# Breaking Down the Language of Hate: NLP-based Classification of Antisocial Behaviour on Twitter

Anishamol Abraham
*Assisstant Professor, Dept. of CSE*
*Amal Jyothi College of Engineering*
Kottayam, India
anishamolabraham@amaljyothi.ac.in

Ryan Savio Shibu
*UG Student, Dept. of CSE*
*Amal Jyothi College of Engineering*
Kottayam, India
ryansavioshibu2023@cs.ajce.in

Sidharth Subin
*UG Student, Dept. of CSE*
*Amal Jyothi College of Engineering*
Kottayam, India
sidharthsubin2023@cs.ajce.in

Sharon Mathew Sunny
*UG Student, Dept. of CSE*
*Amal Jyothi College of Engineering*
Kottayam, India
sharonmathewsunny20232023@cs.ajce.in

Shaun S. Kurien
*UG Student, Dept. of CSE*
*Amal Jyothi College of Engineering*
Kottayam, India
shaunskurien2023@cs.ajce.in

***Abstract*—Antisocial behaviour can take many different forms, such as violence, disregard for the safety of others, lack of remorse, unlawful activity, etc. It affects the experience of social media negatively, for the average user. A sentiment analysis based approach for identifying and classifying online antisocial behaviour is presented in the paper (ASB), which would help combat the prevalent anti social behaviour across Twitter, as a social media platform. The proposed framework utilizes data collected from repositories, an NLP technique - Sentimental Analysis, and assigns polarity value to classify tweets into 5 classes of ASB. The framework can be used as a tool for social media monitoring, content moderation, and online safety.**

## I. Introduction

A personality disorder is a persistent pattern of inner experience and behaviour that starts in adolescence or early adulthood, is pervasive and rigid, lasts for a long period, remains stable across time, and produces distress or impairment. It dramatically deviates from what an individual might expect in their culture.

Online antisocial behavior refers to antisocial behavior displayed on social media, blogs, news channels, and other online platforms used for expressing opinions and sharing information. Individuals with antisocial personality often display disregard for others and the law, use abusive and threatening language, and engage in socially unacceptable behavior when using these channels. Little has been done to deter such behavior online. Some platforms tolerate this behavior in the name of freedom of speech, but there is a fine line between free speech and unacceptable social behavior.

To confront antisocial behaviour online, it is imperative to understand its etiology. There are many factors that can lead to a person developing and manifesting ASB. Some of these factors are parental-rejection, maternal-depression, physical neglect, genetic-influence, and poor nutrition intake, etc.

## II. Literature Survey

The work by R. Singh [1] proposes a framework that uses features such as sentiment analysis, topic modeling, and named entity recognition to identify potentially harmful tweets and their authors.

The work by M.S. Neethu [2] proposes the use of machine learning techniques for sentiment analysis in Twitter, evaluating their performance in accurately classifying positive, negative, and neutral tweets.

The work by E.V Altay [3] detecting cyberbullying in social networks by analyzing textual features such as sentiment, emotion, and syntactic structures.

The work by K.Reynolds [4] application of machine learning techniques for identifying cyberbullying behavior on social media platforms by analyzing text-based features and evaluating the performance of various classification models.

The work by M. Dadvar [5]an approach to improve cyberbullying detection on social media platforms by incorporating user context information such as age, gender, and location in addition to analyzing the textual content, and evaluates the effectiveness of this approach using machine learning techniques.

The work by S. Subramani [6] a deep learning-based approach for identifying domestic violence crisis signals in Facebook posts, using features such as linguistic cues and temporal patterns to predict the presence of domestic violence, and evaluates the effectiveness of the approach.

The work by S. Kiritchenko [7] sentiment analysis of short, informal texts such as tweets and text messages, and evaluates the effectiveness of various techniques such as lexicon-based and machine learning-based approaches for accurately classifying the sentiment of these texts.

The work by Y. Kim [8] his paper introduces a convolutional neural network (CNN) architecture for sentence classification tasks, such as sentiment analysis and topic categorization, and evaluates the performance of the model on various datasets.

The work by N. Kalchbrenner [9] proposes a convolutional neural network (CNN) model for sentence modeling, which learns to capture n-gram features of varying lengths and performs well on tasks such as sentiment analysis, paraphrase detection, and natural language inference.

The work by P.Liu [10] presents a recurrent neural network (RNN) model for text classification, which incorporates multi-task learning to jointly learn from multiple related tasks, such as sentiment analysis and topic classification, and evaluates the effectiveness of the approach on various datasets.

The work by B. Gambäck [11] explores the use of convolutional neural networks (CNNs) for hate speech classification on social media, utilizing both word and character-level embeddings to capture semantic and syntactic information, and evaluates the performance of the approach on a hate speech dataset.

The work by S. Agrawal [12] explores the use of deep learning techniques for detecting cyberbullying across multiple social media platforms by employing a combination of textual, visual, and social network features, and evaluates the performance of the approach on a multi-platform dataset.

The work by K. Nigam [13] proposes the use of maximum entropy models for text classification, which learn to assign the most probable label to a given text instance based on a set of features, and evaluates the performance of the approach on various text classification tasks, including topic categorization and sentiment analysis.

The work by P. Badjatiya [14] explores the use of deep learning techniques for hate speech detection in tweets, by proposing a convolutional neural network (CNN) architecture that can learn to identify patterns and features in text, and evaluates the performance of the approach on a publicly available hate speech dataset.

The work by C. Caragea [15] proposes the use of convolutional neural networks (CNNs) for identifying informative messages during disaster events, by leveraging both the textual content and visual features of social media messages, and evaluates the effectiveness of the approach on a dataset of tweets related to natural disasters.

The work by D. T. Nguyen [16] classifies messages into different categories based on their relevance to crisis situations, and evaluates the effectiveness of the approach on a dataset of tweets related to real-world crisis events.

The work by S. Subramani [17] explores the use of deep learning techniques for identifying different classes of domestic violence from online posts, by proposing a deep neural network architecture that can learn to extract relevant features from text data, and evaluates the performance of the approach on a dataset of online posts related to domestic violence.

The work by G. Gkotsis [18] proposes a method for characterizing mental health conditions in social media using informed deep learning, by leveraging domain-specific knowledge to improve the interpretability and accuracy of deep learning models, and evaluates the effectiveness of the approach on a dataset of social media posts related to mental health.

The work by J. Trofimovich [19] presents a comparison of different neural network architectures for sentiment analysis of Russian tweets, by evaluating the performance of various models on a dataset of Russian tweets labeled with positive, negative, and neutral sentiments.

The work by R. S. Mohana [20] proposes an approach for predicting public emotions using machine learning algorithms applied to tweets collected from Twitter API. The authors preprocess the tweets and extract features using TF-IDF, and use several algorithms to classify tweets into positive, negative, or neutral emotions. The authors achieve promising results on a dataset of 5000 tweets, demonstrating the usefulness of Twitter-based sentiment analysis for real-time prediction of public emotions, such as for opinion mining and market research.

## III. Proposed Method

A methodology was developed and employed to ensure the successful attainment of the required outcome. This methodology is depicted in Fig. 1.

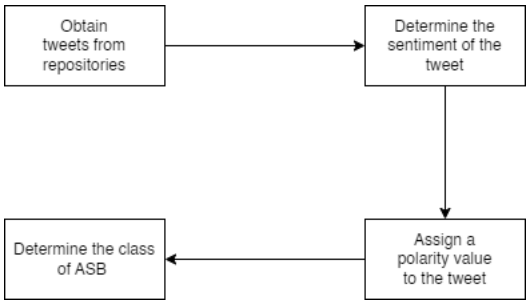


Fig. 1. Methodology

The actual classification of a tweet and identification of antisocial behaviour consists of four basic steps, those being:

- Data Collection
- Sentiment Analysis
- Assign Polarity Value
- Identify Class of ASB

a) Data Collection: Due to some recent changes to Twitter's API, access to their analytics and tweets were locked behind a paywall. To obtain access to tweets, a third-party website, 'kaggle.com' was used to obtain datasets(Fig. 2). These datasets contain raw twitter data, in the form of .csv files, which is then imported into the program and then analysed.

b) Sentiment Analysis: Next, the sentiment of these tweets are analysed by the NRCLex package, which classifies these tweets based on the eight main emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) as well as the two sentiments (negative and positive). Once the data has been analysed, the pre-dominant sentiment present in the sentence can be clearly determined, and the overall sentiment/analysis of a sentence can be judged. Sentiment analysis is limited

to the words that are pre defined in the NRCLex library.

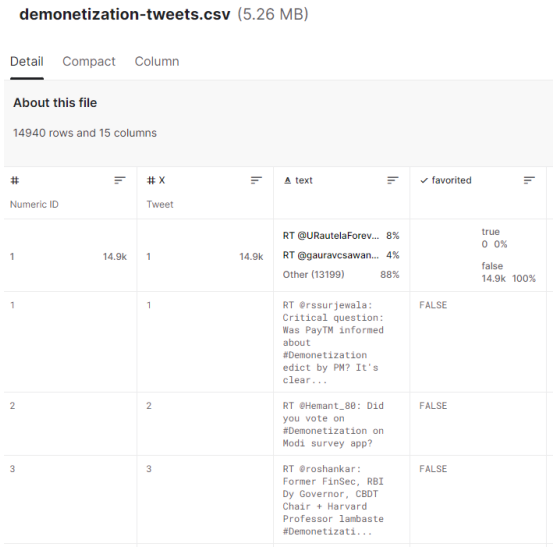


Fig. 2. An example of a kaggle dataset.

c) Assign Polarity Value: Once the sentiment of a sentence has been analysed, the polarity value of the sentence can be generated. The polarity of a sentence is judged by a factor of 'how strong the emotion is'. Say, a sentence with the word 'hate' would register a higher value of negativity, than a sentence with the word 'dislike'. This polarity value is crucial, since this is the value that can assign a value to a sentence, rather than an emotion, making it easier for the program to compute and comprehend.

d) Identify Class of ASB: Anti Social Behaviour is divided into five different classes, each ranging from 0 - 4(Table 1). The higher the class, the more severe the anti social behaviour, with zero being the lowest or non-existent anti social behaviour, and four scoring the highest, or most severe anti social behaviour. Once the tweet is classified into it's respective class, the severity of the ASB prevalent in the tweet can be judged, and further action can be taken by the user to block or restrict the account with the malicious tweet.The following table shows the what each class depicts in the ASB category

Table 1: Classes of ASB

| Class Label | Context |
|---|---|
| Class 0 | Non ASB/General Tweets |
| Class 1 | Failure to conform the social norms |
| Class 2 | Mild ASB found |
| Class 3 | ASB found |
| Class 4 | Lack Of Remose |

## IV. Experiments and Results

The experimental results have confirmed that the tweets are able to be classed into the various class labels which show the extent of antisocial behaviour with a few exceptional cases.Below are given the results of two such cases whereby a negative polarity(Fig. 3) and positive polarity(Fig. 4) is gained.

```
PS D:\NRC> & C:/Users/rsavx/AppData/Local/Microsoft/WindowsApps/python3.11.exe d:/NRC/main.py
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\rsavx\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!

['The', 'world', 'is', 'a', 'dark', 'and', 'dreadful', 'place']

{'dark': ['sadness'], 'dreadful': ['anger', 'anticipation', 'disgust', 'fear', 'negative', 'sadness']}
-0.575
NEGATIVE
Class 2: Mild Anti Social Behaviour
```

Fig. 3. Results for Negative polarity(ASB)

```
PS D:\NRC> & C:/Users/rsavx/AppData/Local/Microsoft/WindowsApps/python3.11.exe d:/NRC/main.py
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\rsavx\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!

['I', 'am', 'feeling', 'happy', 'and', 'satisfied']

{'feeling': ['anger', 'anticipation', 'disgust', 'fear', 'joy', 'negative', 'positive', 'sadness', 'surprise', 'trust'],
}
0.65
POSITIVE
Class 0: No Anti Social Behaviour present
```

Fig. 4. Results for Positive polarity(Not ASB)

## Conclusion

In conclusion, Natural Language Processing has shown promise for multi-class antisocial behavior identification from Twitter data. However, the quality and diversity of the training data, as well as potential biases in the models, remain important areas for improvement.

Further research is necessary to ensure that the models are robust against various forms of bias and generalize well to new data. By accurately identifying and classifying antisocial behavior on Twitter, this framework can help improve the overall quality of user-generated content and promote a safer online environment. Further research can be conducted to expand the framework's capabilities by exploring different preprocessing techniques, feature extraction methods, and machine learning algorithms.

## References


[1] **R. Singh, J. Du, Y. Zhang, H. Wang, Y. Miao, O. A. Sianaki, and A. Ulhaq**, "'A framework for early detection of antisocial behavior on Twitter using natural language processing"

[2] **M. S. Neethu and R. Rajasree**, "Sentiment analysis in Twitter using machine learning techniques"

[3] **E. V. Altay and B. Alatas**, "Detection of cyberbullying in social networks using machine learning methods,"

[4] **K. Reynolds, A. Kontostathis, and L. Edwards** "Using machine learning to detect cyberbullying"

[5] **M. Dadvar, D. Trieschnigg, R. Ordelman, and F. D. Jong**, "Improving cyberbullying detection with user context,"

[6] **S. Subramani, H. Wang, H. Q. Vu, and G. Li**, i, "Domestic violence crisis identification from Facebook posts based on deep learning,"

[7] **S. Kiritchenko, X. Zhu, and S. M. Mohammad**, "Sentiment analysis of short informal texts,"

[8] **Y. Kim**, "Convolutional neural networks for sentence classification,"

[9] **N. Kalchbrenner, E. Grefenstette, and P. Blunsom**, "A convolutional neural network for modelling sentences,"

[10] **P. Liu, X. Qiu, and X. Huang**, g, "Recurrent neural network for text classification with multi-task learning,"

[11] **B. Gambäck and U. K. Sikdar**, "Using convolutional neural networks to classify hate-speech,"
[12] **S. Agrawal and A. Awekar**,r, "Deep learning for detecting cyberbullying across multiple social media platforms,"
[13] **K. Nigam, J. Lafferty, and A. Mccallum**"Using maximum entropy for text classification,"
[14] **P. Badjatiya, S. Gupta, M. Gupta, and V. Varma**, "Deep learning for hate speech detection in tweets,"
[15] **] C. Caragea, A. Silvescu, and A. H. Tapia**,, "Identifying informative messages in disaster events using convolutional neural networks,"
[16] **] D. T. Nguyen, K. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra**, "Robust classification of crisis-related data on social networks using convolutional neural networks,"
[17] **S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, and H. Shakeel**,, "Deep learning for multi-class identification from domestic violence online posts,"
[18] **G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, and R. Dutta**,, "Characterisation of mental health conditions in social media using informed deep learning,"
[19] **J. Trofimovich**,, "Comparison of neural network architectures for sentiment analysis of Russian tweets,"
[20] **R. S. Mohana, S. Kalaiselvi, K. Kousalya**,, "Twitter based sentiment analysis to predict public emotions using machine learning algorithms,"