

Identification of Antisocial Behavior Using Conventional Machine Learning and Deep Learning Algorithms

Poonam Mathan, Dr. Prof Anil Pimpalpure, Dr. Prashant Sen
Research Scholar, Prof & Dean, HOD

Department of Computer Science, School of Engineering, Eklavya University Damoh

Abstract

Online antisocial behavior encompasses actions that harm others in the digital realm. This can include cyberbullying, trolling, spreading misinformation, hacking as well as cyberbullying. People who might not normally participate in antisocial behavior in face-to-face conversations may feel more comfortable engaging in it when interacting online due to the anonymity and apparent distance. Addressing online antisocial behavior requires a combination of strategies, including education, awareness campaigns, and technological solutions. Platforms can implement features to mitigate harassment and abuse, while individuals can practice responsible online behavior and report inappropriate conduct. Detecting and classifying online antisocial behavior (ASB) can be challenging due to the diverse forms it can take and the nuances of online interactions. However, several approaches can be used, Keyword-Based Filtering involves detecting specific keywords or phrases associated with ASB, such as threats, insults, or derogatory language. Labeled data can be used to train machine learning algorithms so they can automatically classify online content as ASB or non-ASB. These models can analyze text, images, or videos for patterns indicative of ASB. User Behavior Analyzer monitor user behavior for patterns associated with ASB, such as frequent use of inflammatory language or repeated interactions with known ASB accounts.

Keywords: Online antisocial behavior, deep learning, machine learning, Antisocial Online Behavior, Keyword-Based Filtering, User Behavior Analyzer.

1. Introduction

ASPD personality disorders, is a complex condition with various contributing factors. While genetics, early life experiences, and environmental

factors can play a role, the exact causes of ASPD are not fully understood. Treatment for ASPD often involves psychotherapy, such as cognitive behavioral therapy, to address problematic behaviors and thought patterns. However, treatment can be challenging, and not all individuals with ASPD seek or respond to treatment.

Online antisocial behavior (ASB) presents a complex challenge due to its various forms and underlying motivations. ASB can encompass a broad variety of actions, from mild forms like trolling and cyberbullying to more severe forms such as online harassment and hate speech. These behaviors can have serious consequences for individuals and communities, including psychological harm, social isolation, and even physical harm in some cases. Addressing online ASB demands a multifaceted strategy that incorporates both technological solutions, social interventions. This can include developing and implementing effective reporting and moderation systems on social media platforms, educating users about online safety and responsible behavior, and promoting positive online norms and values. It also involves understanding the underlying factors that contribute to ASB, such as anonymity, social reinforcement, and the influence of online communities, and finding ways to mitigate these factors.

Studying patterns of behavior, identifying risk factors, and evaluating interventions, researchers can help develop strategies to prevent and mitigate online ASB. Collaborative efforts between researchers, policymakers, and social media platforms are essential to effectively address this complex challenge and create a safer and more inclusive online environment. Top of Form While reporting mechanisms are important, there seems to be a need for more proactive measures to prevent and mitigate such behavior. Developing clearer

guidelines on acceptable behavior and improving detection algorithms could help platforms intervene more effectively. Fostering a culture of positive online interaction and providing support for victims could indeed play a crucial role in mitigating online antisocial behavior (ASB). Creating awareness about the impact of ASB, promoting empathy and respect in online interactions, and implementing effective reporting and support systems can help in reducing instances of ASB and creating a safer online environment.

Twitter is a well-liked social networking site that is well-known for its conversational and real-time information sharing capabilities. Users can publish links, text, photos, videos, and other media as brief messages known as tweets. Twitter has been widely used for various purposes, including news dissemination, social networking, and marketing. Its open nature and ease of use have made it a valuable tool for communication and sharing ideas. A user can participate by sending out tweets, which are 280 words long and can include links to articles, images, videos, and other content. The site promotes user engagement through debates on interesting subjects, but this may also lead to certain unwanted behavior, including abuse, harassment, and bullying [1]. Online ASB has been associated with excessive usage of online platforms and is most common among users between the ages of 18 and 27.

The automated systems in place to prevent certain types of content on Twitter, there's still a significant gap in addressing antisocial behavior (ASB). The reliance on user reports for investigating ASB can lead to underreporting due to fear of retaliation or other reasons. Understanding the factors contributing to ASPD and its manifestation online is crucial for developing effective strategies to mitigate ASB on social media platforms like Twitter.

Antisocial Personality Disorder (ASPD), characterized by disregard for others' rights, irritability, aggression, and irresponsibility, is a highly diagnosed personality disorder influenced by genetics, maternal depression, parental rejection, and socioeconomic factors.

2. Literature review

Natural language processing is a challenging undertaking in and of itself since it requires dealing with unclear text. Depending on the context, the same sentence might have several meanings. The task becomes much more difficult when dealing with internet material, which frequently contains misspellings, uncommon abbreviations, slang, and short words. Despite the challenges, researchers have used several machine learning algorithms to analyze emotion and attitudes [23], forecast online harassment and cyberbullying [24], crisis response and emergency scenario awareness [25], and anticipate domestic abuse crises [26]. In machine learning, Logistic regression, Naive Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Random Forest are among the most often used text categorization techniques. The success of afforested classifiers is highly reliant on feature engineering, even though each of these strategies may apply to various situations. Deep learning is a relatively new phenomenon in machine learning methods that have recently slowed the development of neural networks for several years. Deep learning has shown amazing breakthroughs in computer vision, pattern recognition, and image processing. Recurrent Neural Networks and Convolutional Neural Networks are the two most used designs for deep learning. These two approaches use text word embeddings as inputs and generate feature vectors, manipulable numerical representations.

Convolutional Neural Networks have surpassed standard machine learning techniques in question categorization and sentence-level sentiment analysis. The usage of Recurrent Neural Networks to represent text sequences in a corpus has been proven to enhance multiclass classification performance. RNN variants include Long Short-Term Memory networks (LSTMs), Bidirectional LSTM, and Gated Recurrent Units. These RNN variants have improved performance in Natural Language Processing applications because of their integrated memory architecture for storing long-range associations and historical data.

Based on the literature review, it is observed that the RNN architectures are a more suitable technique

to detect cyberbullying. Very few studies were carried out in this context. Therefore, the researchers apply LSTM, GRU and Bidirectional LSTM models from the RNN architecture to train and test the performance over to a selected dataset and develop an ensemble classifier combining all three algorithms, which can help detect and control cyberbullying posts in social media more efficiently.

I. Antisocial Behavior Online

Understanding the diagnostic criteria for ASPD can provide a framework for identifying behaviors associated with online ASB. The criteria, such as disregard for others' rights and impulsivity, can help in recognizing similar patterns in online behavior. When applied to online behavior, these criteria can help in identifying patterns that suggest ASB, such as cyberbullying, harassment, spreading false information, or engaging in online scams. By understanding these criteria, researchers and professionals can better identify and address ASB in online environments.

When a person exhibits behavior that deviates from accepted social norms, they may be diagnosed with ASPD, a diagnosis that is often made in a clinical context. Being antisocial might entail disobeying laws, conventions, customs, and other socially acceptable behavior. Moreover, the phrase "against the rules and law" can also allude to breaking the law, abusing the legal system, being arrested, etc. For their own entertainment and financial gain, an antisocial person may also lie, trick, and control others. They could be impetuous, careless, and prone to fights.

The intent behind language, especially in online communication where tone and intention can be easily misinterpreted. This is a challenge for machine learning algorithms (models), as they struggle to grasp the subtleties of language and context. Humor, sarcasm, and other forms of nuanced communication can be mistaken for antisocial behavior if not analyzed in the right context. Integrating context-aware analysis and considering the broader context of a conversation or message can help in more accurately identifying antisocial behavior online.

To operate, a machine or computer needs to follow a set of guidelines; this is not always the case with NLP. NLP utilizes a range of methods, and the ML technique has been applied in this study project. It takes a lot of data testing and training, as well as ground truth Validation, to train a machine learning algorithm(model) to identify anti social behaviour in an individual's writing. A psychology graduate was consulted in order to classify the data set and verify the ground truth.

II. ASB Consequences

Understanding the impact of online antisocial behavior on mental health, especially in vulnerable populations like children, is crucial for addressing these issues. Exposure to online antisocial behavior (ASB) during childhood can indeed have significant long-term impacts, affecting academic performance and social relationships in adulthood. Victims may struggle with concentration on academic tasks, leading to poor grades and attendance. This can create a cycle of negative effects, impacting various aspects of their lives. Understanding these impacts is crucial for addressing and mitigating the effects of ASB.

Understanding the impact of online antisocial behavior on mental health and the broader public health system is crucial, especially considering the increasing prevalence of such behaviour. The method described in this chapter can assist internet platforms in automatically identifying this type of activity on a broad scale and preventing its spread.

III. Natural Language Processing(NLP)

NLP is the study of how to teach computers to comprehend and produce human language which is inherently complex due to its nuances and context-dependent meanings. Python, as a programming language, can process text as a series of characters but lacks inherent understanding of the meanings of words or sentences. NLP techniques and libraries in Python enable developers to analyze and manipulate text data.

Deep learning has indeed revolutionized natural language processing (NLP) by enabling more accurate and sophisticated models for understanding and generating human language. Two of the most influential architectures in this domain are

Recurrent Neural Networks (RNNs) and Transformers.

3. Method and material /methodology

This research includes two major phases to detect cyberbullying posts and develop an efficient classifier to identify them.

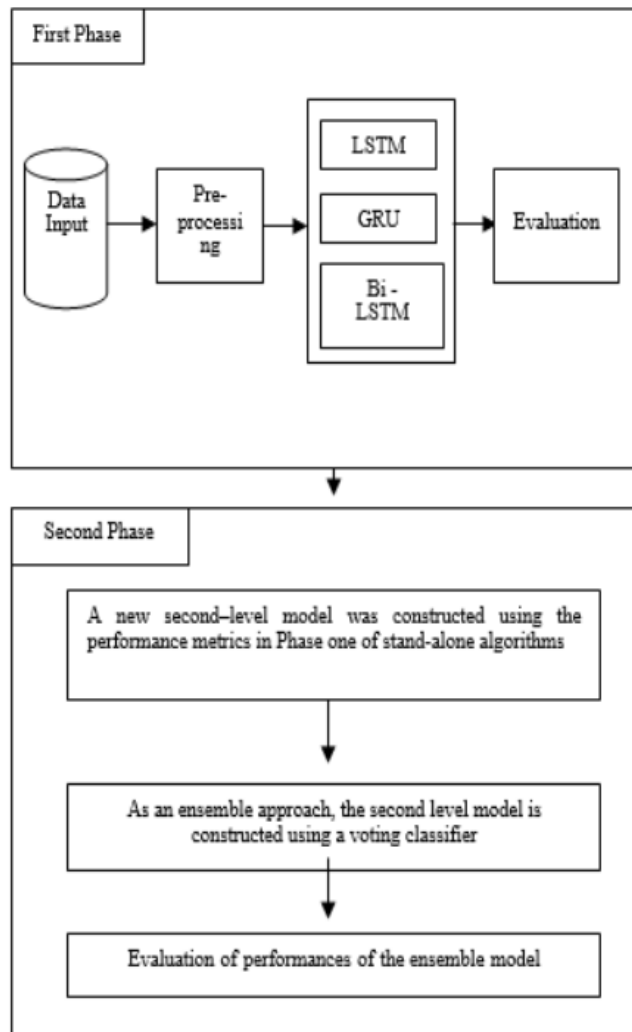


Figure 1 illustrates the proposed methodology.

A. Dataset

Munki Albright uploaded the Twitter dataset on Kaggle. The dataset consists of 19,934 tweets, including suspicious, cyberbullying, hate and suicidal classifications. This research was focused on cyberbullying texts. Sentiment analysis was implemented to classify the cyberbullying posts. For that, sexism (Class 2), racism (Class 1) and either (Class 0) classification were analyzed.

DATASET DESCRIPTION

a. Preprocessing and Feature Extraction

When a text is obtained based on Fig. 1, the data is extensively evaluated using and according to the following procedures: Stopword elimination, tokenization, sentence segmentation, and punctuation removal are some of the accessible features. These measures were used to reduce the quantity of the data, and as a consequence, we were able to delete any unnecessary information. In support of this method, we built a preprocessing program that removes punctuation and some non-letter characters from each text. Finally, the letter case of each document was decreased. This approach produced a sliced document text using an n-gram word-based tokenizer based on the length of n.

b. Tokenization

This approach use tokenization to deal with situations in which a given text will be separated into tokens. The objects listed below are also considered tokens. There are letters, numbers, and punctuation marks present. In addition, a nonsensical sensitive data element was swapped with a non-sensitive comparable element. We assured that the tokenization procedure was protected and tested following the strictest standards for the protection of sensitive data.

c. Stemming

After going through the tokenization method, the next step is to translate the tokens into another standard format. Simply said, stemming means that we may now convert the words back to their original form but with fewer word types and classes in the data. For example, "Winning," "Won," and "Winner" have been abbreviated to "win." It indicates that stemming may be used to build classes.

Classifiers

Long Short Term Memory

RNN, a family of artificial neural networks with node-to-node connections, has severe disadvantages due to numerous network layers. A recent study has found that the LSTM network offers a solution to this problem due to its chain-like topology, which is

similar to that of many neural network modules in RNN.

The recommended course of action, This is provided in this section and is founded on machine learning and natural language processing. It has the ability to automatically identify online abusive speech (ASB) and empower social media sites like Twitter to take proactive steps to stop its spread. The majority of ASB research has been qualitative in nature, with a primary emphasis on thorough examination of a case study. Study teams are typically small and are selected by hand. These investigations might take a lot of time and resources due to their laborious nature. People spend the majority of their time online in today's environment. The way people live has altered as a result of Internet access. Daily activities like work, socializing, banking, shopping, and leisure are becoming primarily done online due to increased screen usage. New approaches to the understanding of personality and behavior traits are required in order to adjust to this changing way of life. This research effort uses human behavior and personality analysis to develop deep learning and machine learning algorithms for online ASB detection using data collected from Twitter.

Once these tweets and labels have been labeled, an authorized individual must confirm them. The one who is capable of diagnosing psychological illnesses and has a deep understanding of them is the qualified person in this situation. Data from tweets in a medical setting were labeled and cleaned using NLP techniques. Next, models for machine learning and deep learning were trained and assessed for optimal results. We constructed the finished model in this chapter by experimenting with the five most popular machine learning techniques. The machine learning methods in question were Naïve Bayes, SVM, RF, Decision Tree, and Logistic Regression Bidirectional LSTM, bidirectional GRU, bidirectional RNN, and CNN were tested in relation to deep learning models. SVM outperformed all other classic machine learning algorithms, while all deep learning models outperformed SVM by a small margin. After the model is constructed, it can be included into social

media sites and other internet platforms. Both stored and real-time stream data will get good results from the model.

When antisocial semantics are identified in a text segment (tweet, post, or news story), in the context of real-time streaming data, it typically means that the content of the text is displaying characteristics or traits associated with antisocial behavior. This could include language that is aggressive, disrespectful, manipulative, or otherwise harmful in nature. Detecting such content in real-time is important for identifying and addressing antisocial behavior as it occurs, allowing for timely interventions or responses.

Extraction of Data

Twitter was used to gather a total of 55,810 tweets between October 2018 and February 2019. Twenty-five,500 tweets remained after duplicates and retweets were removed; they were used in the studies. A tweet is a 280-character message shared by a user on the Twitter network. Three categories exist for psychological disorders: identifying mental and behavioral disorders solely from someone's writing can be challenging. Disorders such as depression, anxiety, and personality disorders can manifest in various ways, including in writing, but they often require a comprehensive assessment that considers multiple factors, including behavior, emotions, and personal history. While certain linguistic patterns or content in writing can sometimes suggest the presence of a disorder, a definitive diagnosis typically requires a professional evaluation by a qualified mental health professional. Both behavior and mental condition might fluctuate over time. But personality disorders or qualities remain constant when they persist in a person for an extended length of time.

A. Pre-processing of Data

Preprocessing is a critical step in preparing textual data for natural language processing (NLP) tasks like text classification. Here are the preprocessing steps taken for the tweets in the dataset used for classifying antisocial behavior (ASB) and non-ASB tweets.

B. Stemming

Stemming is an important step in text preprocessing that helps reduce words to their fundamental form, or root, which can improve efficiency and effectiveness of text analysis. It's particularly useful for tasks like information retrieval and text mining, where reducing words to their base form can help in capturing the underlying meaning and improving the accuracy of analysis.

Reducing the size of the text corpus for machine learning is the goal by using the Porter stemmer. Deep learning algorithms used here don't require extensive pre-processing, but excluding duplicates and organizing the data set is still necessary for smooth utilization of deep learning architecture.

Model Building

NLTK offers a comprehensive set of tools for natural language processing (NLP), and vectorization is a critical step for converting text data into a numerical format suitable for machine learning algorithms. Let's delve into the details of Count Vectorization, Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings.

Count Vectorization

Count Vectorization is a simple and intuitive method for transforming text into numerical data. It involves the following steps:

1. **Tokenization:** Splitting the text into individual words or tokens.
2. **Vocabulary Building:** Creating a list of all unique words in the corpus.
3. **Count Matrix:** Constructing a matrix where each row represents a document and each column represents a unique word from the vocabulary. The value in each cell corresponds to the count of the word in the document.

Term Frequency-Inverse Document Frequency (TF-IDF)

Word Embeddings

Word embeddings are a more advanced technique that captures semantic relationships between words by representing them as dense vectors in a continuous vector space. Popular word embedding

methods include Word2Vec, GloVe, and FastText. The second part of this equation calculates the frequency with which this word appears in all of the tweets. Because of its versatility, Keras[218] is a well-liked option for deep learning algorithms. It's great to hear that you experimented with different architectures like CNNs, bidirectional RNNs, and LSTM/GRU variants, as they are known for their effectiveness in text analysis and classification tasks. Word2Vec is also a powerful tool for feature extraction, particularly with regard to natural language processing.

Assessment of Performance

The model's final building step involved assessing the suggested method for locating and classifying ASB tweets. Performance was assessed using evaluation criteria, including recall, F-measure, accuracy, and precision. Evaluating the performance of machine learning and deep learning classifiers in identifying antisocial behavior (ASB) in posts and tweets involves several key metrics. These metrics provide insights into how well the classifiers are distinguishing between ASB and non-ASB content. Here's a detailed overview of common evaluation metrics:

Among them is ASB. To yet, the problem has not received the attention it merits, and little has been done to identify and stop ASB online. Research has been done on trolling and cyberbullying, which can come under the general category of anti-social behavior; however, little research has been done on identifying the additional elements of this type of behavior.

The research has achieved good results in detecting antisocial behavior (ASB) using NLP and machine learning techniques. High accuracy and consistent precision, recall, and F1 scores across different classifiers indicate the effectiveness of your approach.

$$wf(w, d) = \frac{\text{number of occurrences of a word in tweet}}{\text{total number of all words in a tweet}}$$

Classifier	Feature	Accuracy	Precision	Recall	F1 Score
LR	WF	99.75%	99.59%	99.67%	99.61%
SVM	WF	99.81%	99.70%	99.72%	99.70%
RF	WF	98.08%	99.21%	94.71%	96.89%
DT	WF	99.72%	99.52%	99.55%	99.53%
NB	WF	98.83%	98.89%	97.55%	99.03%

Table 1 Using vectorization of word frequency features

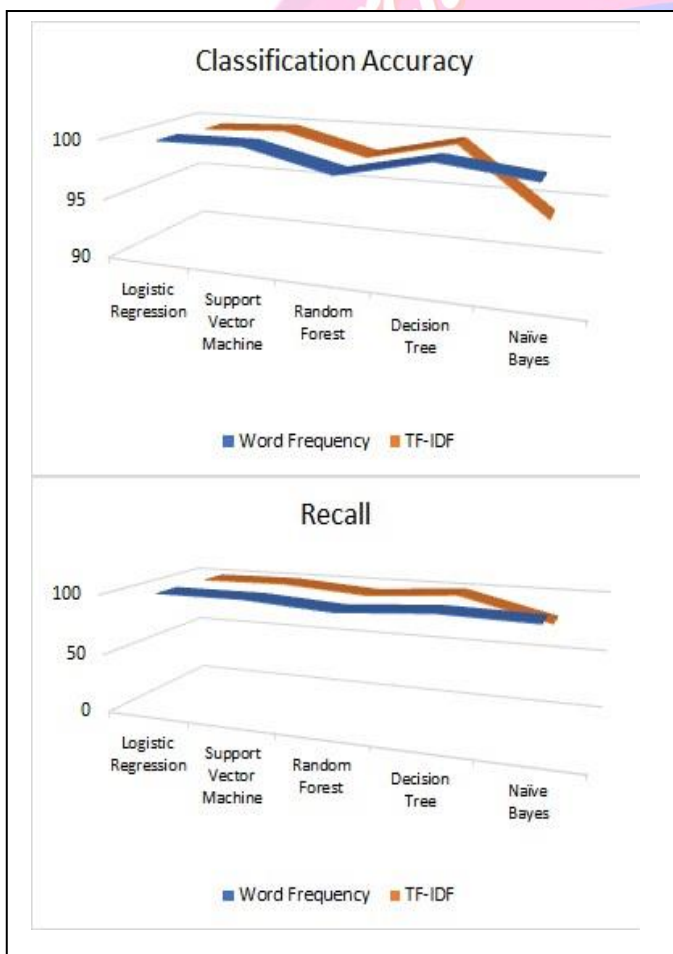


Figure 2a Different Comparison of TF-IDF feature vectors and word frequency

Figure 2 presents the parallels between two distinct vectorization systems' accuracy, recall, precision and F1 score: Considering the information provided, it seems like you want to create a diagram illustrating the performance comparison of different classifiers using Word Frequency and TF-IDF as

features. Here is a basic diagram outlining the information:

This diagram shows that SVM and Logistic Regression performed well across both Word Frequency and TF-IDF, while Naïve Bayes lagged behind in most criteria. SVM was chosen as the classification model due to its overall effectiveness and reliability across different databases.

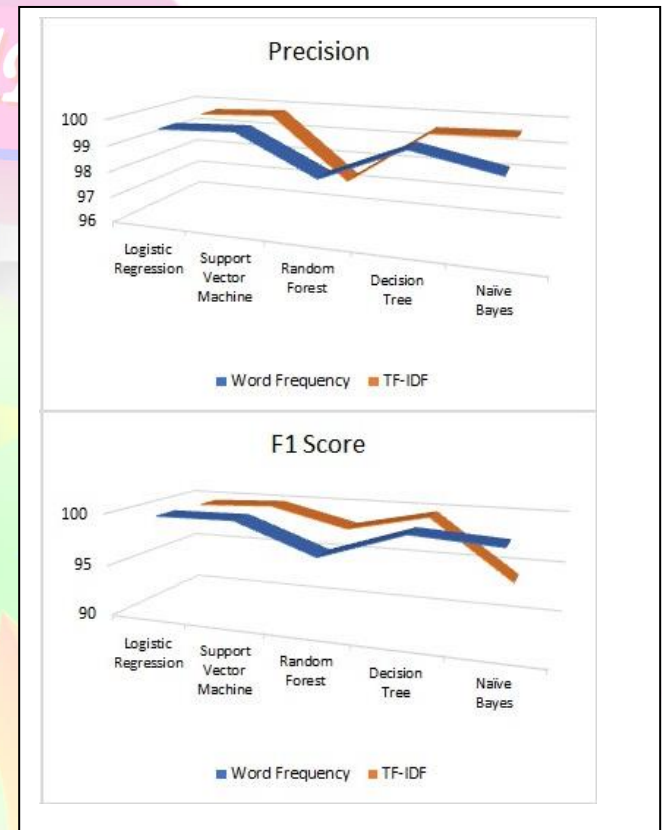


Figure 2b Different Comparison of TF-IDF feature vectors and word frequency

Using Deep Learning to Assess Performance

- Brief explanation the four deep learning models that were employed in the experiment, along with their outcomes:
- **CNN:** CNNs are powerful for learning features from sequential data like text. In your case, using a CNN for tweets makes sense since it can capture local patterns, like phrases or expressions, which are important for understanding the meaning of short text snippets like tweets[22].
- **RNN:** Recurrent Neural Networks (RNNs) are specifically designed to handle sequential data by maintaining a memory of previous inputs, enabling them to capture temporal dependencies in tasks like

language modeling, speech recognition, and machine translation. However, traditional RNNs face the vanishing gradient problem, which hampers their ability to learn long-range dependencies. Here's a detailed explanation:

References

- [1] A. P. Association, Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, 2013.
- [2] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial Behavior in Online Discussion Communities," in *Icwsn*, 2015, pp. 61-70.
- [3] A. M. Gard, H. L. Dotterer, and L. W. Hyde, "Genetic influences on antisocial behavior: recent advances and future directions," *Current opinion in psychology*, 2018.
- [4] E. Flouri and S. Ioakeimidi, "Maternal depressive symptoms in childhood and risky behaviours in early adolescence," *European child & adolescent psychiatry*, vol. 27, no. 3, pp. 301-308, 2018.
- [5] M. Woeckner et al., "Parental rejection and antisocial behavior: the moderating role of testosterone," *Journal of Criminal Psychology*, 2018.
- [6] W. M. McGuigan, J. A. Luchette, and R. Atterholt, "Physical neglect in childhood as a predictor of violent behavior in adolescent males," *Child abuse & neglect*, vol. 79, pp. 395-400, 2018.
- [7] D. B. Jackson, "The link between poor quality nutrition and childhood antisocial behavior: A genetically informative analysis," *Journal of Criminal Justice*, vol. 44, pp. 13-20, 2016.
- [8] A. R. Baskin-Sommers, "Dissecting antisocial behavior: The impact of neural, genetic, and environmental factors," *Clinical Psychological Science*, vol. 4, no. 3, pp. 500-510, 2016.
- [9] J. R. Meloy and A. J. Yakeley, "Antisocial personality disorder," *A. A.*, vol. 301, no. F60, p. 2, 2011.
- [10] P. Liu, J. Guberman, L. Hemphill, and A. Culotta, "Forecasting the presence and intensity of hostility on Instagram using linguistic and social features," *arXiv preprint arXiv:1804.06759*, 2018.
- [11] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, "Trolls just want to have fun," *Personality and Individual Differences*, vol. 67, pp. 97-102, 2014.
- [12] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," *Journal of Information Science*, vol. 36, no. 3, pp. 357-370, 2010.
- [13] J. Guberman and L. Hemphill, "Challenges in modifying existing scales for detecting harassment in individual tweets," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.
- [14] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, "Searching for safety online: Managing "trolling" in a feminist forum," *The information society*, vol. 18, no. 5, pp. 371-384, 2002.
- [15] M. Drouin and D. A. Miller, "Why do people record and post illegal material? Excessive social media use, psychological disorder, or both?," *Computers in Human Behavior*, vol. 48, pp. 608-614, 2015.
- [16] N. Sest and E. March, "Constructing the cyber-troll: Psychopathy, sadism, and empathy," *Personality and Individual Differences*, vol. 119, pp. 69-72, 2017.
- [17] R. Singh, Y. Zhang, and H. Wang, "Exploring Human Mobility Patterns in Melbourne Using Social Media Data," in *Australasian Database Conference*, 2018:Springer, pp. 328-335.
- [18] J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang, "A probabilistic method for emerging topic tracking in microblog stream," *World Wide Web*, vol. 20, no. 2, pp. 325-350, 2017.
- [19] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, and K. Wang, "Convolutional neural network for multi-

class classification of diabetic eye disease," EAI Endorsed Transactions on Scalable Information Systems, vol. 9, no. 4, pp. e5-e5, 2022.

embeddings for hay fever detection from twitter," Health information science and systems, vol. 7, no. 1, p. 21, 2019

[20] R. Singh, Y. Zhang, H. Wang, Y. Miao, and K. Ahmed, "Investigation of social behaviour patterns using location-based data—a melbourne case study," EAI Endorsed Transactions on Scalable Information Systems, vol. 8, no. 31, 2020.

[21] R. Singh et al., "Deep Learning for Multi-class Antisocial Behaviour Identification from Twitter," IEEE Access, 2020.

[22] J. He, J. Rong, L. Sun, H. Wang, Y. Zhang, and J. Ma, "A framework for cardiac arrhythmia detection from IoT-based ECGs," World Wide Web, vol. 23, pp. 2835- 2850, 2020.

[23] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, "Automated epilepsy detection techniques from electroencephalogram signals: a review study. Health Information Science and Systems. 2020; 8 (1): 1–15," ed.

[24] J. Lee, J. S. Park, K. N. Wang, B. Feng, M. Tennant, and E. Kruger, "The use of telehealth during the coronavirus (COVID-19) pandemic in oral and maxillofacial surgery—a qualitative analysis," EAI Endorsed Transactions on Scalable Information Systems, vol. 9, no. 4, 2021.

[25] S. B. Manuck and J. M. McCaffery, "Gene-environment interaction," Annual review of psychology, vol. 65, pp. 41-70, 2014.

[26] L. W. Hyde et al., "Heritable and nonheritable pathways to early callous-unemotional behaviors," American Journal of Psychiatry, vol. 173, no. 9, pp. 903-910, 2016.

[27] K. Samal, K. Babu, and S. Das, "Predicting the least air polluted path using the neural network approach," EAI Endorsed Transactions on Scalable Information Systems, vol. 8, no. 33, 2021.

[28] J. Du, S. Michalska, S. Subramani, H. Wang, and Y. Zhang, "Neural attention with character