International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228146

Inexpensive Detection of Substance Abuse Based on Social Media Data using Machine Learning

Abhinav Potineni
Academies of Sciences, Leesburg, Virginia, USA

Article Info
Volume 8, Issue 2
Page Number : 01-08

Publication Issue :
March-April-2022

Article History
Accepted: 01 March 2022
Published: 05 March 2022

ABSTRACT

Over the past few years, substance abuse has become one of the most severe public health problems in the United States. The annual cost of substance abuse aftereffects in the United States alone is approximately $3.73 Trillion. The societal costs of substance abuse include premature deaths, lost productivity, and increased crime rates. Unfortunately, many victims, especially in lower-income families, don't have access to early detection and early family intervention tools due to limited access to traditional diagnostic tools and rehab specialists. Currently, there is no complete diagnostic pipeline to inexpensively detect substance abuse and automatically inform family members or trusted contacts. To combat this, the experimenter developed the SOS 280 system, which utilizes machine learning techniques in a smartphone application. SOS 280 works through social media monitoring and automatic notification using SMS and GPS location. The SOS280 algorithm primarily uses social media data, namely publicly available Twitter, and Instagram posts, to identify substance abuse-related activity. The experimenter collected and classified data by applying for the Twitter and Instagram Developer API Platforms, mining tweets and posts with specific drug keywords present. The investigator trained a Natural Language Processing (NLP) text classification model to analyze the sentiments on the tweets, then classifying them as positives (containing substance abuse-related keywords) and negatives. The master model is a Bidirectional Encoder Representations (BERT) derivative that uses a transformer-based architecture to detect emotions in sentences and conversations to classify substance abuse instances. In total, the researchers looked at 55,551 tweets and Instagram posts indicative of potentially alarming substance usage. Finally, the experimenter developed a smartphone application to capture trusted contact information and GPS location, send data to a remote server housing the neural network, output the network's detection, and send automated alerts to trusted contacts via SMS and GPS location. The experimenter further validated the system's effectiveness through a partnership with national nonprofit Faces and Voices of Recovery, which works with

23 million addiction recovery victims. SOS280 is an inexpensive, reliable, easy to use, and timely tool for families of young adults in predicting substance abuse.

Keywords : Substance Abuse, Drug Abuse, Natural Language Processing, Social Media Mining, Twitter, Instagram, Bidirectional Encoder Representations, Linguistic Analysis, Language Modeling, Sentimental Analysis, Early Diagnostic Systems

I. INTRODUCTION

The United States is currently going through a substance abuse epidemic, which leads to numerous avoidable deaths in the United States [1]. The use of illicit drugs causes over 100 deaths each day, ahead of motor vehicle accidents as the leading cause of injury deaths. According to a study by the National Center for Drug Abuse, 788,000 teenagers aged 12 to 17 years old met the requirements of an Illicit Drug Use Disorder (IDUD). Various institutions like The Substance Abuse and Mental Health Services Administration have classified substance abuse as a preventable disease. The proposed early intervention is an effective strategy to reduce the impact of substance use and mental disorders in American communities [2]. Of the 15.1 million adults aged 26 or older (1 in 14) that needed substance abuse treatment, only 3 million received treatment. Barriers to adequate support for patients suffering from substance abuse disorders include a lack of early family intervention resources and the general social stigma associated with substance abuse disorders [3]. More importantly, the lack of adequate research is also attributed to inaccurate assessments and the lack of early diagnosis systems. Therefore, it is quintessential to develop early diagnostic tools and techniques to intervene to prevent further escalation. To solve the issue of a lack of systems for early family intervention, the SOS280 team developed the novel, inexpensive, social media-based substance abuse detection smartphone system that offers a two-pronged solution. First, it brings family intervention early into potential substance abuse situations, thereby helping seek early substance addiction treatment. Second, it provides a one-click support button for someone struggling with a potential substance abuse situation. Instead of interventions at the Risky Usage stage of the cycle of addiction, which can cost upwards of $60,000, SOS280 provides intervention at the Initiation & Experimentation phase [4]. With the advent of smartphones, content sharing and behavior sharing have become one click away. Social media presents an excellent opportunity for diagnosis as teenagers and young adults increasingly share information, seek ideas, and interact with their peers through social media [5]. In 2020, 3.6 billion people worldwide used social media, which is projected to reach 4.41 billion by 2025. The SOS280 algorithm uses data from social networking sites such as Instagram and Twitter, particularly popular among younger audiences.

II. SCIENTIFIC PREMISE

The current massive Twitter & Instagram user-bases, combined with rapid increases in new users, coupled with a highly accessible application program interface (API), make it a suitable choice for our substance abuse prediction research [6]. This project focused on building a machine learning-based NLP method that can identify high substance use risk based on social media posts on Instagram and Twitter. The rich dataset retrieved from Twitter and Instagram platforms combined with the recent adoption of machine learning methods in medicine provides a

unique opportunity for offering a solution to the substance abuse pandemic [7]. Our proposed substance abuse prediction model in this work is based on Bidirectional Encoder Representations (BERT), a transformer language model with a variable number of encoder layers [8]. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional unlabeled text representations and is empirically robust.

Smartphone app: SOS280 used send TextMessage() of the SmsManager class to notify trusted contacts via text or GPS location. This method allows for sending messages from SOS280 app without having to use another installed app. SOS280 used an API geolocation tool, enabling the app to learn and display the location, providing an ability to integrate location-based services (LBS) into the app.

III. LITERATURE REVIEW

Researchers are increasingly utilizing machine learning to develop methods to track and monitor social media interactions for public sentiment analysis on various topics [9]. For Instance, a recent study of Twitter data to monitor user engagement with substance use found that approximately 1 in every 2000 tweets sent was about marijuana [10]. It was stated in the national opioid brief that the death counts shifted from opioids for pain to heroin and then to synthetic opioids [11]. These examples suggest social media platform data is a vast data reserve for accurate sentiment analysis and tracking substance abuse. Substance abuse follows a pattern of seven stages, starting from a person's first use and leading to addiction itself [12]. These seven stages are Initiation, Experimentation, Regular Usage, Risky Usage, Dependence, Addiction, and Crisis/Treatment. Research shows that the best way to help someone with potential substance use risk is early family intervention during the Initiation and Experimentation stage. While all previous studies help understand the substance abuse trends, no study to date has been conducted to 1) Predict substance abuse at early stages of abuse life cycle 2) Provide early family intervention. Moreover, there is a lack of an easily accessible, cheap, and effective early diagnostics system for substance abuse, and no research has yet attempted to develop such a diagnostics system [13]. Therefore, with the SOS 280 system, the team aimed to: 1) Use the Natural Language Processing Machine Learning model for real-time monitoring and detection of substance abuse based on social media posts 2) Develop a novel hybrid app-based system to offer support by notifying the trusted contacts via text and sharing GPS location with one click of a button 3) Implement data Confidentiality through Encryption and Consent based access.

IV. METHODS

A. Data Collection

A total of 128,348 tweets and 56,829 Instagram posts were collected from the years 2008 through 2021 through the Twitter API and Instagram API. After discarding the non-English rows, the cleaned dataset ended up with 121,432 tweets and 54567 posts. The rows were classified again into Tweets and Posts having multiple keyword occurrences and Tweets and Posts having single actual drug name and/or drug street name. To generate the training dataset the experimenter randomly selected 2401 Tweets and Posts and manually labeled the class and generated 8000 row synthetic data by keyword replacement. The experimenter ended up with 55, 551 training rows and 8000 testing rows [14]. The experimenter implemented a machine learning-based Natural Language Processing (NLP) technique and Bidirectional Encoder Representations from Transformers to estimate the sentiment score for each tweet.

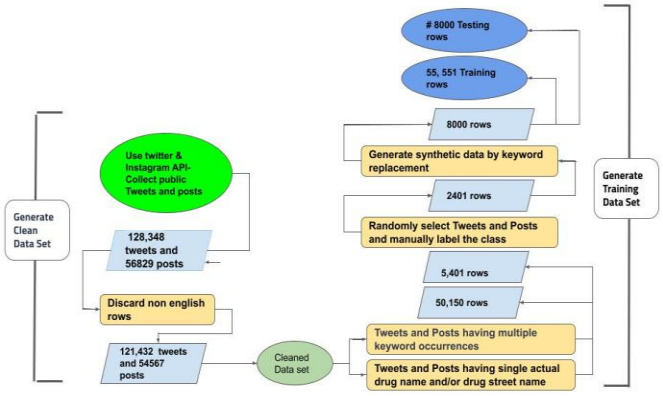


Figure 1: Data collection flowchart

B. Features extraction and Text classification

To identify the substance use risk in Instagram and Twitter users, the experimenter analyzed each user's posts and tweets to extract indicative features from posted tweets and posts. The experimenter used a BERT model with three input embeddings - token embeddings, segment embeddings, and position embeddings. Token embedding is a data-driven tokenization strategy that brings a balance of vocabulary size and out-of-vocab words. New words are generated by combining two essential elements in the Twitter and Instagram posts and adding them to the crucial elements. Segment embeddings are the sentence numbers encoded into a vector needed for the substance abuse model to know whether a particular token belongs to sentence A or B in BERT. Segment embeddings helped add the positional encodings to the existing word embedding, giving the final pre-processed embedding used in the encoder part. In the transformation layer, BERT learned multiple attention mechanisms, called heads, which operate parallel to one another, enabling the model to capture a broader range of relationships between words. The transformation layer also has an "Add & Norm" step in which the "add" aspect refers to a residual connection that adds the input of each layer to the output, and the "norm" aspect refers to Layer Normalization.

The feed-forward layer weights are trained and applied to each token position. It is a highly parallelizable part of the model since it is applied without any communication or inference by other token positions. The experimenter used SoftMax to classify content into positive (substance abuse involved) and negative (neutral/non-substance abuse related) discussions based on probability assignments.

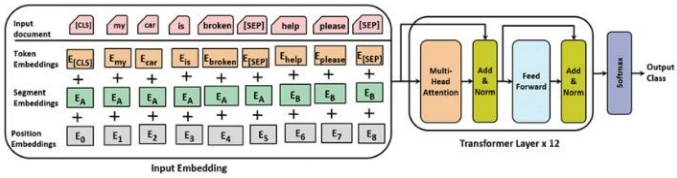


Figure 2: Architecture diagram for BERT classifier.

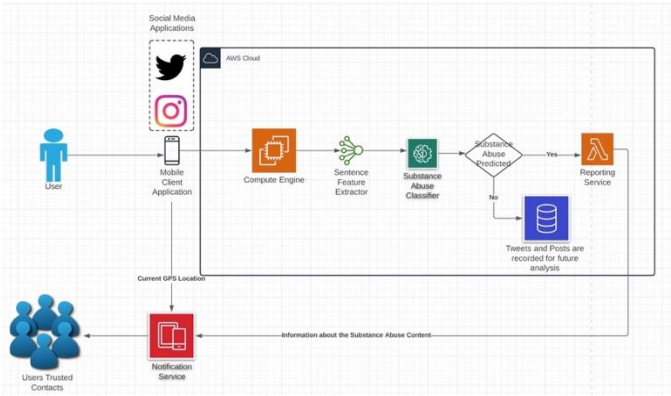


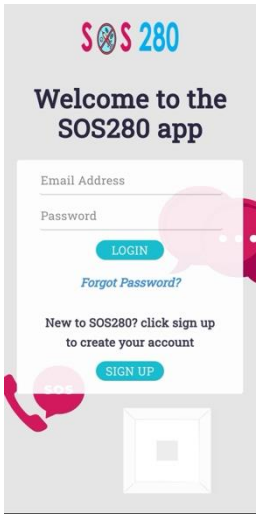
Figure 3: Overall architecture diagram of SOS 280.

V. RESULTS AND DISCUSSION

SOS 280 app is available on google play store for android [15,16]. SOS 280 users can download the app from the app store. Upon download, the user is prompted to create an account by providing a username and password. The app also asks for permission to access the location to be able to share the GPS location upon detection of substance abuse. After creating the account and logging in, SOS 280 screen will display four options - Account, Home, Monitor, and Contacts. SOS 280 users can click on the contacts button to enter the information of up to five trusted contacts. The trusted contacts are then notified (via text message) of the request. Each trusted contact must accept the request to receive notifications explicitly. An SOS 280 user or a family member can click on the Monitor button and add up

to five Twitter and Instagram IDs. The Twitter and Instagram users are asked for approval. Upon approval, the app starts monitoring the Twitter and Instagram accounts 24/7.

The SOS 280 Natural Language Processing (NLP) text classification model algorithm is a Bidirectional Encoder Representations (BERT) derivative that uses transformer-based architecture to detect emotions in sentences and conversations, analyzes the sentiments on the tweets and posts to classify them as positives (containing substance abuse-related keywords) and negatives [17]. These substance use of marijuana, cocaine, K2, heroin, crystal meth, MDMA, hallucinogens, DXM, etc. the street names like dope, percs, white, TNT and Captain Cody, etc. are monitored. As soon as SOS 280 detects a substance abuse tweet or post, it will send a text message and share the GPS location of the Twitter or Instagram owner to all the trusted contacts [18]. Additionally, users can click on the monitor button to manually send an alert to trusted contacts. This button works as a free digital panic button to get the SOS 280 user immediate support via text or GPS when they have trouble reaching out.



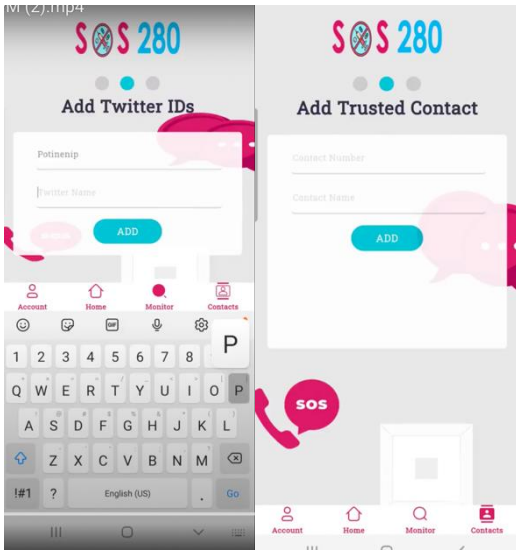


Figure 4: Screenshots of SOS280 deployed on Google play store

VI. CONCLUSION

Through this study, the experimenter clearly demonstrated that social media data could be used to build robust drug abuse risk behavior detection systems based on supervised machine learning and cloud computing techniques. Challenges came from the limited character size of Twitter data, as well as the limited availability of annotated Instagram data. The experimenter partnered with two national non-profits working in the substance abuse prevention space and have a combined reach of 26 million. In future the experimenter would like to extend this app to tackle cyberbullying and additionally include more social media and messaging platforms like Facebook and Snapchat. Facebook has 1.58 billion users, Instagram 500 million and Snapchat 100 million users. This offers a huge potential for analyzing social media behavior.

VII. FUTURE WORK

The SOS 280 team researched multiple ways of countering substance abuse on top of early family intervention and is working on incorporating four

additional features to improve the capability and reach. According to research conducted by Douglas L. Polcin et al., sober access to social support networks and de-addiction centers are crucial for substance abuse survivors to stay sober and not relapse [19]. SOS 280 plans to incorporate the research findings and improve on the app. The first additional feature is to use the geolocation API and provide SOS 280 users access to deaddiction centers & sobriety networks in the vicinity based on user GPS location [20].

As per the research done by Marya T. Schulte et al. l, being aware of the Substance Use and Associated Health Conditions throughout the Lifespan is critical to the de-addiction journey [21]. The second feature is to provide SOS280 users information on health impacts. The information on health impacts will be gathered from the substance abuse and mental health services administration's website (SAMHSA) and maintained on the SOS280 server. The SOS280 team will sync up the database with SAMHSA's website periodically to keep the information current.

Also, one of the most challenging aspects of fighting against substance abuse is relapse following adolescent substance abuse treatment. About 40-60% of individuals relapse within 30 days of leaving an inpatient drug and alcohol treatment center, and up to 85% relapse within the first year. When dealing with highs and lows in sobriety, many turn to inspirational recovery quotes as a source of strength [22]. The third feature is to provide SOS280 users with one motivation quote a day for 365 days of a year to inspire sobriety. The quotes will be procured from good reads and uploaded into the SOS280 database server. One unique quote widget is displayed every day, and additionally, the users have access to all 365 quotes on the main menu. The fourth and last feature is to provide SOS280 access to sobriety merchandise. Vendors are organized by category in the main menu (Clothes, accessories, etc.), and on-click of the vendor's name, the users are redirected to the vendor's website.

VIII. REFERENCES

[1]. Cynthia L. Rowe, "Family Therapy for Drug Abuse: Review and Updates 2003–2010, Wiley Online Library", 2021, Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-0606.2011.00280.x

[2]. Monique Bolognini, "Evaluation of the Adolescent Drug Abuse Diagnosis instrument in a Swiss sample of drug abusers", Wiley Online Library, 2002, Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1360-0443.2001.9610147711.x

[3]. Alfred S Friedman et al., "A Method for Diagnosing and Planning the Treatment of Adolescent Drug Abusers (The Adolescent Drug Abuse Diagnosis [ADAD] Instrument)", Sage Journals, 1989. Available:https://journals.sagepub.com/doi/abs/10.2190/XBU5-9MAB-C2R5-0M96

[4]. Sunny Jung Kim et al., "Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data", Journal of Medical Internet Research, 2017. Available: https://www.jmir.org/2017/10/e353/

[5]. Kevin R. Scott et al., "Opportunities for Exploring and Reducing Prescription Drug Abuse Through Social Media", Journal of addictive diseases, 2015. Available:https://www.tandfonline.com/doi/abs/10.1080/10550887.2015.1059712

[6]. Tareq Nasralah et al., "Social Media Text Mining Framework for Drug Abuse: Development and Validation Study with an Opioid Crisis Case Analysis", Journal of Medical Internet Research, 2020. Available: https://www.jmir.org/2020/8/e18350/

[7]. Mohammad Ali Al-Gradi et al., "Text classification models for the automatic detection of nonmedical prescription medication use from social media", Springer Link, 2021. Available:

https://link.springer.com/article/10.1186/s12911-021-01394-0

[8]. M S Neetu et al., "Sentiment analysis in twitter using machine learning techniques", IEEEXplore, 2013. Available:https://ieeexplore.ieee.org/abstract/document/6726818

[9]. Patty Cavazos-Rehg et al.," Characterizing the Followers and Tweets of a Marijuana-Focused Twitter Handle", Research gate, 2014. Available: https://www.researchgate.net/publication/263514457_Characterizing_the_Followers_and_Tweets_of_a_Marijuana-Focused_Twitter_Handle

[10]. Colin Planalp, MPA et al., "State health data access center, The Opioid Epidemic: National Trends in Opioid-Related Overdose Deaths from 2000 to 2017", 2019. Available:https://www.shadac.org/sites/default/files/publications/2019%20NATIONAL%20opioid%20brief%20FINAL%20VERSION.pdf

[11]. Steve Sussman, "A Lifespan Developmental-Stage Approach to Tobacco and Other Drug Abuse Prevention", Hindawi Publishing Corporation ISRN Addiction Volume 2013, 2013. Available: https://downloads.hindawi.com/archive/2013/745783.pdf

[12]. Stanton, M. Duncan et al., "Outcome, attrition, and family–couples' treatment for drug abuse: A meta-analysis and review of the controlled, comparative studies", American physiological association, 1997. Available: https://psycnet.apa.org/record/1997-05606-004

[13]. Robin Haunschild et al.,"Investigating dissemination of scientific information on Twitter: A study of topic networks in opioid publications", MIT Press direct, 2022, Available: https://direct.mit.edu/qss/article/2/4/1486/108046/Investigating-dissemination-of-scientific

[14]. Ke Wang et al.,"Sentiment Analysis of Peer Review Texts for Scholarly Papers, Research gate", 2018, Available: https://www.researchgate.net/profile/Ke-Wang-123/publication/326137575_Sentiment_Analysis_of_Peer_Review_Texts_for_Scholarly_Papers/links/5bfd005b92851cbcdd746889/Sentiment-Analysis-of-Peer-Review-Texts-for-Scholarly-Papers.pdf

[15]. Abhinav Potineni, "Link to the SOS 280 app in the Google Appstore", Google play, 2022, Available: https://play.google.com/store/apps/details?id=com.mobile.sos280&hl=en_US&gl=US

[16]. Abhinav Potineni, "SOS280 APP Functionality", Youtube, 2022, Available: https://studio.youtube.com/video/9fWM0uuIUS0/edit

[17]. Saif M. Mohammad, "An Interactive Visual Explorer for Natural Language Processing Literature" Aclanthology, NLP Scholar, 2020. Available: https://aclanthology.org/2020.acl-demos.27.pdf

[18]. Woo-Young Ahna et al., Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence, US National Library of Medicine, National Institutes of Health, 2016, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4955649/

[19]. Douglas L. Polcin et al., "What Did We Learn from Our Study on Sober Living Houses and Where Do We Go from Here?", US National Library of Medicine, National Institutes of Health, 2015. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3057870/

[20]. Debasish Basu et al., "Changing pattern of substance abuse in patients attending a de-addiction center in north India (1978-2008)", US National Library of Medicine, National Institutes of Health, 2012. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3410210/

[21]. Marya T. Schulte et al., “Substance Use and Associated Health Conditions throughout the Lifespan”, Public Health Reviews, 2013. Available: https://publichealthreviews.biomedcentral.com/articles/10.1007/BF03391702

[22]. Eric Lacayo, “10 Inspirational Quotes for Recovering Addicts”, Banyan Treatment center, 2020. Available: https://www.banyantreatmentcenter.com/2020/11/16/10-inspirational-quotes-for-recovering-addicts-boca

Cite this article as :

Abhinav Potineni, "Inexpensive Detection of Substance Abuse Based on Social Media Data using Machine Learning ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 2, pp. 01-09, March-April 2022. Available at
doi : https://doi.org/10.32628/CSEIT228146
Journal URL : https://ijsrcseit.com/CSEIT228146