EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Antisocial Behavior Identification from Twitter Feeds Using Traditional Machine Learning Algorithms and Deep Learning.

Ravinder Singh[1,*], Sudha Subramani[1] Jiahua Du[1] ,Yanchun Zhang[1], Hua Wang[1], Yuan Miao[1], Khandakar Ahmed[1]

[1]Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne, Australia.

Abstract

Antisocial behavior (ASB) is one of the ten personality disorders included in 'The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and falls in the same cluster as Borderline Personality Disorder, Histrionic Personality Disorder, and Narcissistic Personality Disorder. It is a prevalent pattern of disregard for and violation of the rights of others. Online antisocial behavior is a social problem and a public health threat. An act of ASB might be fun for a perpetrator; however, it can drive a victim into depression, self-confinement, low self-esteem, anxiety, anger, and suicidal ideation. Online platforms such as Twitter and Reddit can sometimes become breeding grounds for such behavior by allowing people suffering from ASB disorder to manifest their behavior online freely. In this paper, we propose a proactive approach based on natural language processing and deep learning that can enable online platforms to actively look for the signs of antisocial behavior and intervene before it gets out of control. By actively searching for such behavior, social media sites can prevent dire situations leading to someone committing suicide.

Keywords: Antisocial Behavior Disorder, Behavior Classification, Personality Disorder, Online Antisocial Behavior, Deep Learning, Machine Learning.

Received on 27 March 2023, accepted on 05 May 2023, published on 12 May 2023

doi: 10.4108/eetsis.v10i3.3184

1. Introduction

Antisocial behavior is one of the ten personality disorders in 'The Diagnostic and Statistical Manual of Mental Disorders (DSM-5). These ten disorders are characterized into three clusters: antisocial personality disorder falls in Cluster B, Borderline Personality Disorder, Histrionic Personality Disorder, and Narcissistic Personality Disorder [1]. It is a prevalent pattern of disregard for, and violation of the rights of others. A person with antisocial personality disorder fails to conform to social norms concerning lawful behavior. The person can become irritable, aggressive, and consistently irresponsible when dealing with others. The person may lack remorse and mistreat others [1, 2]. Many elements may lead to a person developing antisocial behavior: genetic influences, maternal depression, parental rejection, physical neglect, poor nutrition intake, and adverse socioeconomic or sociocultural factors are few of them [1, 3-7]. These factors can be categorized broadly into three main categories: Neural, Genetic, and Environmental [8]. Antisocial personality disorder (ASPD) is one of the most reliably diagnosed conditions among all personality disorders. Many psychiatrists are reluctant to treat people who suffer from ASPD because there is a widespread belief that it is untreatable; however, there is increasing evidence that it can be treated in certain cases[9].

Online antisocial behavior is a widespread problem and threatens free discussions and user participation in many online communities. It can devastate victims and deter

*Corresponding author: Email: Ravinder.singh@vu.edu.au

them from using these platforms [10]. Online antisocial behavior appears to be an Internet manifestation of everyday sadism. An individual who possesses and displays such behavior online seems to enjoy it at the expense of others ignoring the distress and harassment it may cause [11]. Apart from Sadism, attention-seeking, boredom, a desire to cause damage to the community, and revenge are some of the motivations related to antisocial behavior [12]. Antisocial behavior annoys and interferes with a person's ability to conduct business lawfully. Current measures to discourage antisocial behavior rely mainly on users reporting it directly to platforms [13].

In most cases, victims are reluctant to confront such behavior online because they fear retaliation. Therefore, most cases of antisocial behavior go unnoticed. Online platforms encourage freedom of speech but fail to distinguish between free speech and unacceptable behavior. Current measures do not effectively prevent people from explicitly displaying antisocial behavior, exposing many who may fall into a vulnerable group of people, to be on the receiving end of such behavior.

Twitter is among the most popular social media platforms encouraging people to share views and content. A user contributes in the form of a tweet, which is 280-word text and may contain an image, a video, a link to an article, etc. The platform encourages user participation in discussions on topics of interest, however, this may bring along some undesirable behavior, such as bullying, abuse, and harassment [2]. Online antisocial behavior is prevalent mainly among users aged 18-27 and has also been linked to excessive use of online platforms. Perpetrator seems to enjoy at the cost of others [11, 12, 14]. Ramifications of excessive use also lead to other psychological disorders, and employing measures to curtail their impact on society is imperative [15].

1.1. Motivation

Twitter and other platforms rely on users to report Antisocial behavior to the media. Only based on the user initiative, Twitter intervenes and looks into matters of various things. The platform has automated systems that may prevent the distribution of illegal material, spam, nudity, pornography, etc. but nothing for antisocial behavior [16]. So, on the one hand, the platform connects users to enable the exchange of information, ideas, and other valuable resources. On the other hand, however, it facilitates the spread of antisocial behavior and related issues and puts many people at risk[17]. This paper studies the problem of antisocial behavior on public platforms.

1.2. Our Approach

We in this paper propose an approach based on natural language processing (NLP) and deep learning that can be used, on a scale, to classify tweets containing antisocial behavior with high accuracy and precision. The approach can be implemented to automate the detection of antisocial behavior on Twitter and other online platforms to curb its prevalence. To our knowledge, no prior work has either focused on critical tweet identification or evaluated Deep Learning, and machine learning techniques against different feature extraction approaches for identifying psychological disorders, particularly antisocial behavior, from social media data. Firstly, a benchmark data set of tweets labeled “ASB” and “non-ASB” is constructed. Textual features are then extracted from unstructured data for deep learning, a cutting-edge machine learning architecture, along with traditional machine learning algorithms, to build a prediction model for the automatic identification of ASB. We have treated the problem of recognizing ASB online as a binary text classification task, where a tweet is classified as “ASB” or “non-ASB” based on the textual content. Table-1 shows examples of antisocial and non-antisocial tweets classified by our approach.

Table 1. Examples of Antisocial tweets with corresponding labels.

ID	Online Tweet	Context	Label
T_1	Fuck the law. Laws are merely words on paper backed by the promise of violence if they are not obeyed. Words on paper have no authority over any of us. Just like imaginary lines have no authority.	Failure to conform to social norms.	Antisocial
T_2	i will Personally fight the sun. come at me you fucking bitch i will Destroy You	Irritability and aggressiveness	Antisocial
T_3	Drink driving is fun and I do not mind doing it every now and then as long as I am not caught. I mean to have	Reckless disregard for safety of self or others	Antisocial

	fun is our right.		
T_4	I snubbed the dude and then he fell. he was a looser and I could not care less. No regret	Lack of remorse as indicated by being indifferent to or rationalizing having hurt and mistreated	Antisocial
T_5	I would like to win the mega lotto this Tuesday, and if I do win a lot of people will be very happy because I am a very generous person and I like to help others good luck to me;!!!	Seeking wishes and good luck	Non-Antisocial
T_6	As a cancer survivor I like to help out where I can.	Being nice and offering help.	Non-Antisocial
T_7	So sorry for your loss. I pray for you and your family	Condolences	Non-Antisocial

We evaluated the performance of our proposed approach and compared the results between the Deep Learning methods and other traditional machine learning techniques. Analyses of features helped to identify important words that could distinguish between ASB and non-ASB tweets. The experiment results and analysis benefit researchers interested in researching online antisocial behavior, utilizing social media data.

1.3. Contribution

The main contributions of this research study are: (1) A medium-scaled benchmark antisocial behavior tweet dataset was constructed with labels for antisocial and non-antisocial behavior tweets. (2) We developed a deep learning classification model after evaluating the performance of different DL architectures. (3) We validated the higher performance of Deep Learning models against several traditional machine learning models. (4) We also presented a visually enhanced interpretation of different feature vectors in machine learning (5) Proposed a novel approach to study psychological disorders from social media data using artificial intelligence

The rest of the paper is organized as follows. Section II provides the background on antisocial behavior disorder, etiology, manifestation, and repercussions. Section III presents an approach to ASB tweet identification. Section IV offers details on experiments conducted to evaluate our system with analyses of the results and discussion. Section V concludes our work in this paper and envisages future research directions.

2. Background

2.1. Online Antisocial Behavior

To understand antisocial behavior online, we went through the diagnostic criteria explained in DSM-5 [1]. The term antisocial personality disorder is primarily used in a clinical setting and may be used to describe the behavior of a person against societal norms. To be antisocial may mean to be against rules, laws, standards, and acceptable behavior [2]. Furthermore, against the government and regulation may refer to a failure to obey laws and legal system, engaging in criminal activities, arrest, etc. A person with an antisocial personality may lie, deceive, and manipulate others for self-amusement and profit. The person may get irritable and aggressive quickly and is inclined to engage in fights. The person may also be impulsive, irresponsible, and lack remorse for action [8]. A person's writing can diagnose not all psychological disorders, but a few can and antisocial behavior is one of them. Since it can be interpreted by how a person writes, we can detect such behavior online from tweets, posts, reviews, and comments. In any text, antisocial behavior is expressed by using words and the context of the use of those words. There are several rude and taboo words and short phrases that can be associated with antisocial behavior. It may seem easy for a human to pick up such behavior through text; however, it may not be that easy for a machine [18]. One reason is that some rude words can be used in humor or sarcasm, which may not always be considered antisocial. Also, the context of a text plays an important role in classifying it as an antisocial text.
The use of slang, the order of words, local culture, etc., all play an important role in classifying a text. Some words and phrases that are normal to use in one country may imply rude or antisocial behavior in other. An example is an experience shared by a friend from Australia, who was in a café in the US and asked for a 'White Coffee". This is a standard way of getting a coffee with milk in Australia; however, in the US the guy at the café, who was a person of color, thought that my friend was rude and racist. My friend should have asked for 'coffee with milk' instead of 'white coffee.' Under certain circumstances, it is difficult for even humans to know the exact intentions of a person

from their writings, and we can imagine how hard this could be for a machine.

A machine or a computer relies on a set of rules and instructions to take any action; however, in the case of natural language processing, it could be more straightforward. There are few techniques used in natural language processing, and we, in this research project, have used a machine learning approach. To train a machine learning model to detect antisocial behavior from a person's writing requires a lot of training and testing data, along with ground truth validation. We sought the help of a psychology graduate to label our dataset and ground truth validations.

2.2. Etiology of Antisocial Behavior

Understanding the etiology of antisocial behavior may be the first step toward preventing and eliminating it. Antisocial behavior disorder is part of Cluster B of personality disorders and borderline, histrionic, and narcissistic personality disorders. Individuals who have these behavioral personality disorders appear emotional, dramatic and erratic. These characteristics are common in all these four disorders in the cluster. A person with antisocial behavior often disregards other people's emotions and feelings and often engages in activities that are considered illegal; however, the manifestation of such activities dwindles as the person grows older [1]. There may be many elements that lead to a person developing antisocial behavior. Some of these are genetic influences, maternal depression, Parental-rejection, physical neglect, poor nutrition intake, and adverse-socioeconomic and sociocultural factors. [1, 3-7, 19] These factors can be categorized broadly into three main categories: Neural, Genetic, and Environmental [8, 20, 21]. Antisocial behavior due to neural factors has been studied through structural and functional approaches. Structural studies assess the brain's morphology, and available studies assess its activity. Together these studies try to understand core neural regions that are related to salience detection, affect, and controlled cognition, including the frontal cortex, amygdala, and anterior cingulate cortex [8]

A person's genes are linked to the antisocial behavior he may develop during adolescence[3, 22-24]. Certain types of gene combinations are closely associated with such behavior. A child raised by biological parents diagnosed with antisocial behavior is highly likely to develop antisocial behavior. However, some studies have concluded that if the same child with parents diagnosed with antisocial behavior is raised by adopted parents, who do not suffer from the such disorder, he has a lower chance of developing antisocial behavior [20, 25, 26]. Therefore, genes play an essential role in the onset of antisocial behavior in a person. However, the impact can be mitigated if the individual's environment can be more positive.

Some environmental factors that may trigger or lead to antisocial behavior are exposure to community violence, family dysfunction, and peer influence [8, 27, 28]. Research has shown that being part of a disadvantaged community, living in a poor neighborhood, dependent on social security, being a part of female-headed households, and not having a job may exaggerate or trigger the onset of antisocial behavior [29-32]. Being a part of a broken family and facing maltreatment by either parent, a parent's mental health can also impact an individual's mental health. Apart from parents, abuse by others around an individual can also prompt him to manifest antisocial behavior [33, 34]. Child neglect, in general, has been associated with ASB [35]. The sort of company an individual hangs around with usually influences his behavior and personality, and vice-versa. So hanging around with an individual who manifests antisocial behavior can lead you to display such behavior as well [36-41].

We all know that smoking harms the person who smokes and individuals around him who inadvertently and passively inhale smoke excreted by a smoker. Many studies have linked maternal smoking during pregnancy and severe mental disorders, particularly antisocial behavior disorders in offspring [4, 42-49]. Similarly to smoking, excessive parental drinking is also associated with an offspring developing ASB.

Apart from neural, genetic, and environmental factors, some studies have found a link between poor-quality nutrition and childhood antisocial behavior [7]. A deficiency of B-Vitamin is mainly linked with ASB and other mental health & behavioral disorders [50]. So, many factors can lead to an individual developing general mental health disorders and antisocial behavior. Here, we have discussed some of the crucial factors leading to ASB, and since this is still an active area of research, we may learn more about this personality disorder in the future.

2.3 Manifestation of ASB

Antisocial behavior emerges in disparate forms online. Some of the most common ones are trolling, cyberbullying, threatening, hostile behavior, offensive language, the publication of inappropriate images, etc. Trolling is widespread on social media, magazines, and news websites. Trolls are general visitors to a website and write offensive and inflammatory comments in the public section. Their main aim is to disrupt an online discussion and, at the same time, grab some attention in the process. They disregard the author of the writing they comment on and show no respect to other commenters. They do this by posting comments that are sexist, hateful, racist, and profane in nature. Troll intensity range from subtly provoking someone to outright threatening and abusing [51]. For some, trolling traits are inborn, and they have a history of trolling and engaging in such behavior online. These people seem to enjoy at the cost of others [11, 12, 14]. This type of trolling is associated with sadism [52]. For others, environmental variables, situations, and context can come into play [53]. Negative mood and seeing other people trolling online can also thrust someone into trolling

[54]. A person who otherwise has a charming and normal personality can sometimes be pushed into getting involved in trolling inadvertently. This sort of situation may arise if someone, who is not a troll, feels that he has been pushed around and needs to stand up to it. In the process, the victim himself can start trolling the abuser, either to get him off his back or to teach him a lesson, in the hope of preventing him from engaging in trolling in the future [20, 54]. Studies have found that people who troll focus their effort on a small number of threads and make issues of petty things. They usually write worse than people who do not troll, and, in some instances, their writings are irrelevant to the topic of discussion. Considering gender from a perspective, males are more likely to get involved in trolling compared to females [16, 17, 55]. Over time, these trolls become less tolerant of the online community and get reported and kicked out of the conversation and, in some cases, from the community [2, 56]. The impact of trolling on victims can sometimes be more devastating than if they have experienced similar behavior in real life [57]. Exposure to online trolling can lead the victims to experience psychopathological outcomes such as anxiety, depression, and low self-esteem [58].

While trolls mainly focus on being a nuisance and attracting attention, cyberbullies target individuals. Instead of posting general offensive and inflammatory statements in the public comment section of a website, they post abusive and vicious comments about a single individual. Cyberbullying refers to using an online platform, such as Twitter, Reddit, Facebook, etc., to intentionally and repeatedly harass or harm an individual [59]. Cyberbullies focus on intimidating, shaming, and demeaning their victims. Unlike trolls, cyberbullying does not usually want to attract attention and instead focuses more on targeting an individual and causing distress to them. For this, they post images, text, audio, and video targeted at individuals repetitively. [59] This media is abusive, aggressive, and intentionally drafted to bully someone online. As the use of the online platform has increased, so has cyberbullying, which is prevalent in school-aged children. Depending on the measuring tool applied, 10%-40% of school-age children experience some sort of cyberbullying [60]. Cyberbullying has recently been getting much attention from government authorities and social scientists because of its association with many suicides [61].

Trolling and cyberbullying are the two most prominent manifestations of antisocial behavior online. Threats, misleading and wrong information, offensive language, sexism, racism, and rude and taboo words are other ways antisocial behavior can take form online.

2.4 Repercussions of ASB

Antisocial behavior and its impacts are matured and well-researched; however, online antisocial behavior is a relatively new research area and has recently gained plenty of attention. The perpetrator often displays such behavior via cyberbullying and trolling. Individuals on the receiving end of antisocial behavior get impacted in many different ways, and adverse health impact is one of them. Victims can suffer internalizing problems such as depression, low self-esteem, anxiety, suicidal ideation, and anger [62]. They can also go through externalizing such problems as alcohol abuse, smoking, self-harm, aggression, and hostile behavior towards the external environment [63, 64]. Victims of all ages can experience negative mental health impacts of antisocial behavior; however, individuals who come across such behavior as children have higher chances of developing the psychological disorder and social problems [65-67]. Some studies have linked exposure to antisocial behavior as a child, a drop in academic performance as a young child, and low-income family and social relationships as an adult [68, 69]. Victims are often preoccupied with their antisocial behavior experiences and find it hard to concentrate on academic tasks, leading to poor performance. In addition, to drop-in grades at school, a drop-in attendance has also been linked to exposure to such behavior, leading to a vicious cycle affecting all aspects of academic life.

Adult victims of antisocial behavior report a higher level of anxiety, depression, and severe social difficulties. The thought of experiencing such behavior repeatedly prevents many victims from going out and socializing. This leads to self-confinement & isolation, leading to depression and social problems [70]. Antisocial behavior at work leads to lower employee morale and lost output. Perpetrators often target their victims via inflammatory emails, offensive text messages, and posting inappropriate comments and images. Females, minorities, and new employees are often easy targets for such behavior. Putting measures in place and managing such activities & behavior at the workplace cost organizations many resources, in addition to negative media coverage and higher staff turnover, which again adds to the cost of doing business [71-75].

Many studies have linked victims of antisocial behavior to drug and alcohol use, hyperactivity, and a decline in pro-social behavior [76, 77]. Victims fall prey to drugs and alcohol as a convenient escape to their problems, and excessive use makes them hyperactive and deters them from socializing. Studies have also linked suicidal thoughts and self-harm behavior as one of the ramifications of experiencing antisocial behavior. Self-harm may include cutting, jumping from heights, self-battery, burning, and poisoning, with some industrialized and developed nations experiencing higher-than-average incidents [78-80]. In contrast, some victims may use aggression to get their frustration out and may bully, harass or troll other individuals around them [81].

Despite its relatively brief history, online antisocial behavior has been identified as a severe public health threat. Apart from the direct impact on victims and an indirect impact on their family & friends, antisocial behavior is also a burden on the public health system. The cost of treating individuals with depression, anxiety, and other related psychological disorders adds up and

significantly impacts the growing public health spending [71, 82].

2.5. Obligation to Restraint

Online antisocial behavior needs to be deterred and confined. We may not be able to eliminate it; however, by placing appropriate measures in place, we may be able to confine it to a certain extent or reduce its impact on victims, their families, the workplace, and the public healthcare system. Antisocial behavior is a massive cost to our society. With the advent of the internet, it has become ever easier for many people to indulge in such behavior online. The spike in the incidents of antisocial behavior in general and online antisocial behavior, in particular, can be explained by the fact that perpetrators can stay anonymous online, which usually is not an option for them in a face-to-face situation [83]. In the real world, perpetrators typically have power over victims they exploit to bully and harass. This power could come in many forms, such as social status, physical strength, and workplace seniority. In an online world, these powers may be insignificant. Since a perpetrator can stay anonymous, he can also bully and harass someone higher in a hierarchy regarding workplace seniority, status, and physical strength [84]. Also, in a face-to-face confrontation, a perpetrator can seize abuse and bullying once he recognizes that he has hurt his victim enough. In online bullying, a perpetrator may not know when to stop and can push the victims toward extreme actions such as self-harm or suicide. Pushing someone online to the extent that may lead the victim to commit suicide is a form of suicide baiting, using which an offender encourages a victim to take their life [85]. Online antisocial behavior is a huge cost and burden to our society. Damages caused by this can be seen in families, workplaces, and the public healthcare system. This kind of behavior is never accepted, whether online or offline. Even though large social media platforms and other online platforms are responsible for ensuring their platforms do not become breeding grounds for antisocial semantics, it is everyone's responsibility, who use those platforms to discourage and report such behavior. Our proposed approach in this paper can assist platforms to automatically identify such behavior on a scale and restrain it from spreading.

2.6 Natural Language Processing

Natural language processing is a field concerned with the ability of a computer to understand, analyze, manipulate, and potentially generate human language. By human language, we refer to any language used for everyday communication. This can be in English, Spanish, French, or Mandarin. A programming language such as Python, which we have used in this research, does not naturally know what any given the word means. All it sees is a string of characters. For example, it has no idea what antisocial means. It seems it is a ten-character long word, but the individual character doesn't mean anything to Python. Indeed, the collection of those characters together does not mean anything either. Humans know what an 'A' and a 'S' means, and together, those 10 characters make up the word 'antisocial', and we know what that means. So natural language processing is the field of getting the computer to understand what 'antisocial' signifies, and from there we can get into the manipulation or potentially even generation of that human language [74, 86]. We probably experience natural language processing daily without even knowing. Natural language processing is a broad and evolving field that encompasses many topics and techniques. The core component of natural language processing is extracting all the information from a text block relevant to a computer understanding of the language. There are many techniques for natural language processing and machine learning methods in general and deep learning, in particular, is the most promising of all. Machine learning is a field of study that allows computers to learn without being explicitly programmed.

3. Methodology

This section presents our approach based on natural language processing and machine learning that can automatically detect antisocial behavior online and can enable platforms such as Twitter to proactively prevent it from spreading by having appropriate measures in place. Most of the research on antisocial behavior has been qualitative, focusing mainly on deep case study analysis. Study groups are often chosen manually and are small in number. These studies could be more convenient and may require a lot of resources and time. In today's world, we humans spend most of our time online. Access to the Internet has changed how we live and do things in our daily lives. We spend more time in front of screens today than we ever did. Most of our daily tasks, such as work, social interactions, banking, shopping, entertainment, etc. occur online. Since the way we live and do things have changed significantly, we need new ways to explore and study personality and behavioral traits [17, 34]. The research for this project has been conducted by collecting data from the social media site Twitter. Since this data is generated during our interactions with the outer world, it has much information about our human behavior and personalities. In this research project, we extract such information and use it to build machine learning and deep learning models to detect antisocial behavior online.

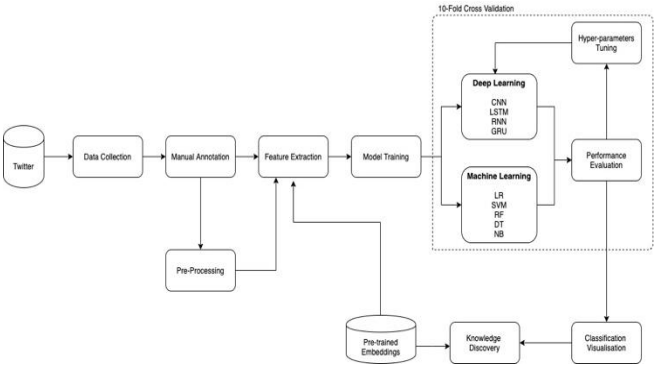

Figure 1. The architecture of our Proposed Antisocial Behavior Detection Approach.

The proposed approach consists of collecting the right tweets and labeling them. Once marked, these tweets and labels need to be verified by a qualified person. The qualified person in our scenario is a person who has a thorough understanding of psychological disorders and can diagnose them in a clinical setting. Once the data, a set of tweets in our case, is properly labeled, we can use natural language processing techniques to clean and pre-process it. These natural language processing techniques are discussed in detail in the following sections. After we have cleaned data, it is used to train and test traditional machine learning & deep learning models to establish which one outputs the best results. For our model building, we experimented with the five most popular traditional machine learning algorithms: Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Naïve Bayes.Regarding Deep Learning Models, we experimented with CNN, Bidirectional RNN, Bidirectional LSTM, and Bidirectional GRU. Among the traditional machine learning algorithms, Support Vector Machine performed the best, however, all the deep learning models performed better than SVM. Once the model is built, it can be integrated into any online platform, including social media. We believe the model will perform well with stored and live-stream data. In the case of live stream data, once a text (tweet, post, news) is triggered to have antisocial semantics, it can either be removed by an algorithm or may require human intervention for further actions. A proactive approach like this can help reduce online antisocial behavior and encourage healthy and clean discussions.

3.1 Data Extraction

We collected 55,810 tweets from Twitter between Oct 2018 and Feb 2019. After removing retweets and duplicates, we were left with 25,500 tweets. Tweets are 280-word text that users share with others on the Twitter platform. When Twitter first started, the limit on the length of these tweets was 140 words and was later increased to 280 words as the platform's popularity soared. We used various phrases such as “I do not care about the law,” “I wish you die soon”, “Go to hell” etc. to search and collect these tweets. Text data collected online is typically in a semi-structured or unstructured form. Our data was no different and was in semi-structured from when first collected. Therefore, some tweets were missing delimiters and did not indicate any punctuation. We used functions from the NLTK library of Python to structure our data set. Once the dataset was in a structured form, we annotated the dataset manually with two categories: Tweets that conveyed antisocial behavior and tweets that did not. Once the dataset was annotated, we wanted to get it verified by someone from Psychology. We hired a psychology graduate to do so. The person thoroughly understood all the personality disorders and could diagnose them in a clinical setting.

Psychological disorders can be classified into personality, behavior, and State of mind. Behavior and State of mind disorders fluctuate and usually cannot be detected in a person's writing. Behavior and State of mind may change from time to time. However, Personality disorders or traits do not fluctuate and stays with a person for a longer period [1, 87]. Since these traits stay with a person longer, they manifest through a person's speech and online writings. Antisocial behavior is a personality disorder that can be reliably detected from online corpora. Our annotator could manually review tweets to see if they qualified as ASB tweets. If a tweet did, it was labeled one. Once labeled, our dataset was ready to be explored further.

3.2 Data Pre-processing

This phase involved removing punctuation from our tweets, followed by tokenization, which means dividing a sentence into individual words. Once tokenization was done, the next step was to remove stop words. Stop words are words that do not contribute much to the meaning of a sentence. Examples of such words are the, is, are, etc. After removing the stop words, we used stemming to cut down words into their shortest form. This is done to reduce the work for our algorithm. All these steps are explained in the following paragraphs.

The first step in the pre-processing phase was to remove punctuation from the tweets. To remove punctuation, we had to show Python, the programming language we used, what punctuation even looked like. We accomplished by using the String package in Python. We care about removing punctuation because period, parentheses, and other punctuations look like just another character to Python. Still, realistically, the period does not help pull the meaning out of a sentence. For instance, for us "I like to research." with a period, is the same as, "I like to research". They mean the same thing for us, but when we give these sentences to a machine learning algorithm, the algorithm says those are not equivalent. We wrote a function to cycle through every character, checked if it was punctuation, and discarded it if it was. This was done to reduce the workload

of our algorithm. By removing punctuations, our algorithm had to deal with fewer characters in the learning process.

Now that we had removed punctuation, we could begin tokenizing our text. Tokenization is spitting some string or sentence into a list of words by white spaces and special characters. For example, we could split the sentence "I am doing research" into four words: 'I', 'am', 'doing', and 'Research'. Instead of seeing the whole sentences, our algorithm could see four distinct tokens, and it knew what to look at. Some of the words were more important than others. For instance, the words 'the', 'and', 'of', and 'or', appear frequently but offered little information about the sentence itself. These are what we call stop words. We removed these words to allow our algorithms to focus on the most key words in our tweets. From the example above, if we remove 'I' and 'am', we are left with 'doing research'. This still gets the most important point of the sentence, but now our algorithm is looking at half the number of tokens. The next step in the process was stemming. Stemming is reducing inflected or derived words to their word stem or root. In other words, it means to chop off the end of a term, to leave only the base. This means taking words with various suffixes and condensing them under the same root word. For example, we can stem words such as 'connection', 'connected,' and 'connective' to one word 'connect.' Stemming shoots for the same goal by reducing variations of the same root word and making our algorithm deal with fewer words. Without stemming, our algorithm must keep all three words: 'connection,' 'connected', and 'connective' in memory, increasing the workload and making our machine learning model less efficient. In simple words, the whole idea of all these steps is to reduce the corpus size for our machine-learning model to deal with. For stemming, we used the Porter stemmer from the NLTK package.

The deep Learning algorithms we implemented in this paper do not require similar pre-processing as traditional machine learning algorithms. We still had to remove duplicates and retweets and had to structure the data set to iron out any abnormalities and website links.

3.3 Model Construction

The natural language toolkit is the most utilized package for handling natural language processing tasks in Python. Usually called NLTK for short, it is a suite of open-source tools initially created in 2001 at the University of Pennsylvania to make natural language processing in Python easier. NLTK is great because it provides a jumpstart to building any natural language processing tasks by providing essential tools that can be chained together rather than making them from scratch. We used the NLTK package of Python for the traditional machine learning algorithms.

Once we had clean text data that we could use to build our machine-learning model, we needed to convert it into a form that our model could easily understand. The process is called vectorization. This is the process of encoding text as numbers to create feature vectors. A feature vector is an n-dimensional vector of numerical features representing some object. In our context, we had to convert individual tweets into a numeric vector representing those tweets. We did this by taking our dataset, which had one line per document, with the cell entry as the actual text message, and converting it into a matrix that still had one line per document. Still, then we had every word used across all documents as the columns of our matrix. And then, within each cell was counting, representing how many times that particular word appeared in that document. This is called a document-term matrix. Once we had the numeric representation of each tweet, we carried it down with our machine-learning pipeline and fitted and trained our machine-learning model. We vectorized text to create a matrix that only had numeric entries that the computer could understand—in our case, counting how many times each word appeared in each tweet. A machine-learning model understands these counts. If it sees a one, a two, or a three in a cell, the model can start correlating that with whatever we're trying to predict. In our case, that was antisocial behavior. We analyzed how frequently certain words appeared in a tweet in context to other words to determine whether the tweet manifested antisocial behavior. We used both Word Frequency (WF) and Term Frequency-Inverse Document Frequency (TF-IDF) vectorization methods in this research. We did this to see a difference in the results with our machine learning model. The count vectorization created the document-term matrix and then counted the number of times each word appeared in that given document, or tweet in our case, and that is what is stored in the given cell. The equation for this is:

$$wf(w,d) = \frac{number\ of\ occurrences\ of\ a\ word\ in\ tweet}{total\ number\ of\ all\ words\ in\ a\ tweet} \quad (1)$$

Term Frequency-Inverse Document Frequency, which is often referred to as TF-IDF, created a document-term matrix, where there was still one row per tweet, and the column still represented a single unique term; however, instead of the cells representing the count, the cell represents a weighting that was meant to identify how important a word was to an individual tweet. We started with the TF term, which is the number of times a term occurred in a tweet divided by the number of all terms in the tweet. For example, if we use "I like research" and the word we focus on is 'research' then this term would be 1 divided by 3 or 0.33. The second part of this equation measures how frequently this word occurs across all the tweets. We started by calculating the number of tweets in the dataset and divided that by the number of text messages that this word appeared in and then took the log of that equation. For example, if we had 20 tweets and only one had the word 'research', then the inverse document frequency means log (20/1). We had two parts of the equation: Term and inverse data frequency. The last step was to multiply both to get a weight for the word 'research' in the tweet. The equation is as follows:

$$W_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (2)$$

Both the matrices have the same shape; the only difference is the values in the cells. After vectorization, we had our data set that algorithms could use to build a machine learning model. Machine learning is the field of study that allows computers to learn without explicitly programming. We do that by training a model using data and an algorithm and then testing its accuracy using more data. To this end, we divided our dataset into buckets to train and validate our model. In this project, we used the K-fold Cross Validation method to divide our data, using tenfold Cross-Validation. The full data set was divided into ten subsets and the holdout's procedure was repeated 10 times. Each time, nine subsets were used to train the model and the tenth subset for testing it. Results were stored in an array, and the method was repeated 10 times with different testing sets each time. In the end, the average of all test results was taken to produce the final result. While building our model, we tried five traditional machine learning algorithms: Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Naïve Bayes. These algorithms were implemented twice using two different vectorization methods: Word Frequency and TF-IDF. As discussed earlier, not much research has been conducted to deter online antisocial behavior, and therefore we felt the need to explore all the available machine learning algorithms to get an optimum result with our dataset.

For deep learning algorithms, we used Keras [88], an open-source neural network library in Python. We experimented with the four most popular algorithms for text analysis and classification: CNN, Bidirectional RNN, and Bidirectional LSTM, and GRU. Like the traditional algorithm we used ten-fold cross-validation to train and evaluate our deep learning models. We used word2vec feature extraction for all the four models

3.4 Performance Evaluation

The last stage in our model construction was to evaluate our proposed approach to detect and identify antisocial behavior post from social media. We adopted Accuracy, Precision, F-Measure, and Recall as the evaluation metrics for our algorithms. These metrics are widely used to evaluate performance for both machine learning and deep learning classifiers Field [83-85], and are appropriate for classifying posts and tweets related to antisocial behavior. We constructed one data set for the identification of antisocial behavior posts and non-antisocial behavior post. Therefore, adopting k-fold cross validation approach was imperative and was used. In this approach the collected dataset was arbitrarily apportioned into k partitions. Out of the k partitions, one was reserved as test subset and the others were combined into training subsets. The whole procedure was carried out k times, which was 10-times in our scenario. The results from all these 10 folds were then averaged to indicate an overall algorithm performance.

4. Experiment and Analysis

4.1. Prediction Performance Evaluation with Traditional Machine Learning

Online antisocial behavior is a relatively new area of research. When social media platforms such as Twitter and Facebook started getting traction, they bought in some of the issues along with them. Antisocial behavior is one of them. To the best to our knowledge, there hasn't been much work done to detect and prevent antisocial behavior online. There are studies on cyberbullying and trolling, which can fall under the umbrella term of anti-social; however, not much has been researched on detecting other aspects of such behavior. Using natural language processing and machine learning techniques, we have done a reasonably good job detecting all forms of antisocial behavior. Following are the results from trying five traditional classifiers and using count vectorization. The accuracy we got was relatively high with all the classifiers used. Precision, Recall, and F1 scores were similar with all these algorithms.

Table 2. Vectorization using Word Frequency Feature Method

Classifier	Feature	Accuracy	Precision	Recall	F1 Score
Logistic Regression	WF	99.76%	99.58%	99.66%	99.62%
SVM	WF	99.82%	99.69%	99.73%	99.71%
Random Forest	WF	98.09%	99.20%	94.71%	96.90%
Decision Tree	WF	99.71%	99.51%	99.56%	99.54%
Naïve Bayes	WF	98.84%	98.88%	97.56%	99.04%

All five algorithms detected antisocial behavior with high accuracy and precision. A tweet that was classified as containing antisocial semantics was the one that contained some sort of swear and rude word to upset or annoy someone. Not all the tweets that were classified positive contained swear words. The sentiment, semantic, and context of the text were also considered while manually labeling and deciding whether the tweet represented antisocial behavior. While classifying, some of the tweets were on the borderline or represented more of sarcasm than antisocial behavior. Such tweets were eliminated and were not used. Since this is one of the first studies trying to detect online antisocial behavior in all its forms, we wanted to keep the things simple for our algorithms and model. The tweets on which we had doubts to whether to classify them as positive or negative, were eliminated from the training

and testing dataset. Since Most of our tweets were quite clearly positive or negative, it made the job of classifying algorithms much easier, as there was limited number of words and phrases that our model had to learn to distinguish between positive and negative tweets. This study can be further extended by adding more complex text to classify, even by some human standards. We assume that adding those sorts of tweets will impact the accuracy and precision metrics, however, it will enable our model to generalize better on any data set.

As mentioned above, we tried our classifiers with TF-IDF vectorization as well. The results are shown below. Support Vector Machine showed the best result when used with count vectorization; however, Random Forest was better when TF-IDF vectorization method was used. Overall, we managed to get good results with both vectorization techniques and were able to detect antisocial behavior from Twitter with high accuracy.

Table 3. Vectorization using TF-IDF Feature Method

Classifier	Feature	Accuracy	Precision	Recall	F1 Score
Logistic Regression	TF-IDF	99.48%	99.64%	98.71%	99.17%
SVM	TF-IDF	99.79%	99.77%	99.58%	99.67%
Random Forest	TF-IDF	97.76%	99.31%	94.14%	96.67%
Decision Tree	TF-IDF	99.64%	99.46%	99.40%	99.43%
Naïve Bayes	TF-IDF	93.97%	98.54%	81.55%	99.45%

The following charts show the similarities between accuracy, precision, recall and F1 score using two different vectorization techniques: Word Frequency and TF-IDF. In both cases, the results are highly similar. Reasons for such similar results could be the size of the dataset and the pre-processing techniques that we used. In regard to the size of the data set, even though we had around 55,000 tweets, more tweets could have bought in more variations in the text data. In regard to the pre-processing techniques, we believe our stemmer did a good job truncating all the important words to their roots, assisting both the vectorizing techniques to perform well. As can be seen from the charts below, Support Vector Machine and Logistic Regression performed the best and Naïve Bayes lagged behind in almost every measuring metrics. We propose the use of Support Vector Machine for our model based on its performance on our dataset and its overall credibility dealing with different types of datasets.

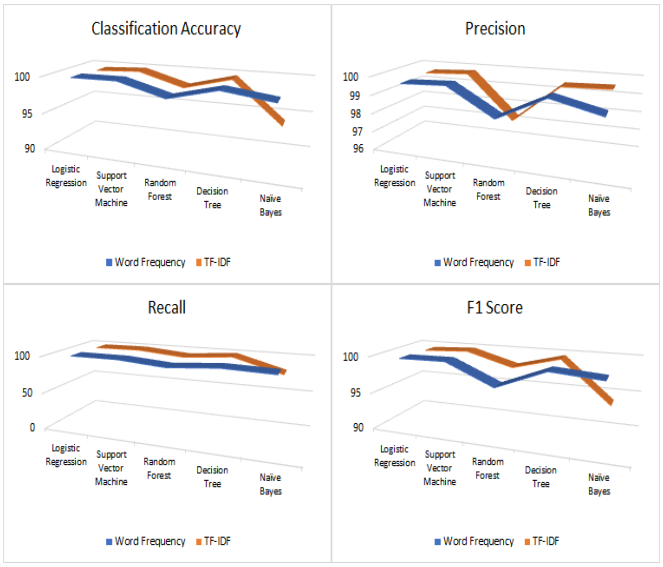

Figure 2. Word Frequency & TF-IDF Feature Vector Comparison

4.2 Performance Evaluation with Deep Learning

This section describes the four deep learning models used to conduct the experiment and the results obtained by using these models. First, we would like to tell the inner working of these models.

- CNN: The first model we tried was CNN and its detailed architecture is demonstrated in [89]. When a pre-processed tweet is fed into a Convolutional neural network, it learns the embedding or the text region internally and captures the semantic coherence information of the tweet. The first layer of CNN is known as the embedding layer and it extracts the n-gram features and stores the word embedding for each word in the text. The convolutional layer contains a disparate number of computational units, each representing an n-gram from the text. Different combinations of n-grams can be experimented with, such as unigram, 2-gram, and 3-gram. The convolutional layers are of variable sizes and the pooling layer transmutes the previous convolutional representation to a higher abstraction level and outputs a fixed-size output. Lastly, a dense layer utilizes the combination of a product feature vector to predict a tweet.
- RNN: The next model that we tried was RNN, and its architecture is described in [90]. RNN handles a flexible-length sequence input and has loops known as the recurrent hidden state. This loop apprehends information from earlier forms. At every step, it receives an input, which is used to update the hidden state. One benefit of RNN over CNN is that its hidden

state integrates and utilizes information from previous time stamps.

- GRU & LSTM: The last two models we experimented with were Bidirectional GRU [91] and Bidirectional LSTM [92]. Both GRU and LSTM are the improved versions of RNNs. They have memory units, that maintain and store historical information, and the gating units that regulate the flow of that information. There is a subtle difference in the architecture of both. LSTMs have three such gates, whereas GRU's have two. We experimented with the advanced version of GRU and LSTM, and these are called Bidirectional GRU and Bidirectional LSTM. Bidirectional features enable these architectures to store both future and historical information. Bidirectional features make GRU and LSTM state-of-art semantic composition machine learning architectures for text classifications. Applications in various fields can be found in [87-89].

For our study, we tried the above four mentioned deep learning architectures. We used 10-fold cross-validation, described in the previous section, to train and evaluate all these models. The results of all these models are shown in the following diagram. We can see the detailed performances in every iteration of the 10-fold cross-validation technique for all four models, and the average of those iterations is also shown.

Table 4. Detailed Deep Learning Classification results with Epoch.

Model	Fold	Epoch	Accuracy	Precision	Recall	F1 Score	Model	Fold	Epoch	Accuracy	Precision	Recall	F1 Score
CNN	1	16	0.998	0.997	0.997	0.997	LSTM	1	22	0.997	0.996	0.996	0.996
	2	9	0.999	0.999	0.999	0.999		2	7	0.998	0.998	0.998	0.998
	3	21	0.998	0.998	0.998	0.998		3	11	0.995	0.994	0.995	0.994
	4	7	0.999	0.999	0.999	0.999		4	6	0.998	0.997	0.998	0.997
	5	7	0.999	0.999	0.998	0.999		5	12	0.997	0.998	0.995	0.996
	6	11	0.999	0.999	0.999	0.999		6	9	0.997	0.998	0.996	0.997
	7	8	0.999	0.999	0.999	0.999		7	16	0.995	0.994	0.994	0.994
	8	24	0.997	0.997	0.997	0.997		8	7	0.995	0.994	0.995	0.994
	9	14	0.998	0.997	0.998	0.998		9	13	0.997	0.997	0.996	0.996
	10	14	0.999	0.999	0.998	0.999		10	8	0.998	0.997	0.998	0.998
AVERAGE		**13.1**	**0.999**	**0.998**	**0.998**	**0.998**	**AVERAGE**		**11.1**	**0.997**	**0.996**	**0.996**	**0.996**
RNN	1	9	0.996	0.996	0.994	0.995	GRU	1	36	0.996	0.996	0.994	0.995
	2	10	0.996	0.997	0.994	0.995		2	7	0.998	0.998	0.998	0.998
	3	11	0.992	0.994	0.988	0.991		3	15	0.996	0.995	0.995	0.995
	4	10	0.997	0.997	0.997	0.997		4	8	0.997	0.997	0.997	0.997
	5	11	0.995	0.994	0.995	0.994		5	6	0.997	0.998	0.995	0.996
	6	10	0.998	0.998	0.997	0.997		6	10	0.997	0.997	0.997	0.997
	7	8	0.998	0.998	0.998	0.998		7	9	0.997	0.998	0.995	0.996
	8	10	0.996	0.996	0.995	0.995		8	9	0.996	0.995	0.995	0.995
	9	10	0.995	0.995	0.994	0.994		9	9	0.997	0.996	0.997	0.996
	10	7	0.998	0.998	0.997	0.997		10	10	0.995	0.994	0.995	0.994
AVERAGE		**9.6**	**0.996**	**0.996**	**0.995**	**0.995**	**AVERAGE**		**11.9**	**0.997**	**0.996**	**0.996**	**0.996**

The above figure compares the accuracy, precision, recall, and F1 scores. It also shows the Epoch, the number of cycles the algorithm went through to learn from the training set. The lower epoch may represent the undertrained model, and higher epoch indicates overfitting. Epoch between10-25 is considered a good outcome. We can see that the average epoch for our study for all these models lies between 9.6 – 13.1. This is an indication that our models learned early on utilizing feature vectors. The following table presents the same results in a more compressed form, however, it shows only the averages, instead of every fold, for all four models used. It can be seen that the accuracy and precision of all these models are close to 100%.

Table 5. Deep Learning Model Evaluation

Deep Learning Model	Feature	Accuracy	Precision	Recall	F1 Score
CNN	Word2Vc	99.86%	99.84%	99.83%	99.83%
LSTM	Word2Vc	99.66%	99.62%	99.60%	99.61%
RNN	Word2Vc	99.61%	99.62%	99.48%	96.67%
GRU	Word2Vc	99.66%	99.63%	99.58%	99.60%

4.3 Traditional Machine Learning and Deep Learning Comparison

In this study, we used both traditional machine learning and deep learning models. Both performed well on our data set of tweets; however, deep learning outperformed traditional methods marginally. The explanation for this outperformance is that deep learning models, unlike most traditional machine learning algorithms, can learn a text's semantics. As explained in the earlier section, these models have memory units and gates that can store and relay such information between different layers of architecture. These units and gates can enable complex data to be stored and communicated within the network, making it possible to handle even the large-scale information and assisting in learning. Learning features such as WC/TF and TI-IDF used in traditional methods cannot store and pass information, and they rely mainly on words and the number of occurrences of these words. SVM was the best performer from the traditional algorithms, and from the deep learning algorithms, CNN outperformed all the other algorithms. So, to build a model based on natural language processing and machine learning techniques, we propose Convolutional Neural Network architecture to classify tweets automatically on a large scale.

4.4 Semantic Coherence Analysis

This sub-section examines the data set to identify essential words in ASB and non-ASB tweets. ASB tweets contain mostly rude, forbidden, and taboo words, representing negative semantics and sentiments. Words such as F**K, mother**k, crime, smoke, lawless, screaming, bitch, fight, nigga, and enforcement are the most prevalent. These are not polite words and are usually avoided in social settings.

One will only use such words during a daily conversation if the intention is to offend and to manifest ASB.

On the other hand, non-ASB tweets are filled with encouraging words such as respect, others, like, beliefs, help, grateful, religion, respected, etc. The contrast in the use of words in both classes can be seen from the word cloud in figure 2. The words in large font are the ones that are prominent in both categories.

Figure 3. Word Cloud Comparison for ASB words and non-ASB words.

There is a clear distinction between the type of words associated with ASB and non-ASB tweets. Some words appear more often in antisocial tweets than in non-antisocial tweets, and other words predominantly do not appear in these antisocial tweets. Traditional machine learning algorithms rely mainly on the meaning of words and how often these words appear in a text, and TF and TF-IDF feature extraction techniques depend on words' meaning. So even though these techniques learn to separate ASB and non-ASB tweets in the experiments we performed, they still need to capture the semantic relationship between these words fully. Deep learning, on the contrary, addresses this issue using the word embedding feature vector technique. In this technique, each word is represented by a feature vector that captures the semantics of a piece of text. Because of this ability of deep learning models, they can learn better from the training data and hence can perform better when implemented.

We also analyzed these words' support in identifying their corresponding class and their likelihood of occurrence in these classes. We picked up common, occurring words from both categories and calculated their support toward identifying tweets in that class. Some of the top words that occurred more than others are shown in the table-5 along with their percentage of occurrence and Z-scores. We performed a Z-test with a p-value <=.05 to establish statistical significance in the event difference.

We established from the experiments that taboo words such as shit, bitch, f**k occurred more often in antisocial tweets than non-anti-social tweets. People who exhibit antisocial behavior online often use taboo words. Some examples of tweets are: (1) *fuck you man, I am goona smash your ugly face right now*. (2) *I WOULD STEAL ALL OF THIS SHIT, I just cant get away wit it. Fuck the law my nigga*. (3) *Don't fuck no bitch that's fucking with your dawg, that's law. If you come up don't forget about your dawgs, that's law. I'm a street nigg@ so it's fuck the law If you broke nigg@ that should be against the lawâ*

In addition, words such as 'Smoke,' 'Law,' 'System,' 'Scared,' and 'broke' are more likely than not to appear in antisocial behavior. These words appear in tweets when individuals posting them claim to have broken the law, scared others, smoke weed, etc. Examples are: 1) *also yes . i know i shouldnt be going 60 in a 35 . fuck the law.* 2) *me & the dogs smoking nothing but nasty *cough cough*. fuck The law and whoever asking.*

Some tweets mentioned when an individual would break the law, hurt someone badly, and threaten others. A few examples are 1*) me & the dogs smoking nothing but nasty *cough cough*. fuck The law and whoever asking*. 2) *nah, your daddy is a real nigga, not 'cause he is hard. Not because he lived a life of crime and sat behind some bars. Cos he'll do this again for ya l!. 3) if you all gonna do what you always do, I'll be killing ya all one by one!*

Table 6. Significant difference in Occurrence of Prominent Words

Words	Anti-Social	Non-Anti-Social	Difference	Z-Score	P-Value
Fuck	0.745	0.010	0.735	95.660	0.0000
Shit	0.450	0.020	0.430	64.048	0.0000
Bitch	0.300	0.010	0.290	50.632	0.0000
Law	0.350	0.050	0.300	47.352	0.0000
Smoke	0.310	0.030	0.280	47.079	0.0000
Kill	0.260	0.010	0.250	46.229	0.0000
Court	0.240	0.010	0.230	43.947	0.0000
System	0.280	0.030	0.250	43.630	0.0000
Hit	0.290	0.040	0.250	42.531	0.0000
Scared	0.230	0.011	0.219	42.510	0.0000
Crime	0.240	0.020	0.220	41.327	0.0000
Broke	0.240	0.020	0.220	41.327	0.0000
Enforcement	0.210	0.010	0.200	40.394	0.0000
Screaming	0.265	0.040	0.225	39.521	0.0000
Nigga	0.180	0.005	0.175	38.178	0.0000
Lawless	0.251	0.050	0.201	35.489	0.0000
Fight	0.220	0.061	0.159	28.877	0.0000
Rules	0.290	0.200	0.090	13.189	0.0000
Like	0.267	0.200	0.067	9.980	0.0000
Want	0.240	0.190	0.050	7.669	0.0000

Words presented in Table-6 have exhibited some features in distinguishing antisocial and non-antisocial tweets. Solely relying on the term frequency may not be a very effective way of classifying these tweets automatically. The justification is that some taboo, bad, and threatening words are often used in non-antisocial tweets to spread awareness or report somebody to authorities. A good classification model should consider the semantic relationship of words in a piece of text rather than relying solely on word count, the approach common in traditional machine learning algorithms using TF-IDF and the Bag of Word approach. We in our experiments, implemented word

embedding features in deep learning, in which a vector feature of 300 dimensions represented each word. Words with similar meanings usually have similar feature vector forms. These 300 dimensions captured the semantics of the tweets and the words used in them. A term used in two scenarios may represent a different meaning if the contexts of these scenarios are disparate. Furthermore, vector features of other but similar words may appear alike and display a strong correlation. To help understand this concept better, we have presented a visualization of the correlation between some of the ordinary occurring words in our dataset.

In figure 3, we have an identical set of words on both the x-axis and y-axis. The 289 small-colored squares represent correlations of words with other words in the figure. The diagonal from the top left to bottom right, made up of dark brown squares, shows the correlation of a term with itself. The darker the color, the stronger the correlation; on the contrary, the lighter the color weakens the correlation between words. As can be seen from the image, the word 'happy' has a high correlation with the words 'amazing,' 'grateful,' and 'thanks.' Similarly, 'asshole' correlates highly with the words 'bitch', 'fuck', and 'shit.' The expression 'behavior' is correlated to 'attitude,' and 'love' is related to 'happy,' 'grateful,' and' thanks.' The white spots within the figure show the opposite. Words 'happy,' 'amazing,' 'grateful,' and 'thanks' have no correlation with 'bitchass,' which shows that these words fall in the opposite semantic bucket to the term 'bitchass.' Light-color squares show no or weak correlation between words.

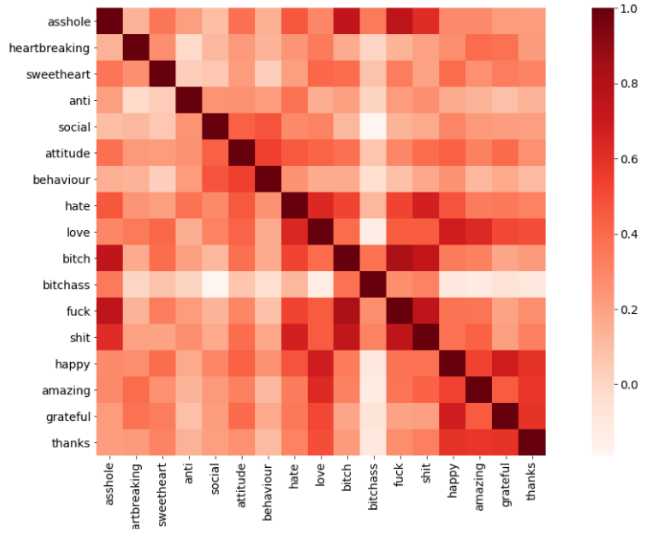

Figure 4. Sample word correlation.

The word embedding features of deep learning models can capture not just the actual meaning of words in a piece of text but also the context in which these words are used, enabling these models to perform better when compared with traditional machine learning models. The point to note here is that about our tweet dataset, the performance of deep learning models is slightly better than the conventional models because the models had to deal with fewer words due to the limited size of tweets, which is a maximum of 280-words, however, when we compare the same models on more extensive texts such as paragraphs or even a large documents, the difference between the performance widens, and deep learning models perform way better than the traditional models.

5. Conclusion & Future Work

This research introduces a data-driven approach to detect and prevent antisocial behavior online. Twitter is responsible for controlling its platform from becoming a breeding ground for antisocial behavior. Similarly, some other online outlets also enable the spread of antisocial semantics that plague the idea of freedom of speech online. It obstructs constructive discussion and leads to many users abandoning participation. At this stage, most of these platforms rely on users reporting such antisocial behavior to these platforms instead of automated detection, which is imperative if the prevention has to work on scale. These platforms may have some measures in place to prevent antisocial behavior online. However, these should be more effective. We, in this research, proposed an approach based on natural language processing and deep learning techniques that can enable online platforms to detect and restrict antisocial behavior proactively. As can be seen, by our results, our model can detect antisocial behavior on Twitter with very high accuracy. The model can be integrated into an online system to depict such behavior on a live data stream. Once detected, appropriate action can be taken, such as deleting the tweet or blocking the user to prevent future incidents.

In this study, we have explored data mainly from Twitter. Further studies can be conducted by collecting data from various online platforms. The diversity of data used will enable models to learn and perform better. Furthermore, we would like to explore other personality and behavior disorders that fall under the same category as antisocial behavior. Diagnostic criteria for these disorders overlap in some instances and can present a challenge in training a model to classify and distinguish these disorders with high accuracy and precision. Future work can also be done by categorizing tweets into different antisocial behaviors and contexts. Depending on the seriousness of the situation, this may lead to offering help to victims by notifying authorities of the dire circumstances. Despite the findings and results, our work has a few limitations. The size of our data set is moderate due to the labor-intensive job of manually labeling tweets. It consists of around 55,000 tweets; however, a larger dataset could have bought in more diversity regarding the feature words and phrases that our algorithms used to learn from. Secondly, we used around 30 different terms to search for tweets on Twitter.

Once these tweets were collected, they were labeled antisocial or non-antisocial. In our subsequent study, we would like to increase the number of these phrases to at least 100. This will bring more diversity to words, phrases, contexts, semantics, and scenarios used to train our classifier. Nevertheless, the findings and the results are valuable in guiding further antisocial behavior studies from social media data using a deep learning approach.

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 *Icwsm*, 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. Woeckener *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.

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

[29] J. M. Beyers, R. Loeber, P.-O. H. Wikström, and M. Stouthamer-Loeber, "What predicts adolescent violence in better-off neighborhoods?," *Journal of Abnormal Child Psychology*, vol. 29, no. 5, pp. 369-381, 2001.

[30] D. L. Haynie, E. Silver, and B. Teasdale, "Neighborhood characteristics, peer networks, and adolescent violence," *Journal of Quantitative Criminology*, vol. 22, no. 2, pp. 147-169, 2006.

[31] T. Huang, Y.-J. Gong, S. Kwong, H. Wang, and J. Zhang, "A niching memetic algorithm for multi-solution traveling salesman problem," *IEEE Transactions on Evolutionary Computation*, 2019.

[32] R. Singh, Y. Zhang, H. Wang, Y. Miao, and K. Ahmed, "Deep learning for antisocial behaviour analysis on social media," in *2020 24th International Conference Information Visualisation (IV)*, 2020: IEEE, pp. 428-434.

[33] T. Braga, O. Cunha, and Â. Maia, "The enduring effect of maltreatment on antisocial behavior: A meta-analysis of longitudinal studies," *Aggression and violent behavior*, 2018.

[34] R. Singh *et al.*, "A Framework for Early Detection of Antisocial Behavior on Twitter Using Natural Language Processing," in *Conference on Complex, Intelligent, and Software Intensive Systems*, 2019: Springer, pp. 484-495.

[35] V. J. Bland and I. Lambie, "Does childhood neglect contribute to violent behavior in adulthood? A review of possible links," *Clinical psychology review*, 2018.

[36] E. Anderson, "The code of the streets," *Atlantic monthly*, vol. 273, no. 5, pp. 81-94, 1994.

[37] E. Aisenberg and T. Herrenkohl, "Community violence in context: Risk and resilience in children and families," *Journal of interpersonal violence*, vol. 23, no. 3, pp. 296-315, 2008.

[38] D. Baskin and I. Sommers, "Exposure to community violence and trajectories of violent offending," *Youth violence and juvenile justice*, vol. 12, no. 4, pp. 367-385, 2014.

[39] S. Javdani, J. Abdul-Adil, L. Suarez, S. R. Nichols, and A. D. Farmer, "Gender differences in the effects of community violence on mental health outcomes in a sample of low-income youth receiving psychiatric care," *American journal of community psychology*, vol. 53, no. 3-4, pp. 235-248, 2014.

[40] E. R. Kimonis, L. C. Centifanti, J. L. Allen, and P. J. Frick, "Reciprocal influences between negative life events and callous-unemotional traits," *Journal of abnormal child psychology*, vol. 42, no. 8, pp. 1287-1298, 2014.

[41] Z. Walsh *et al.*, "Socioeconomic-status and mental health in a personality disorder sample: The importance of neighborhood factors," *Journal of personality disorders*, vol. 27, no. 6, pp. 820-831, 2013.

[42] L. S. Wakschlag, K. E. Pickett, E. Cook Jr, N. L. Benowitz, and B. L. Leventhal, "Maternal smoking during pregnancy and severe antisocial behavior in offspring: a review," *American journal of public health*, vol. 92, no. 6, pp. 966-974, 2002.

[43] P. A. Brennan, E. R. Grekin, and S. A. Mednick, "Maternal smoking during pregnancy and adult male criminal outcomes," *Archives of general psychiatry*, vol. 56, no. 3, pp. 215-219, 1999.

[44] D. M. Fergusson, L. J. Woodward, and L. J. Horwood, "Maternal smoking during pregnancy and psychiatric adjustment in late adolescence," *Archives of general psychiatry*, vol. 55, no. 8, pp. 721-727, 1998.

[45] C. L. Gibson and S. G. Tibbetts, "Interaction between maternal cigarette smoking and Apgar scores in predicting offending behavior," *Psychological Reports*, vol. 83, no. 2, pp. 579-586, 1998.

[46] P. Räsänen, H. Hakko, M. Isohanni, S. Hodgins, M.-R. Järvelin, and J. Tiihonen, "Maternal smoking during pregnancy and risk of criminal behavior among adult male offspring in the Northern Finland 1966 Birth Cohort," *American Journal of Psychiatry*, vol. 156, no. 6, pp. 857-862, 1999.

[47] L. S. Wakschlag and S. L. Hans, "Maternal smoking during pregnancy and conduct problems in high-risk youth: a developmental framework," *Development and psychopathology*, vol. 14, no. 2, pp. 351-369, 2002.

[48] L. S. Wakschlag, B. B. Lahey, R. Loeber, S. M. Green, R. A. Gordon, and B. L. Leventhal, "Maternal smoking during pregnancy and the risk of conduct disorder in boys," *Archives of general psychiatry*, vol. 54, no. 7, pp. 670-676, 1997.

[49] M. M. Weissman, V. Warner, P. J. Wickramaratne, and D. B. Kandel, "Maternal smoking during pregnancy and psychopathology in offspring followed to adulthood," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 38, no. 7, pp. 892-899, 1999.

[50] C. E. Herbison *et al.*, "Low intake of B-vitamins is associated with poor adolescent mental health and behaviour," *Preventive medicine*, vol. 55, no. 6, pp. 634-638, 2012.

[51] A. Binns, "DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities," *Journalism Practice*, vol. 6, no. 4, pp. 547-562, 2012. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1080/17512786.2011.648988?needAccess=true.

[52] E. E. Buckels, P. D. Trapnell, T. Andjelovic, and D. L. Paulhus, "Internet Trolling and Everyday Sadism: Parallel Effects on Pain Perception and Moral Judgment," *Journal of personality*, 2018.

[53] R. B. Cialdini and N. J. Goldstein, "Social influence: Compliance and conformity," *Annu. Rev. Psychol.*, vol. 55, pp. 591-621, 2004.

[54] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *CSCW: proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, 2017, vol. 2017: NIH Public Access, p. 1217.

[55] R. Singh, Y. Zhang, H. Wang, Y. Miao, and K. Ahmed, "Antisocial Behaviour Analyses Using Deep Learning," in *International Conference on Health Information Science*, 2020: Springer, pp. 133-145.

[56] M. Peng *et al.*, "Mining event-oriented topics in microblog stream with unsupervised multi-view hierarchical embedding," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 3, pp. 1-26, 2018.

[57] S. Park, E.-Y. Na, and E.-m. Kim, "The relationship between online activities, netiquette and cyberbullying," *Children and youth services review*, vol. 42, pp. 74-81, 2014.

[58] C. Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions," ed: Walter de Gruyter GmbH & Co. KG, 2010.

[59] C. Chelmis, D.-S. Zois, and M. Yao, "Mining patterns of cyberbullying on twitter," in *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, 2017: IEEE, pp. 126-133.

[60] M. C. McHugh, S. L. Saperstein, and R. S. Gold, "OMG U# Cyberbully! An exploration of public discourse about cyberbullying on twitter," *Health Education & Behavior*, p. 1090198118788610, 2018.
[61] P. Lee, "Expanding the Schoolhouse Gate: Public Schools (K-12) and the Regulation of Cyberbullying," *Utah L. Rev.*, p. 831, 2016.
[62] A. E. Fahy, S. A. Stansfeld, M. Smuk, N. R. Smith, S. Cummins, and C. Clark, "Longitudinal associations between cyberbullying involvement and adolescent mental health," *Journal of Adolescent Health*, vol. 59, no. 5, pp. 502-509, 2016.
[63] B. W. Fisher, J. H. Gardella, and A. R. Teurbe-Tolon, "Peer cybervictimization among adolescents and the associated internalizing and externalizing problems: a meta-analysis," *Journal of youth and adolescence*, vol. 45, no. 9, pp. 1727-1743, 2016.
[64] K. N. Wang, J. S. Bell, E. Y. Chen, J. F. Gilmartin-Thomas, and J. Ilomäki, "Medications and prescribing patterns as factors associated with hospitalizations from long-term care facilities: a systematic review," *Drugs & aging*, vol. 35, no. 5, pp. 423-457, 2018.
[65] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in human behavior*, vol. 26, no. 3, pp. 277-287, 2010.
[66] R. Didden *et al.*, "Cyberbullying among students with intellectual and developmental disability in special education settings," *Developmental neurorehabilitation*, vol. 12, no. 3, pp. 146-151, 2009.
[67] J. Juvonen and E. F. Gross, "Extending the school grounds?—Bullying experiences in cyberspace," *Journal of School health*, vol. 78, no. 9, pp. 496-505, 2008.
[68] T. Beran and Q. Li, "The relationship between cyberbullying and school bullying," *The Journal of Student Wellbeing*, vol. 1, no. 2, pp. 16-33, 2008.
[69] R. M. Kowalski and S. P. Limber, "Psychological, physical, and academic correlates of cyberbullying and traditional bullying," *Journal of Adolescent Health*, vol. 53, no. 1, pp. S13-S20, 2013.
[70] M. Campbell, B. Spears, P. Slee, D. Butler, and S. Kift, "Victims' perceptions of traditional and cyberbullying, and the psychosocial correlates of their victimisation," *Emotional and Behavioural Difficulties*, vol. 17, no. 3-4, pp. 389-401, 2012.
[71] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, "Cyberbullying: Review of an old problem gone viral," *Journal of adolescent health*, vol. 57, no. 1, pp. 10-18, 2015.
[72] C. Wang *et al.*, "A Novel Evolutionary Algorithm with Column and Sub-Block Local Search for Sudoku Puzzles," *IEEE Transactions on Games*, 2023.
[73] M. N. A. Tawhid, S. Siuly, K. Wang, and H. Wang, "Automatic and Efficient Framework for Identifying Multiple Neurological Disorders From EEG Signals," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 76-86, 2023.
[74] Y. Zhao, H. Li, and S. Yin, "A Multi-channel Character Relationship Classification Model Based on Attention Mechanism," *Int. J. Math. Sci. Comput.(IJMSC)*, vol. 8, pp. 28-36, 2022.
[75] J. Yin, M. Tang, J. Cao, M. You, H. Wang, and M. Alazab, "Knowledge-driven cybersecurity intelligence: software vulnerability co-exploitation behaviour discovery," *IEEE Transactions on Industrial Informatics*, 2022.
[76] K. Suzuki, R. Asaga, A. Sourander, C. W. Hoven, and D. Mandell, "Cyberbullying and adolescent mental health," *International journal of adolescent medicine and health*, vol. 24, no. 1, pp. 27-35, 2012.
[77] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychological bulletin*, vol. 140, no. 4, p. 1073, 2014.
[78] C. Hay and R. Meldrum, "Bullying victimization and adolescent self-harm: Testing hypotheses from general strain theory," *Journal of youth and adolescence*, vol. 39, no. 5, pp. 446-459, 2010.
[79] K. Hawton, K. Rodham, and E. Evans, *By their own young hand: Deliberate self-harm and suicidal ideas in adolescents*. Jessica Kingsley Publishers, 2006.
[80] M. Vajani, J. L. Annest, A. E. Crosby, J. D. Alexander, and L. M. Millet, "Nonfatal and fatal self-harm injuries among children aged 10–14 years—United States and Oregon, 2001–2003," *Suicide and Life-Threatening Behavior*, vol. 37, no. 5, pp. 493-506, 2007.
[81] E. K. Englander and A. M. Muldowney, "Just Turn the Darn Thing Off: Understanding Cyberbullying," in *Proceedings of persistently safe schools: The 2007 national conference on safe schools*, 2007.
[82] D. C. Kerr, L. D. Owen, K. C. Pears, and D. M. Capaldi, "Prevalence of suicidal ideation among boys and men assessed annually from ages 9 to 29 years," *Suicide and Life-Threatening Behavior*, vol. 38, no. 4, pp. 390-402, 2008.
[83] *The Sage handbook of qualitative research / ed. by Norman K. Denzin*. Los Angeles [u.a.]: Sage, 2011.
[84] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39-57, 5/24/ 2017, doi: http://doi.org/10.1016/j.neucom.2017.01.078.
[85] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining. Practical Machine Learning Tools and Techniques*. Burlington : Elsevier Science
3rd ed., 2011.
[86] L. Teng and Y. Qiao, "BiSeNet-oriented context attention model for image semantic segmentation," *Computer Science and Information Systems*, no. 00, pp. 40-40, 2022.
[87] D. Jiang, H. Li, and S. Yin, "Speech Emotion Recognition Method Based on Improved Long Short-term Memory Networks," *International Journal of Electronics and Information Engineering*, vol. 12, no. 4, pp. 147-154, 2020.
[88] F. Chollet, "Keras," ed, 2015.
[89] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058*, 2014.
[90] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1017-1024.
[91] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[92] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE workshop on automatic speech recognition and understanding*, 2013: IEEE, pp. 273-278.