The usage of AI in detecting substance abuse-related language on social media

Ayaan Bhatt

Received May 16, 2023
Accepted August 01, 2023
Electronic access August 31, 2023

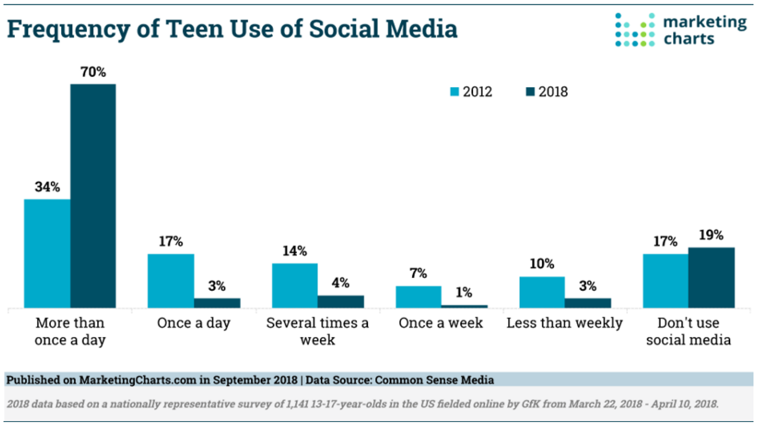
Social media has emerged as one of the largest hotbeds for exposure to substance abuse, and the alarming reality is that many platforms remain largely unregulated. Consequently, the digital landscape poses a significant threat to the well-being of children and young adults. Particularly concerning is the premature exposure to substance-related content facilitated by the widespread use of virtual public platforms such as Twitter among teenagers. Acknowledging this critical issue, in this paper, we propose an enhanced approach to flagging substance abuse-related content on these platforms, leveraging the power of a language-processing word classification model. Through extensive experimentation and analysis, this study demonstrates the potential of employing artificial intelligence (AI) techniques to combat the proliferation of substance abuse-related language on social media. Specifically, our proposed model showcases remarkable accuracy in identifying words where characters have been replaced by symbols or special characters, achieving an impressive accuracy rate of 99.3%.

By utilizing AI, we strive to empower social media platforms and regulators to proactively address the challenges posed by substance abuse within their digital ecosystems. The significance of this research lies in its potential to protect vulnerable populations and foster a safer online environment for children and young adults. Furthermore, this paper contributes to the growing body of knowledge on the utilization of AI in detecting and mitigating the harmful effects of substance abuse-related language on social media.

Introduction

Substance abuse, defined as the excessive misuse of prescription or over-the-counter drugs or the use of illegal drugs[1], is a major problem for youth. While youth substance abuse issues are not new, social media has created new ways for teenagers to be introduced to and potentially even access dangerous substances, like narcotics. Peer pressure and misinformation on social media platforms make teenagers incredibly susceptible and vulnerable to information they intake online. Through platforms like Snapchat, Instagram, and Facebook, teenagers are exposed to idealised portrayals of drug usage by ordinary people and celebrities, which creates cultural pressure to consume drugs themselves[2]. A poll conducted by Columbia University's National Center on Addiction and Substance Abuse, with 2,000 teen respondents, found that teens who used social media regularly had a higher likelihood of abusing substances than those with less-frequent usage. More specifically, cigarettes were 5 times more likely to be purchased, drinking was 3 times as likely, and using marijuana was 2 times as likely. This problem becomes more pronounced when noting that social media use amongst teens is incredibly pervasive, with 70% of teens claiming they use social media daily and 92% of these users checking social media websites more than once a day[3]. This is especially true of Facebook, Instagram and Snapchat, which are leading social media platforms for adolescents. Facebook, Instagram, and Snapchat are young adults' top social media platforms. Twitter is a platform with the majority of the content being in text form, and the users stay relatively anonymous. Adolescents being able to view unfiltered content is hence a significant problem, and there must be a solution that can adapt to users' creativity regarding circumventing conventional text filters. One study suggests participants who were exposed to tobacco content on social media, compared with those who were not exposed, had greater odds of reporting lifetime tobacco use, past 30-day tobacco use, and susceptibility to use tobacco[4]. In another study, the findings show a strong relationship between online exposure to content depicting risky behavior and users' own engagement in risky behavior in the offline environment, suggesting that content on social media may influence behavior[5].

Natural language processing (NLP), a branch of machine learning, is one that has emerged and grown spontaneously over time. Natural language processing examines the meaning of natural language such as English for computers to process it[7]. Text categorization, sentiment analysis, summarization, and text clustering are all applications of natural language processing. In this work therefore, we show how NLP can be used to identify words having characters replaced by symbols/special characters (for eg: "snort" can be written as



Published on MarketingCharts.com in September 2018 | Data Source: Common Sense Media

2018 data based on a nationally representative survey of 1,141 13-17-year-olds in the US fielded online by GfK from March 22, 2018 - April 10, 2018.

Fig. 1 This graph shows the Frequency of Teen Use of Social Media, where we can see that only 19% of surveyed teenagers do not use social media at all. This value is in stark contrast to the 70% of teenagers who use social media more than once a day. This simply serves to show the very high usage of social media amongst today's teens[6].

Table 1 Literature Review in brief

Paper Title	Problems
Identifying substance use risk based on deep neural networks and Instagram social media data	1. Only calculates the "risk" of substance abuse in the person posting 2. Does not factor in the "encryption" of words
An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning	1. Makes extensive use of human classifiers, hence taking longer 2. Very useful, but only for statistics 3. Does not factor in the "encryption" of words

"$n0rt" to avoid detection by conventional filters). Text collection, text preprocessing, word embedding, and machine learning modelling are the four processes in this technique, all four of which are used in this paper. The objective of my research is to come up with a model that is able to "decode" specific words used on social media, that would not be detected/flagged otherwise. This would allow parents/guardians to have a higher degree of control over what content their ward can see.

The rest of the paper is organized as follows, we describe the literature review followed by the methodology and model, then a discussion about the results and the conclusion of this study.

Literature Review

Identifying substance use risk based on deep neural networks and Instagram social media data[8]

This paper presents a similar approach for assessing substance use risk using deep neural networks and data from Instagram. The study explores the potential of using machine learning techniques to analyze user-generated content, particularly images and captions, to identify patterns indicative of substance use risk. By leveraging the vast amount of publicly available data on Instagram, the researchers demonstrate the feasibility of predicting substance use risk levels with promising accuracy. This innovative application of deep neural networks to social media data could have significant implications for early intervention and targeted prevention efforts in the context of substance use disorders. My paper aims to determine what kind of content must be flagged and filtered, rather than calculating the risk of substance abuse. I focus on preventing such content from reaching a young audience, hence preventing the likelihood of misinformation/drug glorification reaching the same.

An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning[9]

This paper presents a comprehensive investigation into drug-abuse risk behavior detection using Twitter data and self-

Table 2 Bad words

Number of characters injected with noise	Accuracy	F1 Score
0	99.81%	0.9980
1	82.66%	0.6598
2	82.66%	0.6598
3	82.66%	0.6598
4	82.66%	0.6598
5	82.66%	0.6598
6	82.66%	0.6598
7	82.66%	0.6598
8	82.66%	0.6598
9	82.66%	0.6598

taught deep learning techniques. The study aims to gain valuable insights into drug-related discussions and patterns on the platform. Leveraging the power of deep learning models, the authors developed an innovative approach to automatically detect and analyze drug-abuse risk behavior in users' tweets. Their findings shed light on prevalent drug-related trends and risk factors on Twitter, offering potential implications for public health interventions and preventive strategies to address drug abuse in the digital space. However, their model, albeit providing more information, does not factor in "noise" in Twitter posts, including but not limited to words with letters swapped out for symbols.

Methodology and Model

The texts to be processed are collected in the first phase (text collection). The data was first collated from different sources[10,11], due to a lack of exact datasets online. Both sets of data were sorted into "bad" or "good" words and assigned labels 0 and 1 respectively. The Keras API and the Pandas package for Python were used to create the actual model as well as clean the data respectively. The second phase (text preprocessing) entails standardising unstructured texts to improve the accuracy of natural language processing. The material gathered contains several difficult-to-understand characteristics, such as mistakes, emojis, abbreviations, and freshly invented terms. The majority are conveyed as if conversing carelessly in terms of language or sentence structure. As a result, preprocessing is conducted according to the need following text processing, which includes transforming the uppercase to lowercase, erasing special characters and emoticons, and text normalisation such as word tokenization and stop word removal. The words are turned into vectors in the third phase (word embedding) so that computers can effectively grasp and analyse natural language. Eventually, a supervised learning model is built in the machine learning modelling step, and training and prediction are conducted utilizing vectorized number-type data. To increase precision, an F1 score is calculated. The F1 score is a machine learning assessment statistic that gauges the accuracy of a model. It combines a model's accuracy and recall scores. The F1 score is used frequently for machine learning classification tasks. In particular, in cases where data is unbalanced i.e. there are a significantly higher number of observations of one class of objects than others, it is often viewed as a better performance measure than metrics like accuracy. In this context, this is seen in this case when there is quite a large difference between the number of "good" and "bad" words.

The code is a Python script that performs text classification using deep learning. The script imports libraries like Pandas, NumPy, TensorFlow, and scikit-learn. It then defines several functions that process data, train models, and test performance. The data used in the script is read from a CSV file named 'GoodBadWords.csv', containing two columns - one column contains bad words and another column contains good words. The script reads this data and combines the bad and good words into a single array. The 'words class' array is then created, which contains labels for each word, i.e., 0 for bad words and 1 for good words. The 'data processing' function returns a Pandas DataFrame with two columns - 'all words' and 'class words'.

The 'custom standardization' function is used to preprocess the text data by removing punctuation and HTML tags and converting all text to lowercase. The 'TextVectorization' class from TensorFlow is used to vectorize the text data.

The 'vectorize text' function is defined to perform vectorisation on the text data. This function takes as input the text to be vectorized, the vectorization layer, and a flag 'train' to indicate whether the function is called during training or testing. The 'train' flag is used to adapt the vectorization layer during training.

The 'recall m', 'precision m', and 'f1 m' functions are used to compute recall, precision, and F1 score, respectively. These functions are defined using TensorFlow functions. The 'model

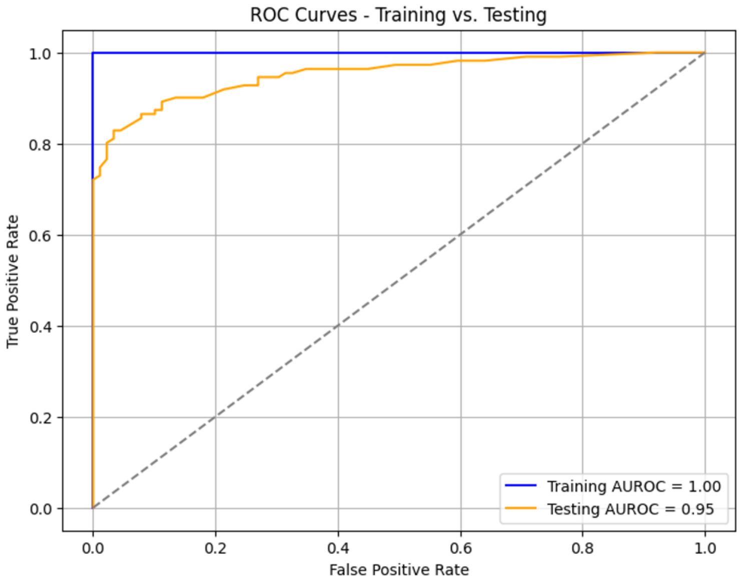


Fig. 2 An "Area Under the Reciever Operating Characteristic" value of 0.95 signifies a 95% accuracy when calculating the number of "true positives" (values that have been classified correctly by the model). The small gap between the Training AUROC curve and the Testing AUROC curve shows the absence of a large amount of overfitting.

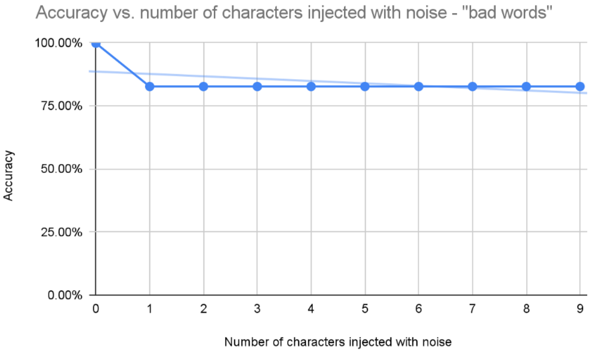


Fig. 3 Accuracy vs number of characters injected with noise – "bad" words. As noise increases, the number of words available decreases, causing a decrease in accuracy, as seen above. The model still maintains a relatively high value.

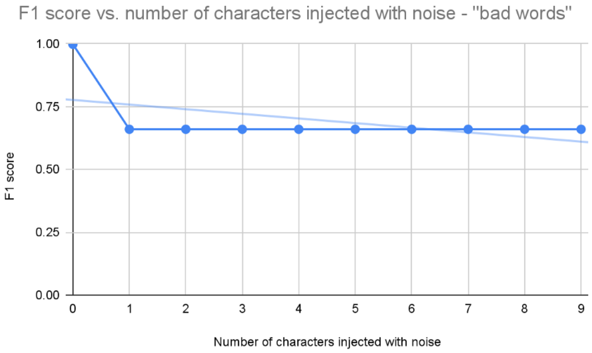


Fig. 4 Score vs number of characters injected with noise – "bad" words. As noise increases, the number of words available decreases, causing a decrease in the F1 score, as seen above. The model still maintains a relatively high value.

sentence classification' function defines the neural network model for text classification. This function takes as input the data returned by the 'data processing' function. The function first one-hot encodes the labels and then vectorizes the text data. The neural network model is then defined, which consists of an embedding layer, convolutional layers, a global max pooling layer, a dense hidden layer, and an output layer. The model is then compiled using categorical cross-entropy loss, the Adam optimizer, and the 'f1 m' metric. The model is trained for 50 epochs on the vectorized text data and one-hot encoded labels. The 'test function' function is used to test the performance of the trained model. This function takes as input the trained model and a test dataset. The test dataset is vectorized using the same vectorization layer used during training. The one-hot encoding is also performed on the test labels. The model is then used to make predictions on the test dataset, and the accuracy, F1 score, and precision-recall-F1 score are computed using scikit-learn functions.

The hyperparameters used in the paper were specified as follows. For the TextVectorization step, the maximum num-

Table 3 New bad words

Number of characters injected with noise	Accuracy	F1 Score
0	99.34%	0.9966
1	66.12%	0.6997
2	66.12%	0.6997
3	66.12%	0.6997
4	66.12%	0.6997
5	66.12%	0.6997
6	66.12%	0.6997
7	66.12%	0.6997
8	66.12%	0.6997
9	66.12%	0.6997

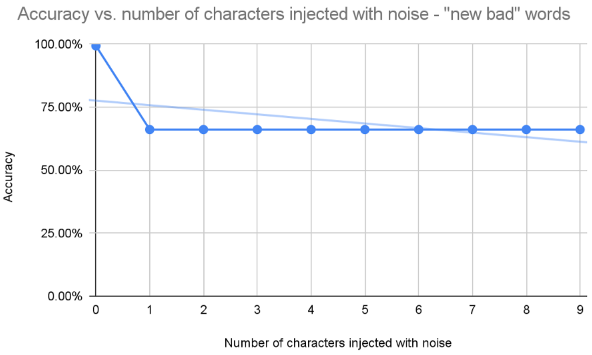


Fig. 5 Accuracy vs number of characters injected with noise – "new bad" words. As noise increases, the number of words available decreases, causing a decrease in accuracy, as seen above. Overall lower value due to a smaller dataset being available.

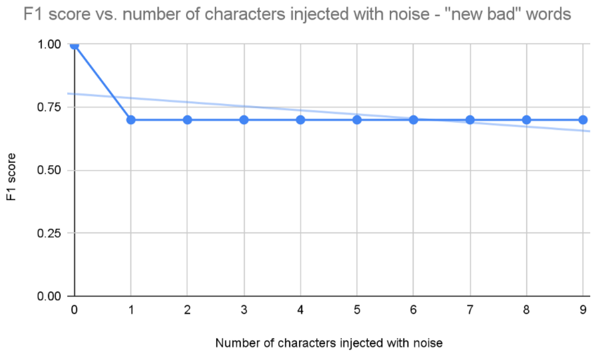


Fig. 6 Score vs number of characters injected with noise – "new bad" words. As noise increases, the number of words available decreases, causing a decrease in the F1 score, as seen above. Overall lower value due to a smaller dataset being available. Higher F1 score achieved by the model.

ber of tokens allowed was set to 20,000. The output mode was chosen to be "int," representing integer-encoded sequences. Additionally, the output sequence length was set to 500 tokens. Moving on to the Model Architecture, it consisted of an Embedding Layer with an embedding dimension of 128. The Convolutional Layers had 128 filters with a kernel size of 7, and "valid" padding was used. The activation function utilized was ReLU, and the convolutional layers had a stride of 3. The Dense Layers consisted of a First Dense Layer with 128 units and a ReLU activation, along with a dropout rate of 0.5. The Output Dense Layer had 2 units, corresponding to the number of classes in the classification task, and used the softmax activation function. For the Training Parameters, the loss function employed was "categorical crossentropy," and the optimizer utilized was "adam." The metrics used for evaluation were accuracy and the f1 score. The model was trained for a total of 50 epochs.

Overfitting is the phenomenon where a model performs well on the training data but fails to generalize to new, unseen data[12]. This model tries to prevent overfitting by implementing dropout layers, such as layers.Dropout(0.5). The dropout rate was set to 0.5, meaning that during training, 50% of the neurons in the respective layers were randomly dropped out, which helps to reduce overfitting. This introduces noise and reduces the reliance of the model on specific features.

Results

The accuracy of recognising "bad" words with noise injected into them decreased as the amount of noise increased, and after a point showed no change. Specifically, the accuracy and the F1 score went from near perfect 99.8% and 0.998 at minimum noise to 82.7% and 0.660 at the highest amount of noise of 9, as seen in Fig. 3 and Fig. 4. There were lower accuracies and F1 scores at higher noise due to the decrease in the number of words available at those higher noise values.

The same trend was seen with the "new bad" words, which consisted of just the words related to substance abuse. Due to a smaller dataset size, the maximum accuracy and F1 score (seen at the lowest noise value of 0) were slightly worse off, at 99.3% and 0.997. A larger dip compared to the "bad" words was seen in the minimum accuracy at 66.1%. However, the F1 score marginally increased to 0.700 when the same is compared here. This is seen in Fig. 5 and Fig. 6.

In a real-life situation, most "bad"/"new bad" words will have a median amount of noise, but the entire range must be accounted for. The most common existing algorithms for overall text filtering include adaboost, logistic regression and support vector machine. These algorithms show accuracies of 44.6%, 49.6%, and 0.8%. My model follows no existing, commonly used framework but shows a higher degree of accuracy of 99.3% when looking at the words related to substance abuse.

Conclusion

The original goal of creating a model that has the ability to identify words that had been typed using special characters had been achieved. In this work/paper, we have built a model capable of identifying noisy words relating to bad words and substance abuse with a maximum accuracy of 99.8% and 99.3%, respectively. Implementation of a similar model on social media platforms can greatly reduce the extent to which adolescents are exposed to usually flawed opinions of substance abuse on social media. This can prevent the spread of misinformation regarding the same and can reduce the exposure to the glorification of it as well. In the future, the basis of this model can be used to include emojis as noise, as this model only looks at special characters. This can be done by using image processing on the emoji to recognise the letter it represents, which in turn can be used to decide if the word should be "flagged" or not. The accuracy of this model can be improved upon as well to help prevent "good" words from being incorrectly classified. Also, the context of the social media post/message can be taken into account. There are many anti-substance abuse stories on these platforms which make use of words that we have classified as "new bad". This can lead to incorrect flagging of this content.

Acknowledgements

I extend my heartfelt gratitude to Dr Kehinde Owoeye for his invaluable guidance, unwavering support, and insightful feedback throughout the journey of writing this research paper. His expertise and dedication have been instrumental in shaping the direction of this study, and I am truly grateful for the opportunity to benefit from his mentorship. This research would not have reached its fruition without his patient encouragement and thoughtful insights.

References

1 N. C. Institute, *NCI Dictionary of Cancer Terms*, `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/drug-abuse.`, Available at:.

2 D. Romer and M. Moreno, *Pediatrics, [online*, **140**, 102– 106.

3 J. Hilliard, *The Influence of Social Media on Teen Drug Use*, `https://www.addictioncenter.com/community/social-media-teen-drug-use/.`, online] AddictionCenter. Available at:.

4 S. Donaldson, A. Dormanesh, C. Perez, A. Majmundar and J.-P. Allem, *JAMA Pediatrics*, **176**, 878.

5 D. Branley and J. Covey, *Computers in Human Behavior*, **75**, 283–287.

6 M. Charts, *Frequency of Teens' Social Media Use, 2018 vs*, `https://www.marketingcharts.com/charts/frequency-teens-social-media-use-2018-vs-2012.`, online] Available at:.

7 Oraclecom, *What is Natural Language Processing?*, `https://www.oracle.com/in/artificial-intelligence/what-is-natural-language-processing/#:~:text=Natural%20Language%20Processing%3F-.`, Available at:.

8 S. Hassanpour, N. Tomita, T. DeLise, B. Crosier and L. Marsch, *Neuropsychopharmacology, [online*, **44**, 487–494.

9 H. Hu, N. Phan, S. Chun, J. Geller, H. Vo, X. Ye, R. Jin, K. Ding, D. Kenne and D. Dou, *Computational Social Networks*, **6**, year.

10 GitHub, *Our List of Dirty, Naughty, Obscene, and Otherwise Bad Words*, `https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en.`, Available at:.

11 R. Kumar, *positive words*, `www.kaggle.com.`, Available at:.

12 GeeksforGeeks, *Underfitting and Overfitting in Machine Learning*, `https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/.`, Available at:.