

RESEARCH

Open Access



Utilizing deep learning and graph mining to identify drug use on Twitter data

Joseph Tassone[†], Peizhi Yan[†], Mackenzie Simpson[†], Chetan Mendhe, Vijay Mago^{*†}  and Salimur Choudhury

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2020 Virtual. 9-10 August 2020

Abstract

Background: The collection and examination of social media has become a useful mechanism for studying the mental activity and behavior tendencies of users. Through the analysis of a collected set of Twitter data, a model will be developed for predicting positively referenced, drug-related tweets. From this, trends and correlations can be determined.

Methods: Social media data (tweets and attributes) were collected and processed using topic pertaining keywords, such as drug slang and use-conditions (methods of drug consumption). Potential candidates were preprocessed resulting in a dataset of 3,696,150 rows. The predictive classification power of multiple methods was compared including SVM, XGBoost, BERT and CNN-based classifiers. For the latter, a deep learning approach was implemented to screen and analyze the semantic meaning of the tweets.

Results: To test the predictive capability of the model, SVM and XGBoost were first employed. The results calculated from the models respectively displayed an accuracy of 59.33% and 54.90%, with AUC's of 0.87 and 0.71. The values show a low predictive capability with little discrimination. Conversely, the CNN-based classifiers presented a significant improvement, between the two models tested. The first was trained with 2661 manually labeled samples, while the other included synthetically generated tweets culminating in 12,142 samples. The accuracy scores were 76.35% and 82.31%, with an AUC of 0.90 and 0.91. Using association rule mining in conjunction with the CNN-based classifier showed a high likelihood for keywords such as "smoke", "cocaine", and "marijuana" triggering a drug-positive classification.

Conclusion: Predictive analysis with a CNN is promising, whereas attribute-based models presented little predictive capability and were not suitable for analyzing text of data. This research found that the commonly mentioned drugs had a level of correspondence with frequently used illicit substances, proving the practical usefulness of this system. Lastly, the synthetically generated set provided increased accuracy scores and improves the predictive capability.

Keywords: Natural language processing, Convolutional neural network, BERT, Twitter analysis, Big data

Background

Introduction

Collecting accurate and up-to-date trend information regarding drug-use is an arduous task [1]. The illicit nature of the topic makes surveying a population difficult, as the potentially illegal nature tends to generate

*Correspondence: vmago@lakeheadu.ca

[†]Joseph Tassone, Peizhi Yan and Mackenzie Simpson have contributed equally to this study

Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay P7B 5E1, Canada



a less honest or unwilling response. This limits the usefulness of the data collected and provides a demand for an accurate system. A prospective solution is in social media, which has been used as a source for studying the mental activity and behavior tendencies of users [2]. Current research has gone so far as to suggest the possible validity in utilizing the information posted online as a substitution for actual surveyed data [3–5]. This fact is not necessarily surprising, as there is widespread utilization and sites such as Twitter are consistently accessed by a significant population of people. As social media is prevalent in today's society, it provides an excellent opportunity for developing a generalized drug detection system, as well as a manner for extracting relevant trends.

Twitter data is not the most consistent or stable information to work with [6, 7]. Inconsistencies within the wording and the lack of discrete variables made analysis and classification a difficult task. Traditional machine learning methods have proved ineffective for our purposes (see “Results” section). As a result, a deep learning approach was used in this research to screen and analyze positively referenced, drug-related tweets. Topic pertaining keywords, such as slang and use-conditions (methods of drug consumption) were used to collect Twitter data. A subset of dataset was then manually labeled with two categories: positive or negative. For clarity, a text such as “smoke weed every day” would register as drug-positive, while a tweet like “all drugs should be illegal” would register as drug-negative. A normal text without reference to any particular drug would also register with a negative result. Following this, a deep learning model using a convolutional neural network (CNN) [8] was trained on the labeled data to classify between positive and negative. A *word2vec* algorithm was used which allowed the embedding of alike words (words having a comparable meaning) to be seen as similar [9]. This helped the CNN care less about the variety of words and instead focus on the semantic meaning of words and their corresponding relationships. Further details are provided in “Methods”.

Training the CNN showed a strong capability for accurately classifying, the details of which can be found in “Classification by CNN-Based Classifier”. Additional classification methods were used to compute the quality of classification; however, the deep learning algorithm with synthetic data was found to outperform them (see “Results” section). Deep learning is not a unique method for performing this task; however, there is little research in utilizing it as a general drug detection system. At the same time, our research found promising results in combining this methodology with synthetically generated data (see “Data pre-processing” section). The following points have been concluded from this research:

- This work verifies that the “possibly_sensitive” tag generated by the Twitter API cannot be used for the classification of drug-positive tweets.
- A CNN model with synthetic data was developed and outperformed other methods in classifying drug-related tweets.
- A novel approach of generating labeled synthetic data improved the accuracy and classifying capability of the model.
- Commonly mentioned drugs had a level of correspondence with frequently used illicit substances.

The remainder of this paper is arranged as follows. The “Related work” describes previous or similarly documented techniques. “Methods” section presents the methodology utilized, with details on the data and pre-processing performed and the quality. “Results” section shows the experimental metrics gained from the SVM, XGBoost, and CNN-based classifiers. Additionally, respective keyword strength and patterns in the data were determined. “Conclusion” section ends the paper and discusses possible future works.

Related work

Health analysis using social media data

Social media data reflects a population's characteristics, including public health information. Many social media platforms such as Twitter and Facebook have a massive user base, constantly generating an enormous amount of messages. For instance, based on a 2019 statistic, 500 million tweets were sent out on a daily basis. Therefore, monitoring and analyzing social media data should be prominent in population-based research, including public health.

In [10], Lamos et al. proposed a way to detect and track influenza in the United Kingdom. Their method was utilizing regression in learning a set of weighted keywords to compute a score, which reflected the influenza rate. Paul et al. [11] assumed that each health-related tweet reflected an underlying ailment and proposed an ailment topic aspect model (ATAM) for syndromic surveillance. The results showed the broad applicability of analyzing Twitter data for public health research. The authors also pointed out the limitation of using Twitter data due to the age of users (Twitter users tend to be teenagers or young adults). Chew et al. leveraged Twitter data analysis to track the trend of sentiment and public attention during the 2009 H1N1 pandemic [12]. Besides the text content and the meta-information, such as keywords count, [3] introduced a predictive model for the classification of healthy and unhealthy populations based on Facebook “likes”. This work also showed that the significant value of Facebook “likes” in public health

prediction and population health-related behaviour analysis. As mentioned above, these researches were focused on population-level health status rather than the study of individual users.

Analyzing individual-level health status of social media users helps doctors or healthcare professionals detect potential patients and provide help. In [13], Twitter data was used for dental pain surveillance. Since dental pain is non-infectious, the purpose of their research was to detect Twitter users with a toothache via data mining. Researchers developed a coding system to analyze the content of the collected tweets. Other similar works include, Coppersmith et al. who built a binary classifier to detect the post traumatic stress disorder of individual Twitter users [14].

Drug abuse detection in social media

Due to the prevalence of social media, research on detecting and monitoring drug abuse-related behaviors have been carried out in recent years. The methods used in these researches can be categorized into traditional statistic methods and machine learning approaches. A semantic prescription drug abuse surveillance platform (PREDOSE) was introduced in [15]. The study scope of the semantic data of PREDOSE is web forum posts. PREDOSE only dealt with three types of data: entities, relationships, and semantic triples. The first stage of PREDOSE was to collect and clean the posts. In the second stage, domain knowledge in drug abuse studies was leveraged to extract and process the information of interest. In the third stage, statistic-based qualitative and quantitative analysis was used to detect the drug user attitudes and behaviors; while temporal analysis was applied to detect the trend of drug abuse.

Sarker et al. proposed a hybrid classification model for automatically monitoring prescription medication abuse from Twitter data [2]. The hybrid classification model was a combination of four traditional supervised learning algorithms, namely: Naïve Bayes, support vector machine, maximum entropy, and a decision tree-based classifier. Since the distribution of abuse and non-abuse tweets was highly unbalanced, the resulting model had a high accuracy yet a poor F1 score.

In [4], the researchers collected tweets with E-cigarette related keywords and manually annotated a small set of data for analysis. The annotated data had five categories, representing the type of corresponding user: individuals, vaper enthusiasts, informed agencies, marketers, and spammers. The classifier used in this work was gradient boosting regression trees. They further studied the importance of each feature regarding the user types. There are some limitations within this work, such as

manual feature engineering and relatively small training dataset.

Social media data contains a great deal of metadata, such as a user's basic information and their interpersonal relationship network. As such, analyzing the high dimensional patterns within this could help enhance the user classification accuracy. In [16], Kursuncu et al. leveraged three levels of features (person-level, content-level, and network-level) in Twitter data for representing a user, where each level of features was called a view. Compositional multi-view embedding (CME) was used for embedding the three levels of features. Experimental results showed that the classification accuracy was improved by using CME.

Hu et al. proposed a deep learning-based Twitter posts drug abuse risk behavior detection system [17]. In their approach, a small number of labeled tweets was used for training the CNN classifier. They further used the CNN to label some of the unlabeled tweets to augment the training dataset. By repeating the above-mentioned steps, the classification accuracy was improved. The problem was that the approach might reinforce the ability of the CNN classifier to detect the patterns of the original manually labeled data, yet miss other patterns which are not in the original labeled data.

Social media text analysis with deep learning

Social media data is worthwhile to mine, as people nowadays tend to express their thoughts through social network platforms [18]. Du et al. [6] proposed a deep learning approach to extract psychiatric stressors for suicide from Twitter data. Keyword-based querying and filtering was used to screen the possible suicide-related tweets from the collected Twitter stream. Following this, a small subset of the candidate tweets was manually labeled (positive/negative) and trained on a CNN. The model was utilized to further select some suicide-related tweets from the candidates. Finally, training with a RNN was completed to perform the stressor recognition task.

Sawhney et al. [19] proposed a long short-term memory recurrent neural network (RNN) to classify suicidal ideation-related social media sentences. They used a two-level embedding approach to prepare the input data for the RNN. The first level of embedding was a 300-dimensional word2vec embedding, while the second level was a sentence level embedding; where a single-layer CNN was used to generate feature maps. They concatenated the pooled feature maps relative to the order of words in the input sentence and used an RNN to do the final classification. The sentence level CNN embedder and the RNN were optimized during training. Severyn and Moschitti used a CNN in Twitter sentiment analysis [20]. As the parameters in a CNN are randomly initialized before

Table 1 Drug related keywords

Drug category	Keyword
Amphetamine	Amy, bennies, benz, dexies, diet pills, get ups, pep pills, wake-ups, amphetamine
Cocaine	Blow, coke, crack, nose candy, cocaine
DMT	Dimitri, dmt
General	Drugs, drug
GHB	Georgia home boy, grievous bodily harm, ghb, liquid ecstasy, liquid e, liquid x
Heroin	Black tar, brown sugar, china white, heroin, mexican brown, skag, white horse
Hydrocodone	357s, dro, fluff, norco, vics, vikes, watsons, hydrocodone
Ketamine	Cat valium, special k, vitamin k, ketamine
Klonopin	k-pin, super valium, klonopin
LSD	Acid, blotter acid, blotter, electric kool aid, lucy in the sky with diamonds, microdot, tabs, lsd
Marijuana	420, blunt, bud, dagga, dope, ganja, grass, green, hashish, hash, hemp, herb, mary jane, pot, weed, marijuana
MDMA	e, ecstasy, happy pill, love drug, molly, vitamin e, xtc, mdma
Mescaline	Blue caps, media luna, mescal, mezcakuba, topi, mescaline
Methamphetamine	Crank, crystal, meth, shards, speed, tweak, uppers, methamphetamine
Mushrooms	Boomers, baps, mushies, shrooms, tweezes, mushrooms
Nitrous_Oxide	Buzz bomb, laughing gas, nitrous, nox, whippets, nitrous oxide
Opioid	Abstral, acetaminophen, actiq, china girl, codeine, dance fever, dilaudid, duragesic, exalgo, fentanyl, hydromorphone, lorcet, lortab, methadone, morphine, murder 8, onsolis, oxy, oxycodone, oxyContin, oxymorphone, percocet, vicodin, opioid
PCP	Angel dust, love boat, peace pill, superweed, pcp
Peyote	Black button, green button, hikuli, hyatari, peyote
Ritalin	Ritalin
Steroids	Gym candy, pumpers, roids, steroids
Synthetic_Cathinones	Bath salts, bloom, cloud 9, cloud nine, cosmic blast, flakka, ivory wave, lunar wave, vanilla sky, white lightning, synthetic cathinones
Xanax	Benzos, xanies, z bars, zanbars, xanax

training and a proper initialization of parameter values is crucial to train a good model; they used a pre-training method as an initialization approach. This was done prior to training the model on their target training dataset. Their pre-training dataset was a set of ten million tweets containing positive words. The main drawback of their approach was that the pre-training process took a significant amount of time (a few days).

Methods

Data source

We created a Twitter developer account which allowed us to employ Twitter's data infrastructure tools and utilize the collected information for research. All the techniques and data mentioned conforms to the Developer's Agreement and Policy enforced by the organization [21]. The data consists of extracted social media information, obtained through Twitter's official public API. Specifically, it is a collection of tweets that were pulled based on 157 keywords; all of which were related to either specific drugs or drug-uses. Drug-use keywords included: "snorted", "snorting", "snort", "pills", "blotter paper", "blotting paper", "tabs", "patches", "injecting", "injected", "inject",

"ingesting", "ingested", "ingest", "smoked", "smoking", "smoke", "chewed", "chewing", "chew", "vaporized", "vaporizing", "vaporize", "vaped", "vaping", "vape", "bong", "pipe", "joint", "needle", "shoot up", "hookah", "grinder", "one hitter", "sinker", "popper", "inhaling", "inhaled", and "inhale" [22]. The remaining drug-only keywords can be viewed in Table 1. These keywords were chosen based on an intelligence report published by the Drug Enforcement Agency (DEA), categorizing drugs by slang and street terms [23].

Twitter data collection ran from October 22 to November 30, 2018. The time period may appear shorter than other research projects [24]. This duration corresponds with the legalization of marijuana in Canada. Therefore, a higher number of relevant tweets could be collected during a period which was advantageous to our research. The data set collected during this time period combined with synthetic data provided a substantial set. Only tweets containing the valid keywords were selected, and misspellings were handled on a case-by-case basis by the Twitter API. The initial set was cleaned with the following filters: removed newlines, contracted extra spaces, removed hashtags, removed emojis, removed reserved words, removed smiley, removed URLs,

removed mentions, removed all punctuation, remove all numbers, converted the text to lowercase, removed stop words, fixed known misspellings, and contracted words. This cleaned dataset consisted of 51 attributes, with 26,184,358 tuples of data (see “Additional files”) and is available for research purposes upon request.

Data pre-processing

The organization of this data demanded a large amount of pre-processing. The tweets that were received were not tagged according to drug-use; therefore, this needed to be determined prior to analysis. In addition, by the sheer volume of the tweets alone, tagging by hand was not a realistic option. As a result, a CNN was trained with a subset of the data to perform the remainder of the task. Many of the tweets also proved to be irrelevant for the purposes of training the neural network; therefore, were removed. The following filters were utilized in generating the dataset following initial collection (see Fig. 1):

- Removed rows with null or empty tweets.
- Removed non-English rows.
- Removed rows with tweets having no keywords.

For the purposes of data analysis, only English tweets were considered, as other languages would have added a level of complexity and required translation. There was the possibility after the initial cleaning that there would either be blank tweets or those without keywords. An example of this would be if a keyword was held in a hashtag or a URL. After these filters were considered, the resulting set consisted of 3,696,150 rows. This may seem like a heavy reduction; however, the filtering procedure is strict in order to ensure data quality. At this stage, a row in the dataset consisted of each collected tweet and the associated metadata attributes attached to it.

The previously mentioned pre-processing tasks standardized the dataset, and initially an attempt at simplifying the information by reducing the number of keywords. The *method of delivery* for a drug was replaced by the literal word “BETA”, while the specific drug was replaced by “ALPHA”. As an example, a phrase like “Smoke weed everyday” would become “BETA ALPHA everyday”. It was theorized that the neural network did not need to interpret the drug that the user was referring to, and the name of the substance could be replaced with a pseudonym. While it was thought this would reduce the burden of training the model, it resulted in a loss of descriptive intent and decreased the model’s accuracy. We therefore decided not to pursue this methodology.

The data was initially extracted using 290 keywords and following the guidelines by the DEA [22], the authors of the paper agreed to reduce the list to 157 keywords.

All keywords were determined based on the most common drugs referenced through a guide released by the DEA [22]. The reasoning behind the reduction was that uncommon slang words like “friend” (for fentanyl) significantly impacted the training performed on the network. Essentially these words were too common in normal speech or in the case of a word like “amp” (amphetamine), was detected as the ASCII characterization of “&”. This resulted in a massive number of drug-negative tweets, risking a skewing in the final metrics. Although it would appear a significant amount of data was removed, these words were uncommon slang terms and the most prevalent keywords still remained.

During the pre-processing task, additional attributes were generated (example: “number of keywords used”) from the text string for possible variable selection and final analysis. Some of these were in addition to the metadata attributes that were collected along with the tweet by the Twitter API. The attributes following pre-processing (either sums or identifiers) included: “id_str”, “text”, “user_followers_count”, “possibly_sensitive”, “timestamp_ms”, “lang”, “original_text”, “user_friends_count”, “alpha”, “beta”, “snort”, “blotter”, “inject”, “ingest”, “smoke”, “chew”, “vaporize”, “vape”, “inhale”, “hitter”, “shoot”, “tabs”, “patches”, “pills”, “bong”, “pipe”, “joint”, “needle”, “hookah”, “grinder”, “sinker”, “popper”, “Amphetamine”, “Cocaine”, “DMT”, “General”, “GHB”, “Heroin”, “Hydrocodone”, “Ketamine”, “Klonopin”, “LSD”, “Marijuana”, “MDMA”, “Mescaline”, “Methamphetamine”, “Mushrooms”, “Nitrous_Oxide”, “Opioid”, “PCP”, “Peyote”, “Ritalin”, “Steroids”, “Synthetic_Cathinones”, “Xanax”, “both”, and “classification”. These attributes together made up each row of the dataset.

The “classification” attribute required the CNN to be trained, meaning testing and training sets needed to be extracted from the pre-processed dataset. Three temporary sets were generated, based either on patterns or a random selection of data:

- *Set 1*: Tweets containing both drug and use-keywords.
- *Set 2*: Tweets containing multiple occurrences of keywords (example: “weed” is mentioned twice).
- *Set 3*: Tweets randomly selected from the cleaned 3,696,150 rows.

These sets were completely unique, with no overlapping data between them. Following this, 2661 rows were randomly selected from the set for manual labeling by the Lakehead University DaTaLab students (Mannila Sandhu and Tanvi Barot). Each tweet was assigned either a 1 (drug-positive) or 0 (drug-negative), depending on the semantic meaning of the text. A text such as “smoke weed every day” would register

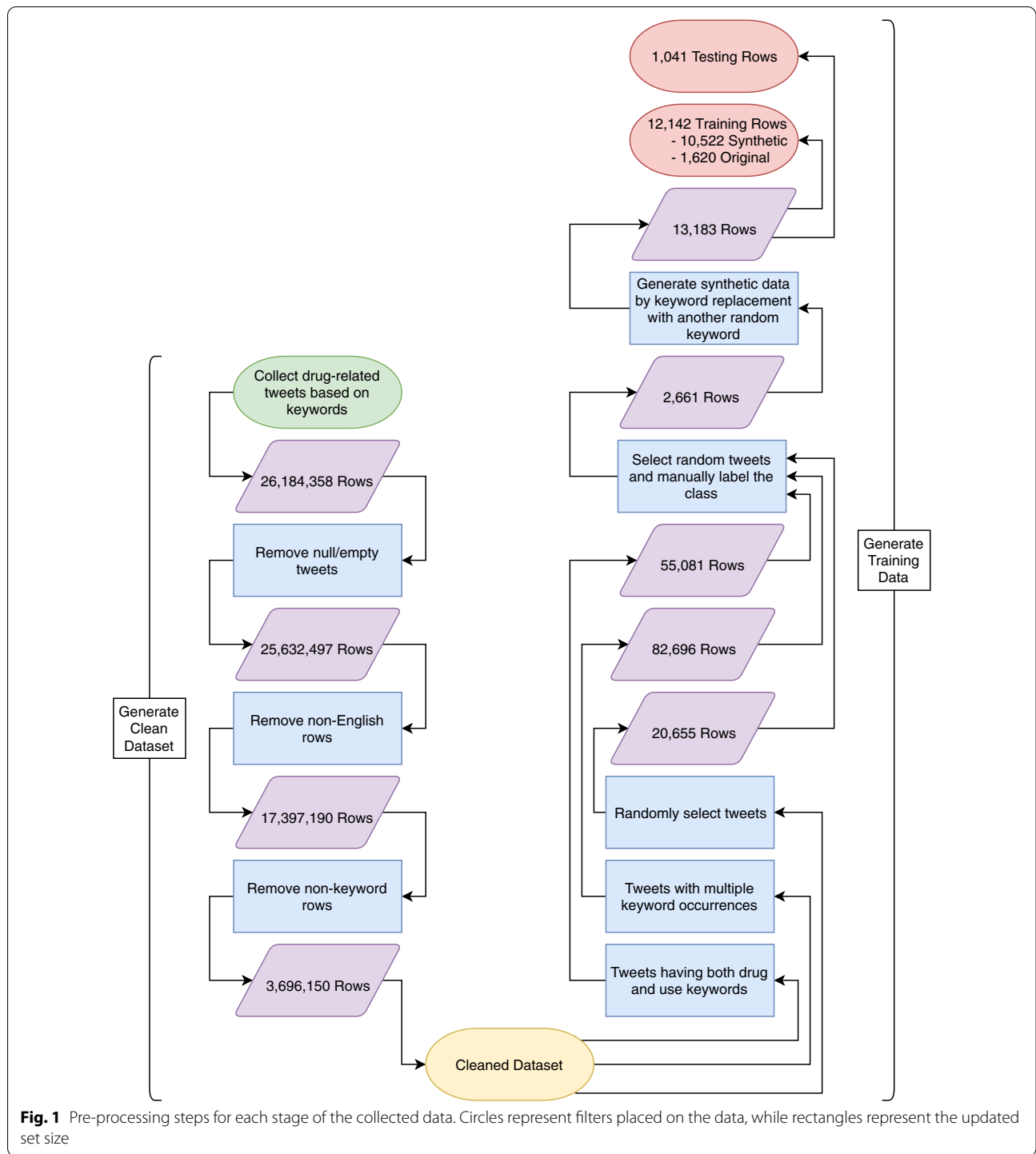


Fig. 1 Pre-processing steps for each stage of the collected data. Circles represent filters placed on the data, while rectangles represent the updated set size

as drug-positive, while a tweet like “all drugs should be illegal” would register as drug-negative. Passive references such as those related to news or simple drug discussion would also register as drug-negative. Likewise, news related tweets or those referring to the observation of drug use would be considered drug-negative.

Essentially, the tweet had to be referring to the active usage of drugs with an implied or directly supportive connotation in order to garner a positive label. The labelled tweets were then verified by social work student Caleb Pears (specializing in addictions research) to ensure all the classifications were consistent.

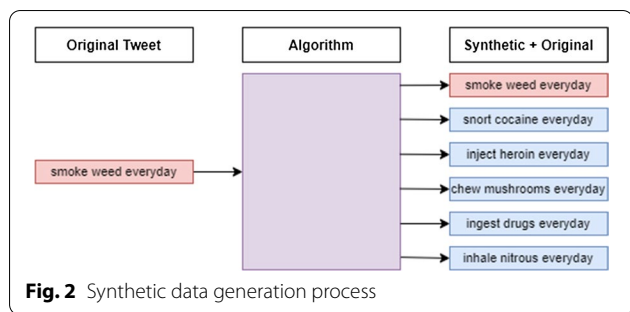


Fig. 2 Synthetic data generation process

Table 2 Example of synthetic data generated from an original tweet

Tweet	Use-keyword	Drug-keyword	Type
Smoke weed everyday	Smoke	Weed	Original
Snort cocaine everyday	Snort	Cocaine	Synthetic
Inject heroin everyday	Inject	Heroin	Synthetic
Chew mushrooms everyday	Chew	Mushrooms	Synthetic
Ingest drugs everyday	Ingest	Drugs	Synthetic
Inhale nitrous everyday	Inhale	Nitrous	Synthetic

This dataset was limited, as the small size meant there was less coverage among the keywords. Training the CNN with this type of data could have allowed a higher probability of misclassification. To ensure proper inclusion and maintain the semantic meaning of the text, synthetic data was generated to compensate. The full process is described through Algorithm 1 and Fig. 2, with a specific example in Table 2. This was done to minimize the sampling bias that could have been present in the neural network. The algorithm simply functioned by looping through the set of tweets, and another loop went through each respective tweet, searching for keywords and replacing them with a random of the corresponding type. Keywords within the “text” attribute of the 2661 rows were replaced respectively by either a random drug (from the 157 mentioned in Table 1) or drug-use keyword. 13,183 rows were contained in this set, among which 12,142 were allotted for training and 1041 for testing. Both training and testing datasets are balanced to help obtain the best results in the case where the proportion of positive and negative tweets in future applications is unknown [25]. As per the “Additional files” section (can be found after reference section), these datasets are available for research purposes upon request. A summary of this methodology, as well as the initial steps, can be seen in Fig. 1.

Algorithm 1: Synthetic training data generation.

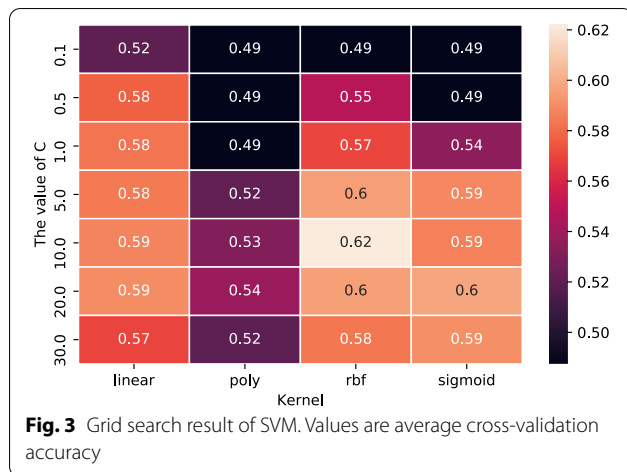
```

Input: The set of tweets: T; the set of use-keywords: U;
       the set of drug-keywords: V
Output: A synthetic set of tweets: W
1 n ← length(T);
2 for i ← 1 to n do
3   t ← i tweet of T;
4   m ← number of synthetic tweets to generate;
5   for j ← 1 to m do
6     w ← t;
7     w' ← Replace use-keyword from w with another
           random use-keyword;
8     w'' ← Replace drug-keyword from w', with another
           random drug-keyword;
9     Add w'' to W;
10  end
11 end
  
```

Data quality

The original dataset consisted of 2661 tweets, each labelled as positive or negative. For verification of quality of labelling, a kappa test [26] was performed to measure reliability. The original labelling was split evenly among two graduate students. Another three graduate students (Punardeep Sikka, Zainab Kazi, and Mohiuddin Qudar) were asked to label the dataset. A kappa statistic was calculated against the original labelling and the three labelers to determine the consistency that the raters agreed on the label. As there were multiple raters a Fleiss’ kappa [27] was performed and resulted in 0.6333, indicating substantial agreement across all raters. It should be noted that this was executed to verify the quality of the original labelling, which had a field expert perform an assessment. The lower result can be attributed to a lack of knowledge of certain less common slang terms among the raters.

The resulting dataset may be considered small; however, falls in line with other similarly published works [28]. The set; however, did have an imbalance with 372 positive and 2289 negative tweets. As such, synthetic generation proved necessary to combat this imbalance. Following the synthetic generation process mentioned in the previous section, the newly formed training dataset contained 6790 drug-negative and 5352 drug-positive tweets. “Results” section further verifies the usefulness of explicitly labelled data from synthetic generation in improving the network’s classification capability. Additionally, of the original 372 drug-positive tweets, only 150 were labelled as “possibly_sensitive” by the Twitter API. This initial result adds weight to the assumption that there is an inaccuracy in utilizing the tag for specifically detecting and classifying drug-related through the Twitter API.

**Table 3 SVM confusion matrix**

	Positive label	Negative label	
Predicted positive	117	8	125
Predicted negative	416	500	916
	533	508	

Support vector machine and XGBoost

Support vector machines (SVM) are widely used in classification problems. However, when the dimension of the input data is large, SVM's are inefficient and take a great deal of time to train. To leverage this tool in our problem, we first used principal component analysis (PCA) to reduce the dimension of our word2vec model. Then we used the same data pre-processing method used for our CNN models to generate training data. We then trained the SVM on the training data with different word vector dimensions. Since the difference of performance is not apparent for different word vector dimension settings, we chose 100 as the word vector dimension. We ran a grid search to select the optimal hyper-parameters for the SVM model. The gamma value is scaled based on: $1/(\text{number_of_features} \times \text{variance_of_data})$. We mainly searched on different kernel methods and the regularization parameter C . The search results are shown in Fig. 3. We finalized the kernel method to be "RBF" kernel, and $C = 10$. The results are summarized in the confusion matrix shown in Table 3. Extreme gradient boosting (XGBoost) is a scalable tree boosting machine learning algorithm which supports parallel computing [29]. We used the same data for the SVM in XGBoost. We searched the optimal hyper-parameters on following parameter space: learning rate {0.01, 0.03, 0.05, 0.07}; maximum depth {3, 4, 5, 6, 7}. We chose the optimal learning rate 0.05 and the optimal maximum depth 5 for

Table 4 XGBoost confusion matrix

	Positive label	Negative label	
Predicted positive	78	15	93
Predicted negative	455	493	498
	533	508	

our final XGBoost model. The confusion matrix shown in Table 4 summarizes the results.

CNN-based classifier

Input embedding

The individual inputs to the CNN were a fixed-size 2-dimensional embedded tweet text (50×400) and each row in the input was a word2vec embedding. The order of rows in the input was correspondent to the order of words in the original tweet text. Unlike many formal English texts, tweets contain many misspellings. If a word2vec model trained on a structured corpus, such as Wikipedia or Google News, was used to embed the words from the Twitter texts, then there would have been a serious out-of-vocabulary (OOV) issue. In this scenario, if a word does not exist in said corpus then it cannot be embedded to the semantic-related vectors. Therefore, a word2vec model (referred to in this case as a Twitter word2vec model) pre-trained on a Twitter corpus [30] was utilized to embed the tweets. The dimension of embedding used in this paper was $K = 400$, meaning each word in the word2vec vocabulary had a correspondent 400-dimensional unique vector. Based on the limitation of characters in each tweet (280 characters), the set length (number of words) of each input text was $L = 50$ [31]. If the number of words in a tweet text was less than 50, then randomized vectors were applied (obeyed uniform distribution, the range was from -0.5 to 0.5) to extend the length of embedded text to 50. For a tweet text that had a length greater than 50, a sliding window of length 50 was employed to get the parts of the text. If any part of the text was labeled as positive (drug-related) by the CNN, a positive label was assigned to the whole text. Figure 4 is a visualized 2-dimensional embedded text example.

CNN architecture

In this paper, a similar CNN architecture proposed by Kim [8] was applied. This CNN architecture had one convolution layer, and the shape of each convolution filter is rectangular. The length of each convolution filter was 400 (same as the dimension of the word2vec embedding we use). The filters were; however, grouped by different heights. There were five groups of filters, where the respective size (height) of each within the groups were 3,

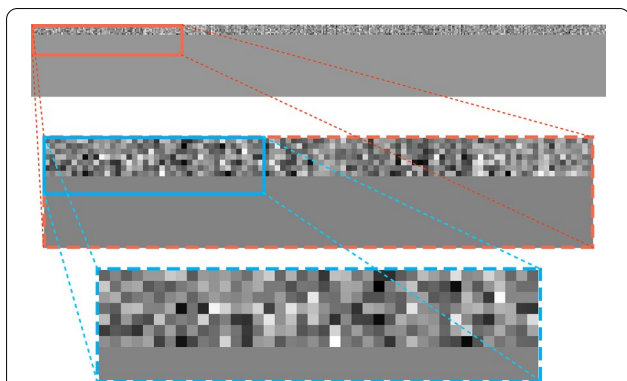


Fig. 4 Visualization of the word2vec embedded text "i think all drugs should be illegal". The first sub-illustration is the original word2vec embedded text (50 × 400), the second and the third sub-illustrations are zoomed parts of the original embedding

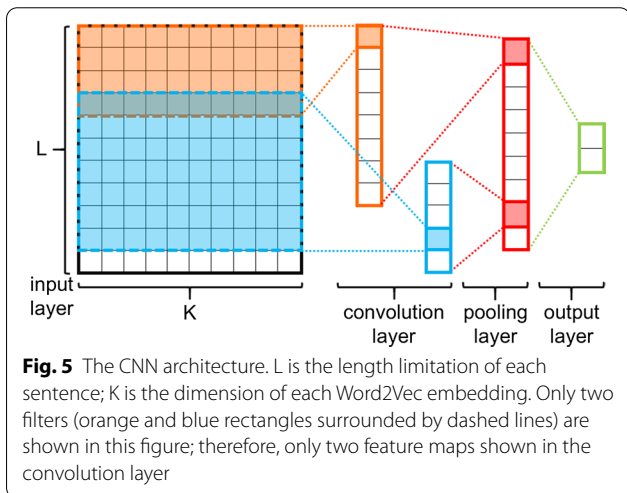


Fig. 5 The CNN architecture. L is the length limitation of each sentence; K is the dimension of each Word2Vec embedding. Only two filters (orange and blue rectangles surrounded by dashed lines) are shown in this figure; therefore, only two feature maps shown in the convolution layer

4, 5, 6, and 7. In other words, the shape of filters could be 3×400 , 4×400 , 5×400 , 6×400 , or 7×400 . Each group had 64 filters, so there were 320 in total. A one-max pooling approach was used to get the maximum values of each feature map and then concatenate them into an array of 320 values. The output layer, which as fully connected to the pooling layer, had two output neurons. Figure 5 is a simplified version of the mentioned CNN architecture.

Training

The training batch size was 64 in the experiment. The Adam optimizer[32] was used to minimize the loss value during training and the learning rate was set to 10^{-3} . Since the number of positive labeled tweets in the training dataset was initially less than the number of negative labeled tweets, a weighted cross-entropy function was used as the loss function (see Eq. 1, where Y represented

the target labels; \hat{Y} represented the predicted scores; and ω was the positive weight).

$$loss(Y, \hat{Y}) = Y[-\log(\hat{Y})]\omega + (1 - Y)[- \log(1 - \hat{Y})] \tag{1}$$

The pseudo-code of the training algorithm is shown in Algorithm 2. T_{batch} is used to represent the set of tweets in a training batch (there is no overlap between any pair of batches and the union of all the training batches is the training tweet dataset T); Y_{batch} is the set of labels of the corresponding training batch (Y is the set of labels of T). E_{batch} represents the word2vec embedded T_{batch} . $W2V$ represents the word2vec embedding dictionary. θ represents the parameters in the neural network model. γ is the learning rate. The value of γ and batch size were chosen via a grid search on our dataset. Due to the lack of training data, a tenfold cross-validation is used to ensure the model converges and avoids over-fitting.

Algorithm 2: Training algorithm (one epoch).

```

Input: The set of tweets:  $T$ ; the set of labels:  $Y$ ; the word2vec model:  $W2V$ 
Output: A trained neural network parameters:  $\theta$ 
1 Randomly initialize the parameters in  $\theta$ ;
2 while  $T$  is not empty do
3   Let  $T_{batch}$  be the next batch of tweets from  $T$ ;
4   Let  $Y_{batch}$  be the corresponding set of labels from  $Y$ ;
5   Remove  $T_{batch}$  from  $T$ ;
6   Let  $E_{batch}$  be an empty set;
7   while  $T_{batch}$  is not empty do
8     Get a tweet  $tweet$  from  $T_{batch}$ ;
9     Remove  $tweet$  from  $T_{batch}$ ;
10    Let  $words$  be a list of the sequence of words in  $tweet$ ;
11    Let  $embedding$  be an empty list;
12    while  $words$  is not empty do
13      Get the first word  $w$  from  $words$ ;
14      Remove  $w$  from  $words$ ;
15      Let  $w2v$  be the embedding of  $w$ ;
16      if  $w$  is in  $W2V$  then
17        | Assign  $w2v$  with output from  $W2V$  model;
18      else
19        | Assign  $w2v$  with random values between  $-0.5$  and  $0.5$ ;
20      end
21      Append  $w2v$  to  $embedding$ ;
22    end
23    Make  $embedding$  a  $[50 \times 400]$  matrix, pad with zeros if the length of  $embedding$  is less than 50;
24    Append  $embedding$  to  $E_{batch}$ ;
25  end
26  Let  $\hat{Y}_{batch}$  be the set of predicted labels of  $E_{batch}$ ;
27  Compute the loss (according to Eq.1):
28   $batch\_loss \leftarrow loss(Y_{batch}, \hat{Y}_{batch})$ ;
29  Use AdamOptimizer to optimize  $\theta$  regarding  $batch\_loss$ ;
30 end
    
```

In the experiment, the CNN model was trained separately on two datasets. One training dataset was the original with 2661 manually labeled samples (372 labeled positively), while the other training dataset was the original+synthetic dataset with 12,142 samples

(5352 positive labeled samples). Since the neural network parameters were randomly initialized (through the Xavier Glorot method [33]), we trained and tested our models twenty times to compute the standard deviation of testing accuracy: CNN model-A is 0.0098, CNN model-B is 0.0125. The best model selected was trained on two datasets, referring to them respectively as CNN model-A (trained on the original+synthetic dataset) and CNN model-B (trained on the original dataset). The testing accuracy of CNN model-A was 82.31% and CNN model-B was 76.35%, making the synthetic model an improvement over the original. The source code is published on Github.¹

Results

Evaluation metrics

The following metrics were used to evaluate each model: accuracy, precision, specificity and recall (sensitivity), F1 score, receiver operating characteristic (ROC), and area under the curve (AUC). Although accuracy is the primary indicator of classification when the number of positive and negative testing samples are balanced, it is not comprehensive. A higher precision represents less false-positive predictions occurring, while a higher recall represents less false-negative predictions occur. F1 score is the harmonic mean of precision and recall: $(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$. We used the F1 score to measure the balance between precision and recall. AUC is derived regarding the ROC curve, which indicates the capability of a model to distinguish between classes. A high AUC value shows that corresponding model has a good distinguishing capability.

Classification by support vector machine and XGBoost

As mentioned previously, training SVM and XGBoost on the original word2vec-embedded tweets does not produce accurate results. Therefore, we used principal component analysis (PCA) to reduce the dimension of the word vectors before embedding the tweets. To compare the effect of different dimension reduction strength, we reduced the dimension of word vectors to different values (between 10 and 100), see Fig. 6. The results indicate that there is no significant correlation between the PCA dimension reduction strength and model performance. Both SVM and XGBoost achieved relatively high recall, but low precision. Although the AUC of SVM appears to be satisfactory, the overall accuracy is relatively low. Therefore, we cannot consider the SVM and XGBoost as good models in this experiment.

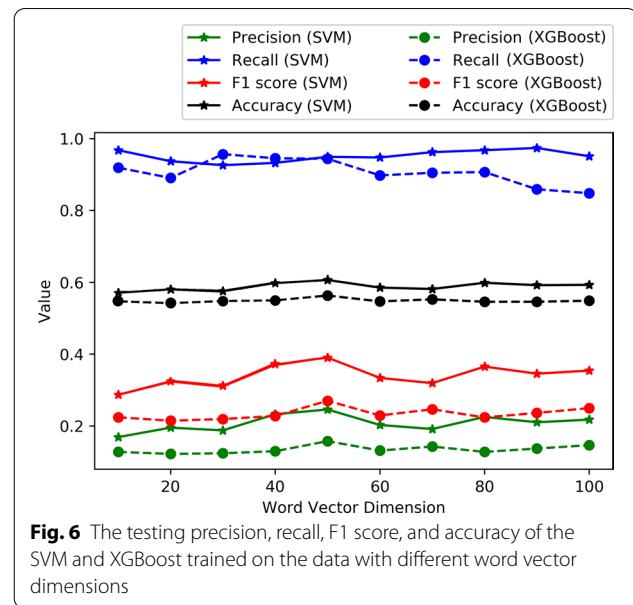


Fig. 6 The testing precision, recall, F1 score, and accuracy of the SVM and XGBoost trained on the data with different word vector dimensions

Classification by CNN classifier

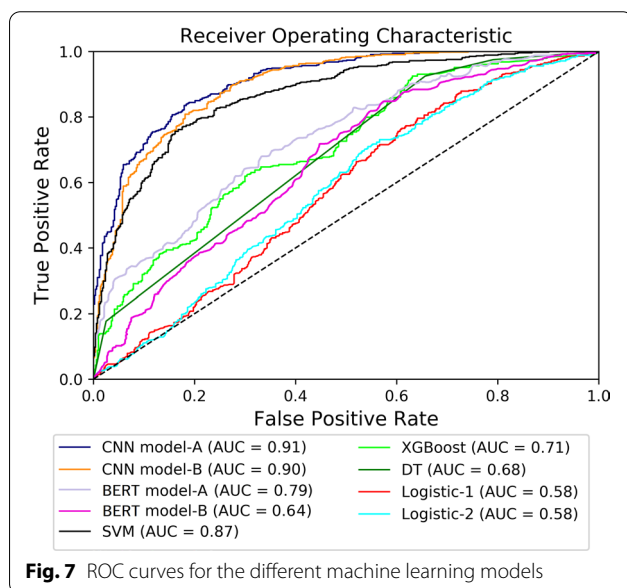
In terms of AUC and accuracy, the CNN (model-A) outperformed the previous models. Results utilizing decision trees and regression based models were not accurate enough for classifying this particular dataset. Instead the very structure and semantic meaning needed to be explored for significant conclusions. The results of all the models are summarized in Fig. 7 and Table 5. All classifiers were trained on a similar sized dataset in order to be properly compared against the neural network-based classifier.

Two CNNs were developed, and it should be of no surprise that the model trained with a larger set had an improved accuracy and AUC. However, a fascinating conclusion was the improved result originating from the synthetic data. While, having a similar AUC, the accuracy of the two models differs by a fair amount. This indicates that the synthetic data had a positive impact on the training of the network. It is possible that this can be attributed to expanding the results, considering keywords that may have been missed in the original training set. Although the ROC itself did not shift dramatically, diversifying the set further would most likely cause a worse classification in CNN model-B. The reasoning is simply that the model does not consider enough keywords, hence why the accuracy is lower. We also trained two BERT-based classifiers (pre-trained on Wikipedia and BookCorpus) [34], namely BERT model-A and BERT model-B. The pre-trained BERT model was implemented

¹ https://github.com/bppvapp/Twitter_NLP.

Table 5 Metrics for the different machine learning models

Models	Accuracy (%)	Precision	Recall	F1 score	AUC
CNN model-A	82.31	0.893	0.784	0.835	0.91
CNN model-B	76.35	0.597	0.906	0.719	0.90
BERT model-A	79.27	0.850	0.713	0.775	0.79
BERT model-B	64.25	0.871	0.338	0.669	0.64
Decision tree	63.40	0.925	0.584	0.716	0.68
SVM	59.33	0.220	0.943	0.356	0.87
XGBoost	54.90	0.146	0.847	0.246	0.71
Logistic-1	57.44	0.873	0.546	0.672	0.58
Logistic-2	54.56	0.954	0.525	0.677	0.58



through Tensorflow-Hub.² Similar to the CNN models, BERT model-A was trained on the original+synthetic dataset, while BERT model-B was trained on the original dataset. BERT model-A achieved better overall results than BERT model-B, but both BERT models failed to outperform the CNN model-A in our task. Moreover, if we ignore the network depth, the BERT algorithm has a time complexity of $O(n^2)$ (n is the length of text, measured in the number of words), whereas, the CNN-based algorithm has a time complexity of $O(n)$. Overall, the CNN (model-A) proved to be the best classifier in terms of classifying drug-related tweets.

Initial analysis on a subset of all the drug-positive tweets classified by the CNN (794,547 of the 3,696,150) was performed. As previously mentioned, the drugs

were broken down into categories (by keyword). During pre-processing, the sums of each of these categories was taken based on the specific occurrence of a keyword within the tweet. The purpose of this exercise was to best determine the drugs referred to most frequently, as well as the occurrence of individual drug-uses. The results of this analysis can be viewed in Figs. 9 and 10 respectively. As seen, the most common reference by a large margin was towards marijuana, with cocaine being a much lower second. An interesting facet of these numbers is they almost correspond with the literal drug activity displayed in true society [35]. Figure 8 was taken from a national survey, quantifying the most actively used illicit substances. If this is compared against the results in Fig. 9, then it can be seen that many of the common street drugs used were also the most commonly mentioned. The same idea can be said towards drug-uses, as smoking is the primary intake medium for marijuana. Though not all drugs align perfectly, many drugs such as marijuana, cocaine, and methamphetamine are still close.

Currently, Twitter utilizes a tag known as “possibly_sensitive” for declaring data that may be considered inappropriate to some readers. However, the tag does not specifically state why a tweet is targeted or if it can be utilized in classifying drug-positive tweets. Figures 9 and 10 show that the tag most likely cannot be used for this task. While drug-related tweets appear to be considered more sensitive than not, there are still a significant number that are missed. Therefore, one can most likely deduce that the “possibly_sensitive” tag is not a replacement for the classification done by the neural network.

Additional analysis

Keyword strength determination

Following a similar process to that in [36]; after the data had been classified with the CNN, association rule mining was performed. The purpose of this process was to best determine the important relationships within the data. To begin the tags “possibly_sensitive” and “drug_negative” were removed from the data. The former meant that the tweet was deemed by Twitter as a possibly sensitive tweet. The latter was classified by the CNN as having a negative association to drugs respectively. These tags were pruned as they clouded important rules within the data. The “possibly_sensitive” tag is present on every tweet that was put through the CNN, while the “drug_negative” is the opposite of the “drug_positive” tag. The balance between the two being 78.503% and 21.497% of the 3,696,150 tweets. The remaining tags that were considered in the association rule mining are “drug_positive” and the parent terms, meaning a term such as “weed” would be considered as “marijuana” in this example. In Table 6 the statistics

² https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1.

Drug	Lifetime (2016)	Lifetime (2017)	Past Year (2016)	Past Year (2017)	Past Month (2016)	Past Month (2017)
ILLICIT DRUGS	130,628	134,704	48,501	51,795	28,564	30,476
Marijuana	118,524	122,943	37,570	40,935	23,981	25,997
Cocaine	38,880	40,550	5,071	5,943	1,874	2,167
Crack	8,776	9,599	882	930	432	473
Heroin	4,981	5,295	948	886	475	494
Hallucinogens	41,490	42,072	4,903	5,125	1,390	1,438
LSD	25,861	26,069	1,896	2,243	374	512
PCP	6,450	6,098	103	96	21	37
Ecstasy	18,459	19,173	2,485	2,487	619	587
Inhalants	24,404	25,187	1,660	1,759	600	556
Methamphetamine	14,533	14,722	1,391	1,633	667	774
Misuse of Psychotherapeutics	nr	nr	18,671	18,077	6,207	5,956
Pain Relievers	nr	nr	11,517	11,077	3,350	3,239
Tranquilizers	nr	nr	6,060	5,944	1,953	1,727
Stimulants	nr	nr	5,647	5,839	1,735	1,841
Sedatives	nr	nr	1,531	1,351	497	352
OPIOIDS (HEROIN USE OR PAIN RELIEVER MISUSE)	nr	nr	11,824	11,401	3,649	3,549

Fig. 8 2017 illicit drug use survey results [35]

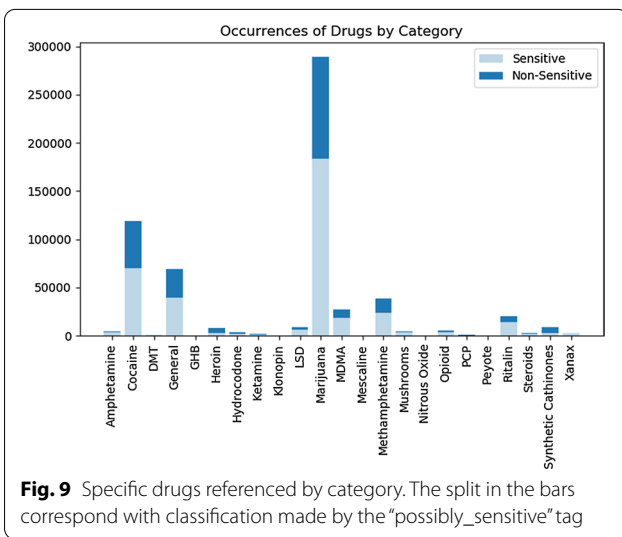


Fig. 9 Specific drugs referenced by category. The split in the bars correspond with classification made by the “possibly_sensitive” tag

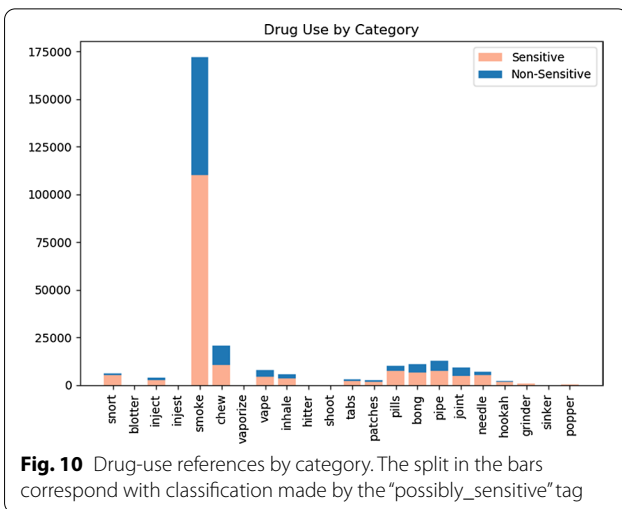


Fig. 10 Drug-use references by category. The split in the bars correspond with classification made by the “possibly_sensitive” tag

that describe the amount of tags per tweet are shown. Frequency of specific tags is shown in Fig. 11.

A maximum of 5 tags was chosen for the mining of association rules, as it only excluded sets found in 80 tweets. Sensitivity analysis was then performed for the number of rules generated depending on the minimum support and confidence settings. This sensitivity can be seen in Fig. 12 from which a minimum support of 0.0003 and minimum confidence of 0.3 were chosen. The top 5 generated rules for the chosen minimum confidence and support can be seen in Table 8, where they are sorted by confidence. Full statistics for the 23 rules generated can be seen in Table 7. A network representation of the rules can be seen in Fig. 13. We applied the HITS algorithm [37] that is designed for finding hubs and authorities in the context of websites to the rule set. The non-zero hubs in order of precedence were found to be “methamphetamine”, “pipe” and “opioid”. The non-zero authorities in order of precedence were “opioid”, “pipe” and “methamphetamine”. This shows that there appears to be no relationship between the rules that have “drug_positive” in the consequent, meaning they independently hold. This is due to the fact that the only “loop” present in the portion of the network containing the rest of the rules is between “marijuana” and “drug_positive”. Examining Table 8, we can see that the strongest relationships revolve around largely the trio “opioid”, “pipe” and “methamphetamine”. Since this is lacking “drug_positive”, we can conclude that discussion around the topics of opioids

Table 6 Statistics that describe the amount of tags per tweet

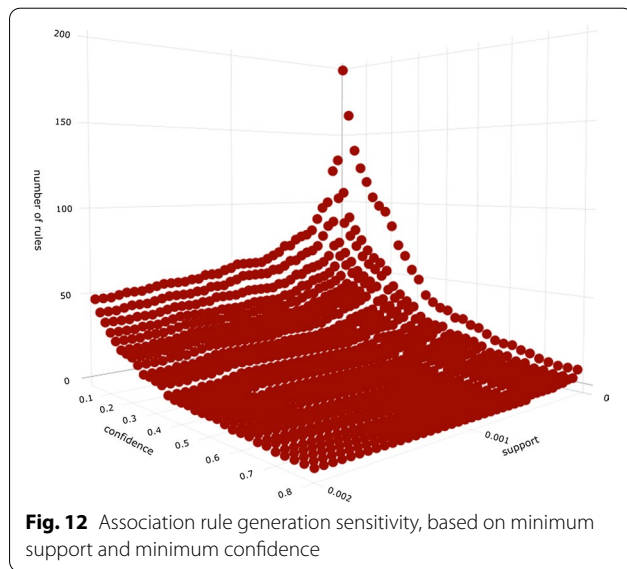
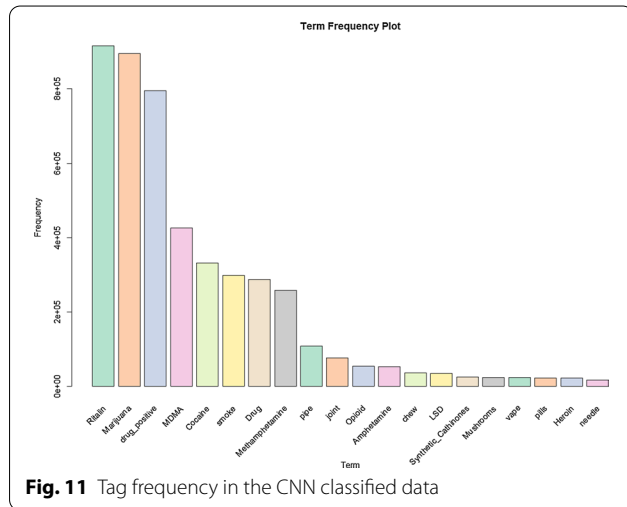
Statistic	Number of tags in tweet
Min.	1.0
1st Qu.	1.0
Median	1.0
Mean	1.309
3rd Qu.	2.0
Max	11.0

Table 7 Statistics for mined association rules

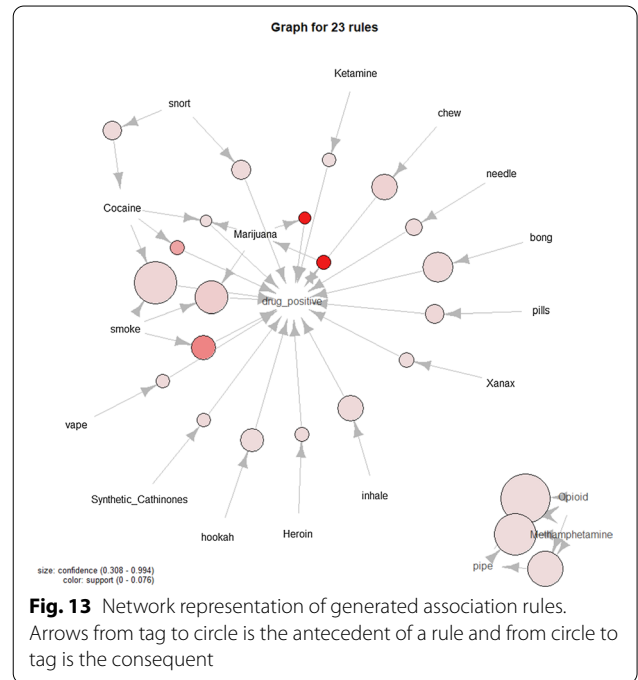
Statistic	Support	Confidence	Lift	Count
Min.	0.0003728	0.3081	1.433	1378
1st Qu.	0.0006541	0.3482	1.591	2418
Median	0.0020700	0.4359	2.063	7651
Mean	0.0116810	0.5090	6.202	43175
3rd Qu.	0.0048150	0.6037	3.096	17797
Max.	0.0762532	0.9945	58.176	281843

Table 8 Top 5 mined association rules by confidence

Rule	Support	Confidence	Lift	Count
Opioid, pipe → methamphetamine	0.0005367749	0.9944862	14.230069	1984
Cocaine, smoke → drug_positive	0.0041372791	0.8685182	4.040260	15292
Methamphetamine, pipe → opioid	0.0005367749	0.8562797	58.175785	1984
Methamphetamine, opioid → pipe	0.0005367749	0.7342709	25.161546	1984
Marijuana, smoke → drug_positive	0.0089966611	0.6865773	3.193889	33253



and methamphetamine is likely part public concern about the problem. Relations between “marijuana” and “cocaine” in regards to “drug_positive”, indicates that these are the most frequent drugs that Twitter users partake in, and are willing to discuss on the platform.



An extension to this is that “smoke” is involved in both of these relations indicating that it is the preferred vehicle for delivery of these substances.

Discussion

The purpose of this work was to classify drug-related tweets and extract feature related information from the results. We achieved a high classification rate while implementing a novel methodology in adding synthetic data to the training process. As it stands, the results have shown that simple analysis without the CNN is limited and possibly flawed. Regarding the CNN, as there are an infinite number of possible misspellings, the OOV problem cannot be eliminated. However, a character-level CNN [38] can be used to match some of the OOV words to the most similar words in the word2vec vocabulary, thus reducing the amount of OOV words. Furthermore, as manually labeling drug-related training data is very expensive, the neural network model can be pre-trained

on some large and easy-to-reach dataset (such as the Twitter dataset for negative/positive sentiment classification). A smaller learning rate can then be implemented to train the neural network model on the current dataset.

It should be noted that this research may possess a limitation due to data adequacy and societal biases. While too many regions have legalized certain substances, there is still a stigma behind drug usage. As such, social media users may be less inclined to discuss this topic, and to the contrary, may simply speak about it negatively. This is important to specify as our data may be either over or under the true number of actual drug-users. While our results seem to show a correlation between actual drug usage (see Fig. 8), there may be a further verification study needed to test the validity of utilizing social media for this task. Additionally, our current methodology does not consider features such as emojis and hashtags. This was an issue with the pre-trained Twitter word2vec model, which did not include these special characters. Lastly, conventional machine learning algorithms rely on feature engineering and simply applying word embeddings to these models may limit their performance.

Conclusion

This research aims to use machine learning algorithms to identify the mention of drug use on the Twitter platform. We have demonstrated that the CNN model has the predictive capabilities if additional synthetic data is used. It should be noted there is a limitation with the current model, which is due to a lack of diversified data and a constraint from physical labelling. The model was trained for a particular period when certain drugs were more prevalent. As an example, the Canadian government had legalized Marijuana, making is highly discussed topic. More training data will be needed, and possible changes in the way people post may need to be considered in the future. Regardless, the model was accurate given the input set and even matched actual drug use trends in society. This may exhibit the possibility of using social media more often as a surveying tool, obtaining metrics and real behavioral trends over performing separate manual measurements.

Subword embedding was not considered for this work [39] and will most likely prove useful where the morphological structures of a particular term hold importance. This methodology will be considered in future work to improve the model. Though our results are promising it may be worth considering other state-of-the-art methods such as transformers, ELMo, and Bi-LSTM. We will also explore training BERT model for word-level and sentence-level embedding on the Twitter corpus. Additionally, due to the exponential growth of social media data,

real-time data processing is essential in practice [40]. Providing solutions to the challenges such as dynamic updates in the training dataset and the filtration of spam tweets [7] is the next step.

Abbreviations

CNN: Convolutional neural network; SVN: Support vector machine; ATAM: Ailment topic aspect mode; PREDOSE: Prescription drug abuse surveillance platform; CME: Compositional multiview embedding; RNN: Recurrent neural network; DEA: Drug enforcement agency; PCA: Principal component analysis; OOV: Out of vocabulary; AUC: Area under the curve; ROC: Receiver operating characteristic.

Acknowledgements

The authors would like to sincerely thank Punardeep Sikka, Zainab Kazi, Mohiuddin Qudar, Mannila Sandhu, and Tanvi Barot for their time in manually labelling the test and training samples employed by the CNN. Additionally, they would like recognize Longfei Zeng, Dillon Small, and the overall Lakehead DaTALab for providing support and the initial dataset. Lastly, they would like to acknowledge Caleb Pears for acting as a substance and addictions consultant, and verifying the keyword selection.

About this supplement

This article has been published as part of BMC Medical Informatics and Decision Making Volume 20 Supplement 11 2020: Informatics and machine learning methods for health applications. The full contents of the supplement are available at <https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-11>.

Authors' contributions

JT, PY, and MS conceptualized the project, developed the methodology, performed formal analysis, wrote the initial draft, and created visualizations of all the results and techniques utilized. CM and JT initiated data collection and conducted curation. VM and SC provided supervision, oversight, and edited the proceeding drafts with JT and PY. All authors have read and approved the final manuscript.

Funding

Publication costs are funded by NSERC Discovery Grant (RGPIN-2017-05377), held by Dr. Vijay Mago.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due Twitter's Developer Agreement and Policy that the distribution of more than 1,500,000 Tweet IDs to any entity (inclusive of multiple individual users associated with a single entity) within any given 30 day period is forbidden. The restriction is lifted if this is done on behalf of an academic institution for the sole purpose of non-commercial research [21]. In order to conform to this, please submit a request to the following link—http://www.datalab.science/request_du_dataset.php.

Ethics approval and consent to participate

Dr. Vijay Mago has been granted approval to conduct this research by the Lakehead University Research Ethics Board (FWA00012950). There is no personally identifiable data (biomedical, clinical, or biometric) being collected from the participants in this research. Therefore, consent to participate is not required for this publication.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 November 2020 Accepted: 16 November 2020
Published: 30 December 2020

References

- Johnson T. Sources of error in substance use prevalence surveys. *Int Schol Res Not*. 2014. <https://doi.org/10.1155/2014/923290>.
- Sarker A, O'Connor K, Ginn R, Scotch M, Smith K, Malone D, Gonzalez G. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug Saf*. 2016;39(3):231–40.
- Gittelman S, Lange V, Crawford CAG, Okoro CA, Lieb E, Dhingra SS, Trimarachi E. A new source of data for public health surveillance: Facebook likes. *J Med Internet Res*. 2015;17(4):98.
- Kim A, Miano T, Chew R, Eggers M, Nonnemaker J. Classification of Twitter users who tweet about e-cigarettes. *JMIR Public Health Surv*. 2017;3(3):63.
- Shah N, Srivastava G, Savage DW, Mago V. Assessing Canadians health activity and nutritional habits through social media. *Front Public Health*. 2019;7:1.
- Du J, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, Xu H. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Med Inform Decis Mak*. 2018;18(2):43.
- Robinson K, Mago V. Birds of prey: identifying lexical irregularities in spam on Twitter. *Wirel Netw*. 2018;2018:1–8.
- Kim Y. Convolutional neural networks for sentence classification. Preprint. 2014; [arXiv:1408.5882](https://arxiv.org/abs/1408.5882).
- Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Preprint. 2013; [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- Lamos V, De Bie T, Cristianini N. Flu detector-tracking epidemics on Twitter. In: Joint European conference on machine learning and knowledge discovery in databases. London: Springer; 2010. p. 599–602.
- Paul MJ, Dredze M. You are what you tweet: analyzing Twitter for public health. In: 5th International AAAI conference on weblogs and social media; 2011.
- Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE*. 2010;5(11):14118.
- Heavilin N, Gerbert B, Page J, Gibbs J. Public health surveillance of dental pain via Twitter. *J Dent Res*. 2011;90(9):1047–51.
- Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in Twitter. In: 8th international AAAI conference on weblogs and social media; 2014.
- Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, Anand G, Carlson R, Watkins KZ, Falck R. Predose: a semantic web platform for drug abuse epidemiology using social media. *J Biomed Inform*. 2013;46(6):985–97.
- Kursuncu U, Gaur M, Lokala U, Illendula A, Thirunarayan K, Daniulaityte R, Sheth A, Arpina IB. What's UR type? Contextualized classification of user types in marijuana-related communications using compositional multiview embedding. In: 2018 IEEE/WIC/ACM international conference on web intelligence (WI). New York: IEEE; 2018. p. 474–9.
- Huang X, Di Lorio S, Dinh T, Chun SA. Deep self-taught learning for detecting drug abuse risk behavior in tweets. In: Computational data and social networks: 7th international conference, CSoNet 2018, Shanghai, China, December 18–20, 2018, proceedings, vol. 11280. London: Springer; 2018. p. 330.
- Serrat O. Social network analysis. In: Knowledge solutions. London: Springer; 2017. p. 39–43.
- Sawhney R, Manchanda P, Mathur P, Shah R, Singh R. Exploring and learning suicidal ideation connotations on social media with deep learning. In: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis. Brussels: Association for Computational Linguistics; 2018. p. 167–75.
- Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, Chile 2015. p. 959–62.
- Twitter: Developer Agreement and Policy. <https://developer.twitter.com/en/developer-terms/agreement-and-policy.html#f-be-a-good-partner-to-twitter>. Accessed 25 May 2018.
- Agency DE. Drugs of abuse: A DEA resource guide. US Department of Justice; 2017.
- Agency DE. Slang terms and code words: a reference for law enforcement personnel. US Department of Justice; 2016.
- Bian J, Yoshigoe K, Modave F. Mining Twitter to assess the public perception of the "Internet of Things". *PLoS ONE*. 2016;11(7):e0158450.
- Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS ONE*. 2013;8(7):67863.
- McHugh M. Interrater reliability: the kappa statistic. *Biochemia medica* : Casopis Hrvatskoga društva medicinskih biokemičara / HDMB 22:276–82; 2012. <https://doi.org/10.11613/BM.2012.031>.
- Fleiss JL, et al. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378–82.
- Ma J, Gao W, Wong K-F. Rumor detection on Twitter with tree-structured recursive neural networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), pp. 1980–1989. Association for Computational Linguistics, Melbourne, Australia; 2018. <https://doi.org/10.18653/v1/P18-1184>. <https://www.aclweb.org/anthology/P18-1184>.
- Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining. New York: ACM; 2016. p. 785–94.
- Godin F, Vandersmissen B, De Neve W, Van de Walle R. Multimedia lab @ acl wnut ner shared task: named entity recognition for Twitter microposts using distributed word representations. In: Proceedings of the workshop on noisy user-generated text. Association for Computational Linguistics, Beijing; 2015. p. 146–53.
- Chaturvedi I, Cambria E, Poria S, Bajpai R. Bayesian deep convolution belief networks for subjectivity detection. In: 2016 IEEE 16th international conference on data mining workshops (ICDMW). New York: IEEE; 2016. pp. 916–23.
- Kingma D, Ba J. Adam: a method for stochastic optimization. In: International conference on learning representations 2014.
- Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th international conference on artificial intelligence and statistics; 2010. p. 249–56.
- Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. Preprint. 2018; [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- SAMHSA: 2017 National Survey on Drug Use and Health (NSDUH). US Department of Health & Human Services; 2018.
- Belyi E, Giabbanelli PJ, Patel I, Balabhadrapathruni NH, Abdallah AB, Hameed W, Mago VK. Combining association rule mining and network analysis for pharmacovigilance. *J Supercomput*. 2016;72(5):2014–34. <https://doi.org/10.1007/s11227-016-1714-y>.
- Li L, Shang Y, Zhang W. Improvement of hits-based algorithms on web documents. In: Proceedings of the 11th international conference on world wide web. WWW'02. Association for Computing Machinery, New York, NY, USA; 2002. p. 527–35. <https://doi.org/10.1145/511446.511514>.
- Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Advances in neural information processing systems. NIPS'15 Proceedings of the 28th international conference on neural information processing systems, Montreal, Canada; 2015. p. 649–57.
- Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist*. 2016;5:135–46.
- Shah N, Willick D, Mago V. A framework for social media data analytics using elasticsearch and Kibana. *Wirel Netw*. 2019;2019:1–9.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.