropriate class label. Six classifiers were assessed: SVM, XGBoost, RF, NB, LSTM, and BERT. The classifiers were trained under the supervision of tweets with binary labels, enabling them to recognize patterns and make predictions based on the assigned class labels.

For supervised classification, the dataset was divided into training and test sets at proportions of 80% and 20%, respectively. The considerable size of the dataset facilitated this approach, enabling the creation of an independent test set to evaluate the performance of the model. Sufficient data remained available to allow for both training and validation, during which the validation set was used to fine-tune the hyperparameters. The performance of the model was assessed to achieve the best possible outcomes. Six different machine learning models were used in our experiments with textual feature sets: SVM, XGBoost, RF, NB, LSTM, and BERT. Aiming to identify the most precise and effective model, a TF-IDF feature set was employed in combination with N-gram vectorization. We also employed the BERT base model, which is a widely adopted language model that enables fine-tuning for specific classification tasks through

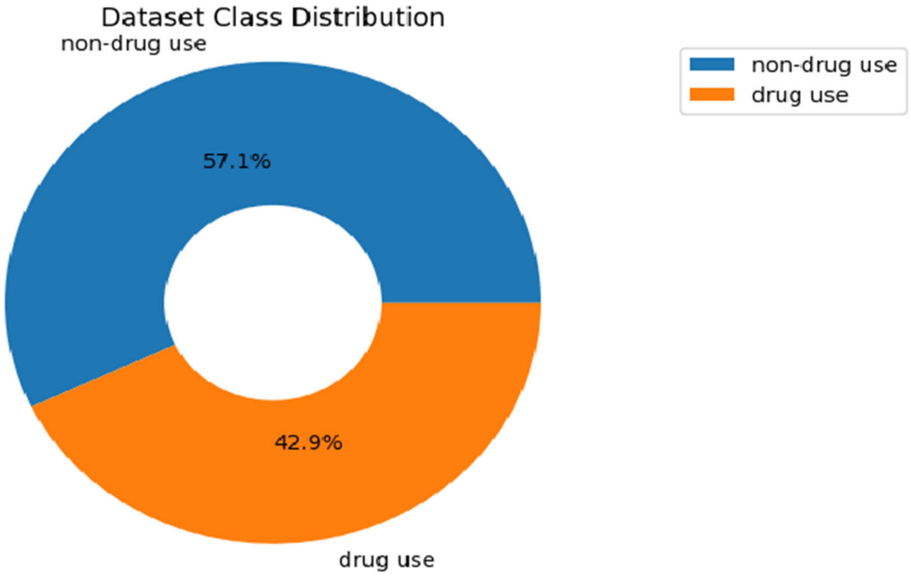
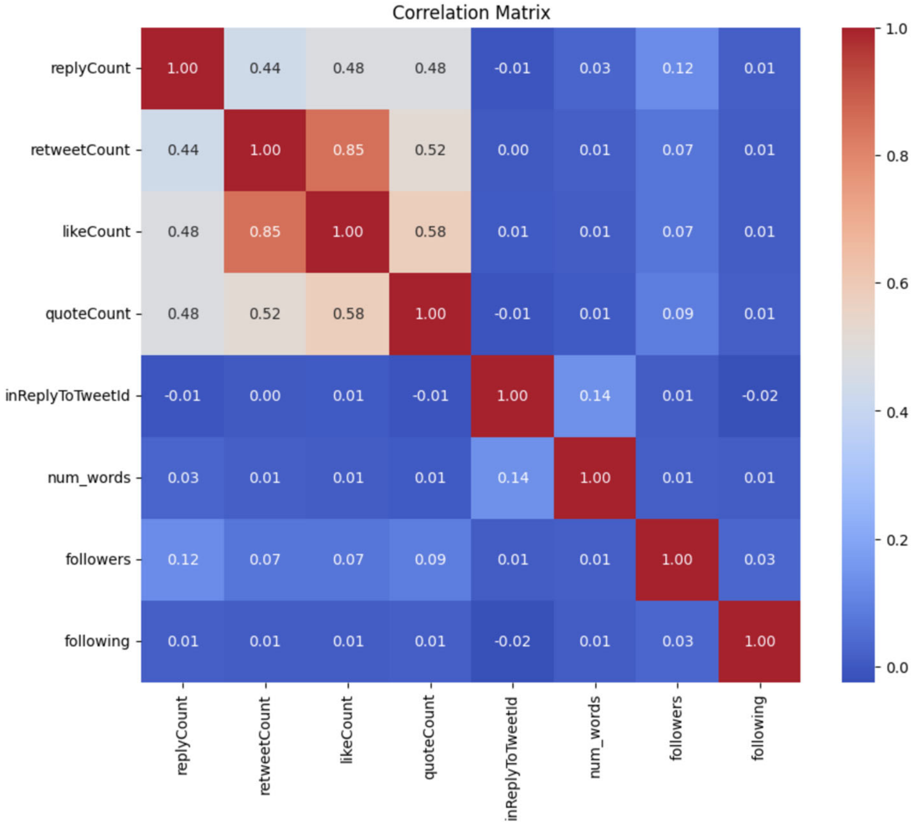


Fig. 4 Dataset class distribution



	replyCount	retweetCount	likeCount	quoteCount	inReplyToTweetId	num_words	followers	following
replyCount	1.00	0.44	0.48	0.48	-0.01	0.03	0.12	0.01
retweetCount	0.44	1.00	0.85	0.52	0.00	0.01	0.07	0.01
likeCount	0.48	0.85	1.00	0.58	0.01	0.01	0.07	0.01
quoteCount	0.48	0.52	0.58	1.00	-0.01	0.01	0.09	0.01
inReplyToTweetId	-0.01	0.00	0.01	-0.01	1.00	0.14	0.01	-0.02
num_words	0.03	0.01	0.01	0.01	0.14	1.00	0.01	0.01
followers	0.12	0.07	0.07	0.09	0.01	0.01	1.00	0.03
following	0.01	0.01	0.01	0.01	-0.02	0.01	0.03	1.00

Fig. 5 Correlations between numerical variables

NLP. The BERT base model was configured with the following parameters: 12 transformer blocks, 768 hidden layers, and 12 attention heads.

4.3.2 Ensemble Method

This section introduces a novel transformer-based ensemble method for the detection of drug-related content, which integrates multiple feature types extracted from tweets. This ensemble model combines the outputs of two predictive models: "stemmed" text from tweets and numerical and categorical features such as "replyCount," "retweetCount," "likeCount," "quoteCount," "outlinks," "inReplyToUser," "mentionedUsers," "hashtags," "verbs," "nouns," "adjectives," "num_words," "followers," and "following."

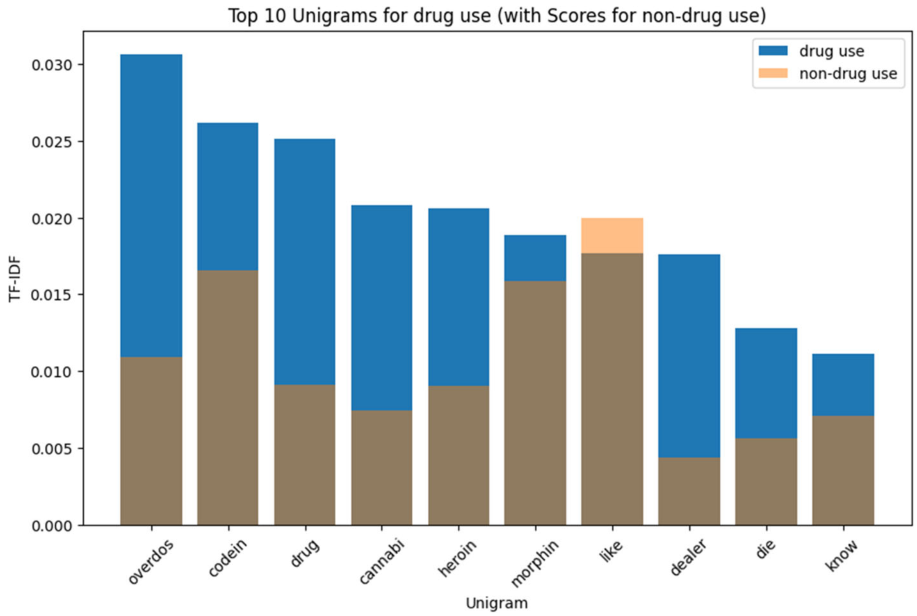


Fig. 6 Top 10 unigrams in drug-related tweets

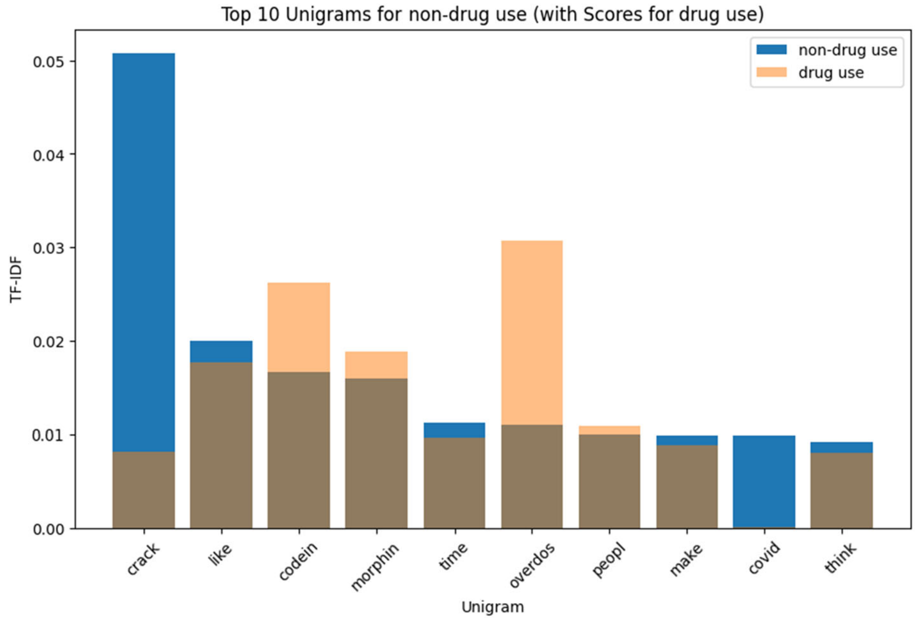


Fig. 7 Top 10 unigrams in non-drug-related tweets

The proposed model classifies drug-related tweets using a specific flow. First, the BERT tokenizer tokenizes the input texts. Subsequently, the tokenized material is passed through the fine-tuned BERT model, where it is further processed to extract representations in the context. To avoid overfitting, the BERT output is subjected to a dropout layer. Simultaneously, separate DNN layers process the categorical and numerical features, starting with a feature fully connected (FC) layer. This is because the DNN model can learn the transformations of non-textual features.

Next, the BERT model's outputs are concatenated with the outputs from the DNN model. These concatenated features are then passed through the ensemble FC layer, where representations from the text and feature inputs are combined. Finally, the application of a sigmoid activation function to the ensemble output generates a probability value between zero and one.

The inputs required by the model comprise a list of text samples and a tensor carrying additional features. The output is a probability value that indicates whether the input text is likely drug-related. The experimental results revealed that the optimal hyperparameters for BERT are a batch size of 64 over five epochs, with a learning rate of 2e−5. The BERT base model utilizes 12 layers of transformer blocks with a

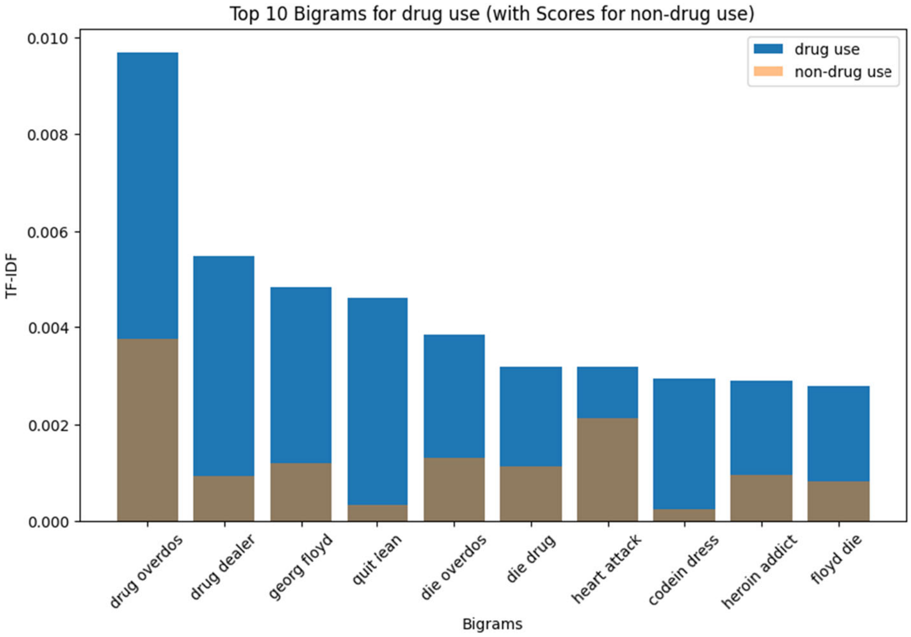


Fig. 8 Top 10 bigrams in drug-related tweets

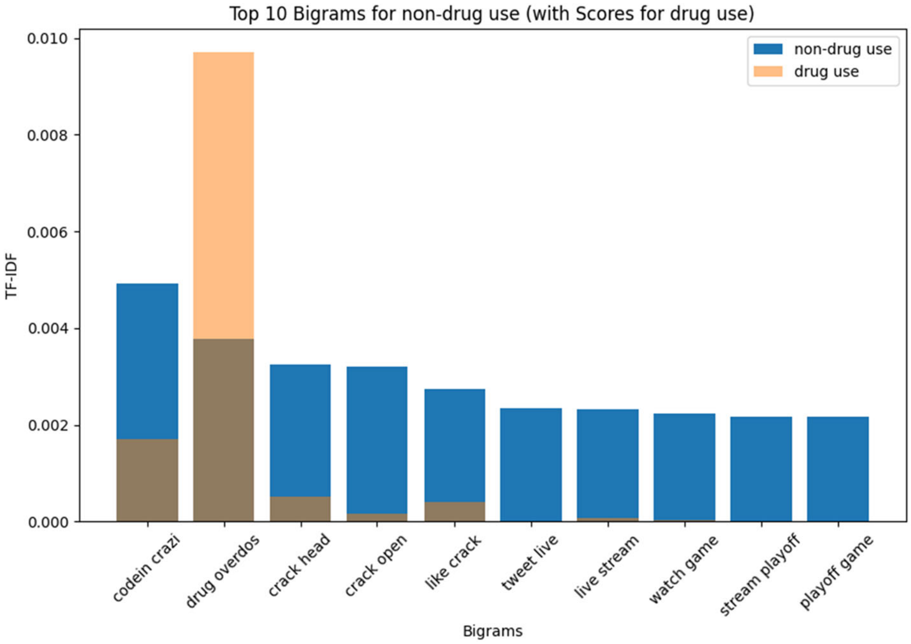


Fig. 9 Top 10 bigrams in non-drug-related tweets

hidden size of 768 and 12 self-attention heads, providing approximately 110 million trainable parameters.

5 Results

This section presents the results of our evaluation of the models' performance, with specific reference to the accuracy of their classification of tweets as drug- or non-drug-related. Table 2 displays the values achieved for the $F1$-score, accuracy, precision, recall, and AUC metrics. Given the balanced nature of our dataset, the precision, recall, and $F1$-score metrics were regarded as ideal for assessing these models.

To assess the performance enhancement achieved by the proposed method, we conducted experiments using various machine learning models. For a comparative analysis, we employed a state-of-the-art transformer-based model to detect drug-related content, as proposed in [33]. Furthermore, we evaluated the efficacy of our approach by comparing it with other commonly used classifiers, including SVM, XGBoost, RF, NB, and LSTM, all of which were enriched with textual features. Table 2 presents the results, where

Table 2 Performance evaluation of all classification models

No	Algorithm	Features	Metrics				
			$F1$-Score	Accuracy	Precision	Recall	AUC
1	SVM	Textual	0.9017	0.8995	0.8819	0.9225	0.8995
2	XGBoost	Textual	0.9000	0.8982	0.8803	0.9206	0.8984
3	RF	Textual	0.8693	0.8662	0.8459	0.8940	0.8664
4	NB	Textual	0.8216	0.8087	0.7641	0.8886	0.8094
5	LSTM	Textual	0.8939	0.9074	0.8790	0.9094	0.9077
6	BERT	Textual	**0.9044**	**0.9159**	**0.8881**	**0.9213**	**0.9165**
7	Ensemble method	Textual Categorical Numerical	**0.9112**	**0.9166**	**0.8953**	**0.9341**	**0.9180**

the BERT base model augmented with textual features as configured in [33], achieved precision, recall, and $F1$-score of 0.8881, 0.9213, and 0.9044, respectively. Our proposed ensemble method, which combines the BERT base model with textual features and a DNN model utilizing numerical and categorical features, clearly outperforms the other machine learning models. The precision, recall, and $F1$-score were 0.8953, 0.9341, and 0.9112, respectively. These results signify not only an improvement in accuracy measures when using our proposed method but also respectable enhancements in precision, recall, and F-measure.

These improvements were expected because there are potential reasons for misclassification errors in the BERT base model. Although the BERT base model with textual features excels at capturing context-dependent word meanings, it lacks the common-sense and pragmatic inference capabilities inherent in human understanding. In addition, humans have a remarkable capacity for generalization and can discern deeper connections beyond the immediate context of words. They weave words into sentences, considering factors such as social group, culture, age, and life experiences, to achieve comprehensive comprehension. Such nuances can be challenging, even for humans from diverse backgrounds or age groups to fully grasp. Although capturing common sense remains a challenge, enhancing the context of a given tweet by incorporating additional user-level information (e.g., details from the user's profile, followers, and following) offers the potential for improved performance. Our experiments unequivocally demonstrated the distinction between our proposed method and previous state-of-the-art models employing machine and deep learning.

Table 3 BERT base model performance for both classes

	Precision	Recall	$F1$-score
Non-drug-related	0.92	0.88	0.90
Drug-related	0.88	0.92	0.90
Accuracy			0.90
Macro avg	0.90	0.90	0.90
Weighted avg	0.90	0.90	0.90

Table 4 Ensemble method performance for both classes

	Precision	Recall	$F1$-score
Non-drug-related	0.93	0.89	0.91
Drug-related	0.89	0.93	0.91
Accuracy			0.91
Macro avg	0.91	0.91	0.91
Weighted avg	0.91	0.91	0.91

Tables 3 and 4 support our argument that integrating the BERT base model with textual features in a DNN model incorporating numerical and categorical features results in higher accuracy and outperforms the BERT base model equipped with textual features alone. The incorporation of numerical and categorical features, such as "replyCount," "retweetCount," "likeCount," "quoteCount," "outlinks," "inReplyToUser," "mentionedUsers," "hashtags," "verbs," "nouns," "adjectives," "num_words," "followers," and "following," gives the model access to supplemental information, thereby enhancing its ability to understand and predict.

By combining these features instead of relying solely on the text, the model can capture a broader array of contextual cues. Consequently, the model exhibits enhanced recall (false negative rate) and precision (false positive rate), achieving improvements of up to 0.0128 and 0.0072, respectively. For instance, the inclusion of diverse features, such as retweetCount or likeCount, can serve as indicators that tweets with high retweet or like counts contain news, awareness, or health information related to drug use rather than instances

Table 5 Confusion matrix of BERT base model classification results (%)

	Non-drug use	Drug use
Non-drug-related	0.93	0.07
Drug-related	0.09	0.91

Rows represent actual classifications, while columns display model predictions

Table 6 Confusion matrix of ensemble method classification results (%). Rows represent actual classifications, while columns display method predictions

	Non-drug use	Drug use
Non-drug-related	0.94	0.06
Drug-related	0.07	0.93

of drug-related misbehavior. Furthermore, certain hashtags that we treated as categorical features serve as indicators of specific communities; for example, hashtags like #420, #Happy420, #4/20, and #20thApril are significant to drug dealers (the number '420' being a coded reference to marijuana). By including these diverse features, we expanded the model's capacity for informed and accurate predictions, clearly demonstrating the benefits of including a wide range of features. These findings demonstrate that the proposed method not only improves the performance, but also bolsters the robustness of the model.

In addition, we conducted an extensive evaluation of the classification accuracy and error prediction of both the BERT base model and the ensemble method. As illustrated in Table 5, the results reveal intriguing insights. Specifically, the confusion matrix highlighted that 9% of the drug-related tweets (positive) were erroneously classified as non-drug-related tweets (negative). Conversely, 7% of the non-drug-related tweets (negative) were inaccurately classified as drug use-related tweets (classified as positive). Upon closer examination, it became evident that most of these misclassified tweets contained similar content but varied in their viewpoints. Table 6 offers further insights, demonstrating a significant reduction in the false negative rate (FN) of up to 2% (from 9 to 7%) when employing the ensemble method in comparison with the BERT base model. This underscores the effectiveness of the proposed method in mitigating misclassification issues with different types of features (textual and non-textual), particularly in addressing the false-negative rate, which is the most sensitive error rate in the context of drug use.

Our experiments clearly demonstrate that our ensemble strategy, which amalgamates the outputs of multiple deep learning models with different types of features, exhibits greater robustness in performance compared to a standalone transformer-based approach. The second observation is consistent with previous studies, in which conventional classifiers demonstrated that ensemble learning often surpasses the efficacy of individual classifiers. However, the improved classification performance is accompanied by a notable increase in computation time, especially during the training phase, because the hyperparameters of each model must be meticulously tuned for optimal performance. The ensemble model generally outperformed the individual components for a range of measures. However, this effect has not been consistently observed, even in traditional machine learning models. We expect that the use of more sophisticated ensemble techniques will allow deep-learning models to operate on features at multiple levels to achieve even higher levels of performance.

The experimental results revealed the optimal hyperparameters for the BERT base model: a batch size of 64 over five epochs and a learning rate of 2e−5. The BERT base model employed 12 layers of transformer blocks with hidden sizes of 768 and 12 self-attention heads, resulting in approximately 110 million trainable parameters. Additionally, to expedite the training process, we leveraged the GPU provided by Google Colab to train the model.

It is important to acknowledge the potential limitations of this research, which stem from data adequacy and societal biases. Despite the legalization of certain substances in many regions, significant stigma associated with drug use remains. Consequently, social media users may be less inclined to discuss this topic openly, and when they do so, their discussions may tend to be negative in nature. This is a crucial point to highlight, as it could have caused our data to either overestimate or underestimate the actual number of drug users. This issue often arises when relying solely on the text of a tweet without considering other features such as the user's profile, hashtags, and emojis.

6 Conclusion

In this paper, we present a comprehensive systematic literature review (SLR) that effectively identifies and elucidates the limitations and gaps in existing techniques for detecting online drug-related content in OSNs. Furthermore, we enhanced the body of knowledge by meticulously constructing and presenting a manually annotated dataset specific to drug-related content on Twitter. This extensive dataset encompasses 112,057 tweets and associated metadata from 2008 to 2022. This study provides insights into the methodologies employed for data collection, EDA, classification, and evaluation. To facilitate future research, this dataset has been made publicly accessible through the IEEE Dataport.

The main contribution of this study is the development of two deep learning approaches for detecting online drug-related content. These models are based on an ensemble method using the transformer architecture. This paper proposes a feature extraction process that incorporates three distinct types of features: content-based, categorical, and numerical. These features were categorized and input into two separate deep-learning models, after which their outputs were combined using an ensemble approach. Notably, the ensemble approach outperformed the baseline classification models. Therefore, the results convincingly demonstrate that the developed online drug-related content detection models offer a practical and effective solution for identifying drug-related content within OSNs.

An innovative aspect of our study is the integration of two models with multilevel features, each capable of functioning independently with different types of features. By incorporating diverse features, the risk of misclassifying tweets is reduced. Relying solely on the texts of tweets may result in incorrect classifications. This ensemble approach allowed us to differentiate tweets that discussed drug use from news, awareness, or health perspectives from those that promoted or suggested evidence of illicit drug use. The proposed comprehensive approach showed the potential to improve precision and recall by decreasing the rates of false negatives and false positives, thereby improving the overall performance of the model.

It is of utmost importance to emphasize that monitoring drug-related tweets and their creators can facilitate early warning systems and offer avenues for initiatives that can predict and prevent drug use. In future studies, we will explore various features individually and in combination to further enhance the effectiveness of the proposed model. One of the objectives of this research was to establish a practical basis for future investigations, especially in the area of developing transformer models, such as XLNET and GPT, for detecting online drug-related content.

Author Contributions All authors have read and agreed to the published version of the manuscript.

Funding The authors would like to thank the Deanship of scientific research in King Saud University for funding and supporting this research through the initiative of the DSR Graduate Students Research Support (GSR).

Data Availability The dataset has been made available on: https://doi.org/10.21227/77am-e529. (Accessed on September, 17, 2023).

Declarations

Conflicts of interest The authors declare no conflict of interest.

References

1. Alsulimani, T.: Social media and drug smuggling in Saudi Arabia. J. Civ. Legal Sci. (2018). https://doi.org/10.4172/2169-0170.1000249
2. Chaffey, D.: Global social media statistics research summary. Smart Insights (2023).
3. Prieto Curiel, R.; Cresci, S.; Muntean, C.I.; Bishop, S.R.: Crime and its fear in social media. Palgrave Commun. (2020). https://doi.org/10.1057/s41599-020-0430-7
4. Al-Otaibi, M.: 8 held for drug dealing through social media. Saudi Gazetti.
5. UNODC/WHO Program on Drug Dependence Treatment and Care.: World Health Organization. Accessed: Jul. 07, 2023. [Online]. Available: https://www.who.int/initiatives/joint-unodc-who-programme-on-drug-dependence-treatment-and-care
6. AlSayyari, A.; AlBuhairan, F.: Relationship of media exposure to substance use among adolescents in Saudi Arabia: results from a national study. Drug Alcohol Depend. **191**, 174–180 (2018). https://doi.org/10.1016/j.drugalcdep.2018.01.025
7. Bigeard, E.; Grabar, N.; Thiessard, F.: Detection and analysis of drug misuses. A study based on social media messages. Front. Pharmacol. (2018). https://doi.org/10.3389/fphar.2018.00791
8. Most popular social networks worldwide as of January 2023, ranked by number of monthly active users. Statista.
9. Kaggle.: Accessed 07 July 2023. [Online]. Available: https://www.kaggle.com/
10. Google Dataset Search.: Accessed 07 July 2023. [Online]. Available: https://datasetsearch.research.google.com/
11. IEEE Data Port.: Accessed 07 July 2023. [Online]. Available: https://ieee-dataport.org/
12. Substance Abuse and Mental Health Services Administration.: Accessed 07 July 2023. [Online]. Available: https://www.samhsa.gov/
13. An official website of the United States government.: Accessed 07 July 2023. [Online]. Available: https://catalog.data.gov/dataset
14. UC Irvine Machine Learning Repository.: Accessed 07 July 2023. [Online]. Available: https://archive.ics.uci.edu/datasets
15. Harvard Dataverse.: Accessed 07 July 2023. [Online]. Available: https://dataverse.harvard.edu/
16. University of California, Riverside (UCR) Library Search.: Accessed 07 July 2023. [Online]. Available: https://search.library.ucr.edu/discovery/search?vid=01CDL_RIV_INST:UCR
17. THE DATALAB.: Accessed 07 July 2023. [Online]. Available: https://thedatalab.com/
18. Sarker, A.; Gonzalez, G.: A corpus for mining drug-related knowledge from Twitter chatter: language models and their utilities. Data Brief **10**, 122–131 (2017). https://doi.org/10.1016/j.dib.2016.11.056
19. Meng, H.W.; Kath, S.; Li, D.; Nguyen, Q.C.: National substance use patterns on Twitter. PLoS ONE (2017). https://doi.org/10.1371/journal.pone.0187691

20. Lokala, U.; Daniulaityte, R.; Carlson, R.; Lamy, F.; Sheth, A.: Social media data for exploring the association between Cannabis use and depression. figshare (2021).
21. Tassone, J.; Yan, P.; Simpson, M.; Mendhe, C.; Mago, V.; Choudhury, S.: Utilizing deep learning and graph mining to identify drug use on Twitter data. BMC Med. Inform. Decis. Mak. (2020). https://doi.org/10.1186/s12911-020-01335-3
22. Fodeh, S.J.; Al-Garadi, M.; Elsankary, O.; Perrone, J.; Becker, W.; Sarker, A.: Utilizing a multi-class classification approach to detect therapeutic and recreational misuse of opioids on Twitter. Comput. Biol. Med. (2021). https://doi.org/10.1016/j.compbiomed.2020.104132
23. Hu, H. et al.: An ensemble deep learning model for drug abuse detection in sparse twitter-sphere. In *Studies in Health Technology and Informatics*, IOS Press, 2019, pp. 163–167. https://doi.org/10.3233/SHTI190204
24. Nasralah, T.; El-Gayar, O.; Wang, Y.: Social Media Text Mining Framework for Drug Abuse: An Opioid Crisis Case Analysis. (2020). https://doi.org/10.2196/preprints.18350
25. Kim, S.J.; Marsch, L.A.; Hancock, J.T.; Das, A.K.: Scaling up research on drug abuse and addiction through social media big data. J. Med. Internet Res. (2017). https://doi.org/10.2196/JMIR.6426
26. Xie, J.; Zhang, Z.; Liu, X.; Zeng, D.: Unveiling the hidden truth of drug addiction: a social media approach using similarity network-based deep learning. J. Manag. Inf. Syst. **38**(1), 166–195 (2021). https://doi.org/10.1080/07421222.2021.1870388
27. Roy, A.; Paul, A.; Pirsiavash, H.; Pan, S.: Automated Detection of Substance Use-Related Social Media Posts Based on Image and Text Analysis. 2017. [Online]. Available: https://www.drugabuse.gov/drugs-abuse/commonly-abused-drugs-charts
28. Jenhani, F.; Gouider, M.S.; Ben Said, L.: Hybrid system for information extraction from social media text: drug abuse case study. In *Procedia Computer Science*, Elsevier B.V., pp. 688–697 (2019). https://doi.org/10.1016/j.procs.2019.09.224
29. Tsai, F.C.; Hsu, M.C.; Chen, C.T.; Kao, D.Y.: Exploring drug-related crimes with social network analysis. In *Procedia Computer Science*, Elsevier B.V., pp. 1907–1917 (2019). https://doi.org/10.1016/j.procs.2019.09.363
30. Shaheen, Z.; Wohlgenannt, G.; Filtz, E.: Large Scale Legal Text Classification Using Transformer Models (2020) [Online]. Available: http://arxiv.org/abs/2010.12871
31. Phan, N.; Bhole, M.; Ae Chun, S.; Geller, J.: Enabling real-Time drug abuse detection in tweets. In *Proceedings—International Conference on Data Engineering*, IEEE Computer Society, pp. 1510–1514 (2017). https://doi.org/10.1109/ICDE.2017.221
32. Hu, H.; et al.: An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. Comput. Soc. Netw. (2019). https://doi.org/10.1186/s40649-019-0071-4
33. Al-Garadi, M.A.; et al.: Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med. Inform. Decis. Mak. (2021). https://doi.org/10.1186/s12911-021-01394-0
34. Fan, Y.; Zhang, Y.; Ye, Y.; Li, X.; Zheng, W.: Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from Twitter and case studies. In *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, pp. 1259–1267 (2017). https://doi.org/10.1145/3132847.3132857
35. Al Dhanhani, S.S.: Framework for Analyzing Twitter to Detect Community Suspicious Crime Activity. Academy and Industry Research Collaboration Center (AIRCC), pp. 41–60 (2018). https://doi.org/10.5121/csit.2018.80104
36. Ding, T.; Bickel, W.K.; Pan, S.: Multi-view unsupervised user feature embedding for social media-based substance use prediction," (2017).
37. Rodrawangpai, B.; Daungjaiboon, W.: Improving text classification with transformers and layer normalization. Mach. Learn. Appl. **10**, 100403 (2022). https://doi.org/10.1016/j.mlwa.2022.100403
38. Qasim, R.; Bangyal, W.H.; Alqarni, M.A.; Ali Almazroi, A.: A fine-tuned BERT-based transfer learning approach for text classification. J. Healthc. Eng. (2022). https://doi.org/10.1155/2022/3498123
39. Bilal, M.; Almazroi, A.A.: Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews. Electron. Commer. Res. (2022). https://doi.org/10.1007/s10660-022-09560-w
40. Pintas, J.T.; Fernandes, L.A.F.; Garcia, A.C.B.: Feature selection methods for text classification: a systematic literature review. Artif. Intell. Rev. **54**(8), 6149–6200 (2021). https://doi.org/10.1007/s10462-021-09970-6
41. Mackey, T.K.; Kalyanam, J.: Detection of illicit online sales of fentanyls via Twitter. F1000Res (2017). https://doi.org/10.12688/f1000research.12914.1
42. Mackey, T.K.; Kalyanam, J.; Katsuki, T.; Lanckriet, G.: Twitter-based detection of illegal online sale of prescription opioid. Am. J. Public Health **107**(12), 1910–1915 (2017). https://doi.org/10.2105/AJPH.2017.303994
43. Hu, H.; Moturu, P.; Dharan, K.N.; Geller, J.; Di Iorio, S.; Phan, H.: Deep learning model for classifying drug abuse risk behavior in tweets. In *Proceedings—2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, Institute of Electrical and Electronics Engineers Inc., pp. 386–387 (2018). https://doi.org/10.1109/ICHI.2018.00066
44. Li, J.; Xu, Q.; Shah, N.; Mackey, T.K.: A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study. J. Med. Internet Res. (2019). https://doi.org/10.2196/13803
45. Prieto, J.T.; et al.: The detection of opioid misuse and heroin use from paramedic response documentation: machine learning for improved surveillance. J. Med. Internet Res. (2020). https://doi.org/10.2196/15645
46. Al Amin, S.; et al.: Data driven classification of opioid patients using machine learning-an investigation. IEEE Access **11**, 396–409 (2023). https://doi.org/10.1109/ACCESS.2022.3230596
47. Smith, A.: "23 essential Twitter statistics to guide your strategy in 2023, (2023).
48. NLTK Library.: Accessed 07 July 2023. [Online]. Available: https://www.nltk.org/index.html
49. Sahoo, K.; Samal, A.K.; Pramanik, J.; Pani, S.K.: Exploratory data analysis using python. Int. J. Innov. Technol. Explor. Eng. **8**(12), 4727–4735 (2019). https://doi.org/10.35940/ijitee.L3591.1081219
50. Kulkarni, A.; Shivananda, A.: Natural language processing recipes. Apress (2019). https://doi.org/10.1007/978-1-4842-4267-4
51. Wosom. Accessed 26 May 2023. [Online]. Available: https://wosom.ai/
52. Al-Ghannam, R.; Ykhlef, M.; Al-Dossari, H.: Annotated drug use tweets. Accessed 17 Sep 17 (2023). [Online]. Available: https://doi.org/10.21227/77am-e529