UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Ketevani Kvirikashvili

# Study of aggressive behavior on social media

Master's Thesis (30 ECTS)

Supervisor(s): Swapnil Mane
Rajesh Sharma
Suman Kundu

Tartu 2023

## Study of cyber-aggressive behavior on social media

**Abstract:**
Antisocial behavior rises since every online environment encourages social engagement. Recently, the expression of aggression in social networks has increased a lot, which also causes a lot of adverse effects, such as mental health problems or some other controversies. Hence we perform the first ever user aggressive behavior analysis on Twitter social media official microblogging site, which has no restriction on aggressive behavior. Using the proposed pipeline, we study the user's aggressive behavior. The pipeline is based on three stages such as data collection, aggression detection, and user profiling. In this study, we detailed analyzed the aggressive behavior of users are depends on their aggressive feeds and events. Further, our analysis revealed that user engagement is higher in aggressive posts.

**Keywords:**
Online Social Media, Aggression Detection, Aggressive Behaviour Analysis

**CERCS:**
P170 Computer science

## Tüübituletus neljandat järku loogikavalemitele

**Lühikokkuvõte:**

Antisotsiaalne käitumine kasvab, kuna iga veebikeskkond soodustab sotsiaalset kaasatust. Viimasel ajal on sotsiaalvõrgustikes palju suurenenud agressiivsuse väljendus, mis põhjustab ka palju kõrvalmõjusid, nagu vaimse tervise probleemid või mõni muu vaidlus. Seetõttu viime läbi esimest korda kasutajate agressiivse käitumise analüüsi Twitteri sotsiaalmeedia ametlikul mikroblogi saidil, millel pole agressiivsele käitumisele piiranguid. Kavandatava torujuhtme abil uurime kasutaja agressiivset käitumist. Konveieri põhineb kolmel etapil, nagu andmete kogumine, agressiooni tuvastamine ja kasutajaprofiilide koostamine. Selles uuringus analüüsisime üksikasjalikult, kuidas kasutajate agressiivne käitumine sõltub nende agressiivsetest voogudest ja sündmustest. Lisaks näitas meie analüüs, et kasutajate seotus on suurem agressiivsete postituste puhul.

**Võtmesõnad:**
Internetis sotsiaalmeedia, agressiooni tuvastamine, agressiivse käitumise analüüs

**CERCS:**
P170 Computer science

# Contents

# 1 Introduction

Online social media has penetrated our lives, and it is one of the main sources to get information from around the world. The information published on social media may impact peoples' personal lives. As time goes on and technology advances, more and more social media becomes popular, and people are more likely to get harmful messages or information from it. Social media could be defined as a forum where users can get information on different topics, discuss and debate about them or seek out the help they need. [Ame15]. Social networks are widespread worldwide and are the best way to establish communication between people. Social network sites(SNS) make life easier for people of almost all professions. Many people benefit from sharing information, communicating with each other, or for other educational purposes. It attracts special attention among scientists because the data it gives could be used for various scientific works, and many things can be predicted by properly using it.

The ability to express one's opinion freely in social networks has caused many obstacles. There is no control, and anyone with internet access can publish/post whatever they wish. One of the most common problems is the expression of aggression. In many cases, Individuals might violate the Freedom of all individuals and respect for fundamental rights. For example, the person can hide his personality and create fake accounts, and by using it, he can express aggression towards anyone. Also, automated social network accounts are called 'bots' with characteristics similar to a real person's profile and have a significant impact on people [HWGD$^+$21]. They can cause aggression in people around a specific topic or story by posting the same things frequently, misleading the users, or having psychological pressure on them. Aggression on social media can be expressed by politicians too, and they can completely change public discourse and it may cause some threats to society. Politicians' comments and their supporters' behavior can cause any threat and escalate political activities, mainly when political candidates use Twitter or any other social media to conduct political debates.

Aggression detection has been particularly popular in recent years. It is one of the most important and new research topics in natural language processing. Many researchers have tried to identify hateful and aggressive posts or comments on social media. People attempted to seek methods for detecting abusive behavior(such as cyberbullying, hateful speech, or aggressive attitudes towards someone) and to prevent them in different ways.[DS21] The concept of aggression has been defined and understood in various ways by researchers and philosophers, and there is no single, definitive explanation. According to researcher [EK19a], aggression may not always be considered a negative action aimed at causing harm, but it can also be a form of defensive behavior. As mentioned, Freud and Lorenz saw aggression as a genetic impulse rather than negative behavior. According to [MD18] study, aggression may not be expressed directly and can instead be hidden. It can be classified into the following three classes: Overtly Aggressive (OAG), Covertly Aggressive (CAG), and Non-aggressive (NAG) and it is not always easy to differentiate

between them.

In this paper, we have analyzed user behavior on aggressiveness. We collected 3,39,390 tweets with respect to 3 events(section 3.2) that happened during the period of collected data. They generated a lot of aggressive social media content, specifically on Twitter. We gathered the data from January to July and included only these three events. Firstly, we partially annotated 1k data with the guidelines discussed in advance. Then, we evaluated annotation quality using inter-annotator agreement. After that, we used a prediction model to predict the sentiment of the rest tweets. This study aims to identify the sentiment of Twitter posts by creating an accurate machine-learning model. The following research question (RQs) have been identified to analyze this issue:

- RQ1: Do my feeds make me aggressive?
- RQ2: Do event feeds makes event specific aggressive?
- RQ3: Do user engagement is more toward aggressive posts?

In this paper, the following sections are as follows: In section 2, we are talking about related work on the aggression and aggression detection model. What method did other researchers use, and what results they had? Section 3 explains how the data was collected and annotated. Section four will describe the methodology of our work. Then in section 5 aggression detection model will be analyzed in detail, as how data was preprocessed and what models were used. Then the next section will explain how the aggression intensity was calculated. Section 7 will provide the details about the users. In sections 8,9,10, each research question will be examined and answered. Then section 11 will briefly discuss the work, and in the final section, we will summarize the research and discuss future work.

# 2 Prior Work

A lot of researchers attempted to seek methods for detecting abusive behavior(such as aggressive attitudes towards someone, cyberbullying, and hateful speech) and to prevent them in different ways [DS21]. In this section, we will analyze various research studies about aggression.

## 2.1 Aggression detection model

The signs of offensive language might be seen not only in social media but in some other places, such as video games, and it increases aggression. The research showed us that it might lead society to public health risk [HG14]. So, to understand why people may behave aggressively, GAM(General aggression model) has been created and developed effective strategies for addressing violence [DAB11]. GAM combined multiple mini theories of aggression into a single, unified model. It means that GAM provided a way to understand how aggression can arise from various motives and how to prevent and reduce it. The GAM model not only explained how to predict aggressive behavior but also illustrated that exposure to violent media can lead to increased aggressive thoughts.[AB18] This method was dominant among other methods in previous years. However, some researchers found that detecting aggression may need to be improved to provide a comprehensive understanding of aggressive behavior [FD12]. So, The research introduced its weaknesses of it and why it could be a better approach.

## 2.2 Methodology of detecting and preventing aggression

Throughout the years, researchers have used many methods and techniques to detect aggression in cyberspace and built many manual checking websites [WHL$^+$22]. These approaches involve BERT [KT20, KSDR21a], Deep-learning approach (RNN-LSTM) [KT20, BDBD20], [SH21] etc.

To detect aggression on Twitter [HCK20] introduced the first real-time detection framework. They used different machine learning methods, such as s Hoeffding Trees, Adaptive Random Forests, and Streaming Logistic Regression. Using the Random forest method, [NNF$^+$19] applied a similar approach to detect hate speech. Another work [SCA20], investigated different machine learning algorithms to detect aggressive behavior, and the Naive Bayes method attained an accuracy of 92% and recall of 95%. Many platforms, such as Instagram, can't detect aggressive comments to someone's posts on social media. So [NBL$^+$19] used various methods, such as the Naïve Bayes Classifier, to classify the comments on Instagram.

One of the researchers [CKB$^+$17] presented their work using a corpus of 1.6 million tweets collected over three months, and they used Feature extraction methods (Network-based, User-based, Text-based) to evaluate their work. They used two methods to assess

workers' reliability: (i) the inter-rater reliability measure and (ii) control cases. And the agreement between them was 0.54. The results showed that network-based features are an excellent way to determine aggressive behavior, while text-based features can't give such accurate results.

Another study [KT20] used a Deep learning approach (RNN-LSTM) and BERT to identify sarcasm in social media. The authors evaluated Twitter and Reddit communication datasets and discovered that the BERT model performed better on both datasets, even if they were small or grammatically incorrect.

In their work, [SH21] developed a corpus using a hierarchical annotation schema. Researchers used several classification algorithms to identify aggressive and non-aggressive behavior. With the combination of CNN and the BiLSTM model, they achieved the perfect f1 score of 0.87 for the identification task and 0.80 for the classification task. In this work, the hierarchical annotation method identified the aggressive behavior and classified it into different categories: Religion, gender, political or verbal aggression.

As discussed previously, a lot of methods exist when it comes to aggression detection in social media. However, there are some obstacles if the text is multilingual. Because some words may have positive meanings in one language, but in another, these words may be considered aggressive. [KSDR21a] developed an LSTM auto encoder-based model only for aggressive comments, and they trained it with non-aggressive comments. Their data was bilingual(English and Hindi), and the increment has been achieved in different social media. Another study was conducted for English, Hindi, and Bangla language and the results revealed that Bert classifiers are the better way to identify aggression in the text.[BDBD20] Another problem that may occur is that some models may not be able to detect the sarcasm in the tweets, so the researchers [AWCM20] introduced a deep learning-based approach in a Hindi-Eglish dataset using the methods such as g CNNs, LSTMs, Bi-directional LSTM, but the Bi-directional LSTM showed the best performance with an accuracy of 78.49 %. [RHY$^+$21] study revealed that logistic regression is the best classification algorithm to identify sarcasm using deep learning features.

## 2.3 Effects of expressing aggression in cyberspace

One of the studies indicated that Social media aggression has a significant effect on people, and expressing negative comments toward other participants increases a tendency to exhibit aggressive behavior from the other side. [EK19b]

Also, If the aggression is towards someone in social media, it may develop into a more severe concept called cyberbullying [GIPS22]. So it is essential to analyze the effects of both cyber aggression and cyberbullying.

The study showed that aggression might be strongly associated with committing cyberbullying among Srilankan adults on Facebook [GP20]. Results indicated that addressing aggression and behavioral issues is vital to decrease this problem.

People, especially children victims of aggressive behavior in cyberspace, may suffer from mental health problems such as depression and anxiety.[SSR21]. So in their studies, [HZZL21] examined how excessive use of social media may cause anxiety, depression, or even suicide. After conducting a cross-sectional STAR questionnaire survey and applying a binary logistic regression model, it has been revealed that 64.32 % of college students suffered from aggression from others through the use of the internet and computer games. Another study also showed that Overindulgence in information and communication technology might have the consequence of cyber-aggression [HCR+20]. Cyberbullying can have indirect effects, such as suicide, and may be more destructive than any direct effects. [ZHB20] So, huge preventive measures must be taken to avoid this problem.

Moreover, [OVM21] shows that cyber aggression may also affect adolescents' quality of life and school satisfaction. So, as it is discussed and experimented, children who are the victims of cyber aggression are more likely to have low performance in school and life satisfaction.[OVM21]. Bullies often try to depersonalize victims and decrease their self-confidence. As it was discussed, gratitude, a feeling to increase moral emotions among adolescents, is a preventive measure.

Sometimes spreading fake news and exposure to it can be a root problem for aggression and the moral disengagement of people.[MHM22]. So, Widely spreading misinformation and reacting to this information may change users' behavior on particular subjects. [WHL+22] in the study found out that users' frequency of tweeting posts increases after exposing the information; however, they also researched that sentiment hasn't changed, but the use of swear words has been increased in some of the groups(target or baseline).

So, As it is visible different techniques exist to understand and detect hate speech and aggression in social media. However, only some studies are conducted on people's behavior after exposure to misinformation. This study focuses on how hateful comments and aggression on social media may affect people's behavior.

# 3 Dataset

## 3.1 The Twitter social network

Twitter is a micro-blogging official trending social networking site. Twitter has a new official policy "Twitter is freedom of speech, but not freedom of reach." According to Twitter guidelines, it does not restrain users from aggressive behavior. It has mechanisms to identify possible sensitive content such as graphic violence, adult content, violent sexual conduct, gratuitous gore, and hateful imagery. The platform does not explicitly consider the text of the post for aggression identification rather, it is based on a manual reporting system. Twitter is restricted to users with up to 280 characters to post. The site is employed based on hashtags which are the representation of specific topics.

## 3.2 Data Collection

We collected several aggressive tweets from the $1^{st}$ of January 2022 to the $15^{th}$ of July 2022. During this period several events happened in this period, which attracted a lot of aggressive tweets around this topic. In particular, we selected the following three events.

1. The youth took to the streets to express their displeasure over the Government's newly launched Agneepath scheme, which led to massive violent protests across India, mainly in the states of Bihar and Uttar Pradesh. The Agneepath scheme aims to reduce ballooning salary and pension bills by recruitment of soldiers into the Army, Navy, and Air Force on a short-term contract basis.
   Possible hashtags: #agneepathprotest #agneepathyojana #AgnipathScheme #agnipathschemeprotest #agnipathschemeprotests #agnipathprotest #Agnipath #AgnipathProtest #AgnipathProtests

2. The exodus of Kashmiri Pandits from the Muslim-majority valley of Jammu and Kashmir resulted from frequent brutal murders and genocide of Kashmiri Hindus or Pandits.
   Possible Hashtags: #StopPakSponsoredTerrorism #KashmirAgainstTerrorism #AakhirKabTak #KashmiriPandits #kashmirihindus

3. A controversial statement by political leader Nupur Sharma led to Hindu-Muslim disputes, massive protests, and violence.
   Possible Hashtags: #MuslimsUnderAttackinIndia #HindusUnderAttack #NupurSharma #NupurSharmaControversy #KanhaiyaLal.

We referenced potential hashtags of each event which are manually selected from Twitter to collect relevant tweets using the Twitter API. Twitter has a non-commercial API version 2 for academic research that allows the collection of up to 10 million tweets in a

| | |
|---|---|
| Time-period | $1^{st}$ Jan 2022 to $15^{th}$ July 2022 |
| Total tweets | 339,390 |
| English tweets | 175,606 |
| Unique users from English tweets | 60,366 |

Table 1. Description about the initial event-related data extraction, such as period, total tweets extracted, number of English-language tweets, and number of unique users associated with the data.

month. Initially, event-related data is extracted (Table 1) which also includes detailed information about the tweets (text, language, retweet, like count, mentions, hashtags, etc.) and users (username, location, profile URL, a profile description, etc.). For aggressive behavior analysis in the same period, we collected all user data (tweets and following information) involved in the events. Then, for further analysis, tweets of the following users are also collected for the same period. Finally, more than 20 million tweets were collected, but the locally stored data was too difficult to handle. So for this experiment, we stored millions of NoSQL data records in a compressed manner using MongoDB NoSQL database.

## 3.3 Annotation

This section presents procedures and guidelines for social media text annotation. Annotation data is the actual Twitter post, which is in English. For annotation, we defined two classes, aggressive (AG) and non-aggressive (NAG). Annotators adhere to a specific ideology that people are aggressive on social media directly and indirectly, i.e., aggressive, by using positive sentiments or non-swear words. Aggression has different targets, such as physical threat, sexual threat or aggression, and identity threat or aggression (gender, geographic, political, ethnic, communal, and racial aggression). Sometimes aggression and abuse co-occur in many instances where we have considered abuse (not banter or teasing) as an aspect of aggression. We follow the previous work of [KRBM18], and [BSK$^+$20] for annotation.

For annotation, we selected four annotators to improve the quality of the annotations. The annotators are from computer science and engineering backgrounds, including two male bachelor students and one male and one female postgraduate student. Annotators individually annotated the same set of data with the guidelines discussed. The post is labeled as NaN if the annotator is ambiguous about the tweet.

We evaluated annotation quality using inter-annotator agreement, which measures how well all annotators can make decisions for a particular class. We used Feiss kappa metrics for inter-annotator agreement evaluation. Feiss kappa is a statistical method used

to measure agreement between multiple raters [Fle71]. Agreement scores range from 0 to 1, where one is perfect, and 0 is no agreement. For our annotated data, the Feiss kappa agreement (reliability) score was 0.7873, which is a substantial agreement.

We used annotated data as a supervised dataset to train and test the aggression detection model. The model's performance does not differ for 5K, and 6K annotated tweets. We decided to stop annotation on about 6K tweets. In the final annotated dataset, the total tweets are 6000, from which 2602 are aggressive (AG) and 3398 are non-aggressive (NAG); Figure 1 shows the statistics of the dataset.

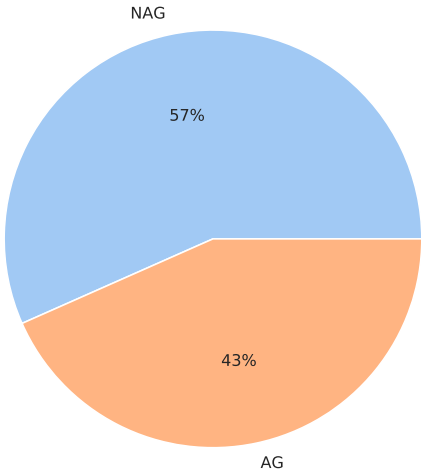



Figure 1. Proportion of Aggressive and Non-aggressive classes in the dataset.

# 4 Methodology

To address our research questions, we need to analyze users' activities by timestamp. So our first task is to prepare an aggressive dataset and collect user data. With the help of the proposed aggression dataset, we build an aggression detection model to predict whether a user's post is aggressive (AG) or non-aggressive (NAG). In order to analyze the user's aggressive behavior, we proposed an evaluation method for the calculation of users' aggression intensity in period l. Aggression intensity is used for user profiling in period l. We used this pipeline flow to study aggressive behavior on social media. A pipeline is composed of the four components listed below:

1. **Collection and preparation of data**: We collected user-wise data for a period of 6 months and created an aggressive dataset using the manual annotations described in Sections 3.2 and 3.3.

2. **Aggression detection**: We trained and tested an aggression detection model for Twitter (in section 5).

3. **Aggression Intensity calculation**: We calculated the aggression intensity score to analyze user's aggressive behavioral activity based on their posts in period l (week/day/hour) (in section 6).

4. **User profiling**: We profile each user with a timeline based on aggressive activity. We used aggression intensity scores to represent user profiles in vector form (in section 7).

Figure 2 shows the flow of methodology, which consists of the pipeline to address our research questions.

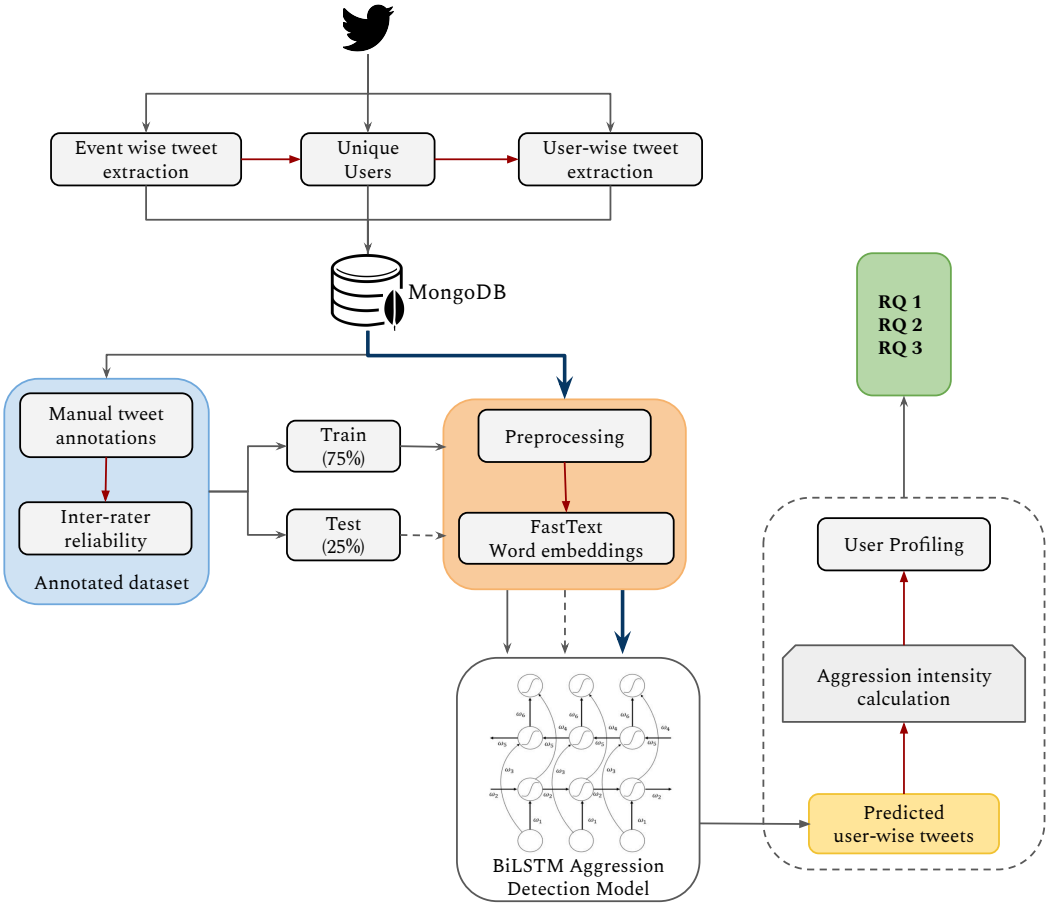



Figure 2. Flow of the proposed methodology to answer various research questions. The method includes various stages: data generation and annotation, training aggression detection models, user aggression intensity calculation, and user profiling.

# 5 Aggression Detection

## 5.1 Data Pre-processing

All data are preprocessed before being used in the model. We removed punctuation marks, numbers, and URLs as they have no significance. Additionally, all letters in English texts are converted to lowercase. We removed all English stop words, white spaces, and new line characters. A common practice is to mention users, even multiples, in tweets. These names are then used to identify possibly vulnerable users but have little or no significance in identifying aggression. Therefore, we removed them for the aggression detection model. Then we performed word lemmatization, which converts inflected to their original form and helps to preserve redundancy.

## 5.2 LSTM and Bidirectional LSTM

As we discussed in the section 1, aggression depends on the overall context of the sentence rather than specific words. Therefore, we used the architecture of LSTM (Long Short-Term Memory) to understand the context of sentences with long-term dependencies [DFW+20]. Figure 3 shows the architecture of a basic LSTM cell. LSTM mimics human brain activity for previously trained data. The memory of LSTM helps to retain the essential parts of the sentence and rejects the insignificant parts through the first layer of LSTM which is the forget gate ($f_t$). It passes through a sigmoid function ($\sigma(.)$) where the output is 0 for forget and 1 for remember (Eq. 1).

$$f_t = \sigma(W_f.[h_{t-1}, X_t] + b_f) \tag{1}$$

The second layer is the input layer ($i_t$), which retrains the remembered data (Eq. 2). The output of the forget gate is multiplied by the cell state of the previous LSTM ($C_{t-1}$). This result is added to the product of the output of the input gate and the result of the $\tanh$ function of the previous hidden state ($h_{t-1}$) becomes (Eq. 4) the state of the cell at timestamp t ($C_t$). After passing the cell state ($C_t$) through the $\tanh$ function, the result is multiplied with the output gate ($O_t$) to become the hidden state at timestamp t ($h_t$). Therefore, the last layer $C_t^r$ contains the combination of the previous and current cell states, which is the memory representation of the timestamp (0 to t) is forwarded to the next LSTM state.

$$i_t = \sigma(W_i.[h_{t-1}, X_t] + b_i) \tag{2}$$
$$C_t^r = \tanh\left(W_c.[h_{t-1}, X_t] + b_c\right) \tag{3}$$
$$C_t = f_t * C_{t-1} + i_t + C_t^r \tag{4}$$
$$o_t = \sigma(W_o.[h_{t-1}, X_t] + b_o) \tag{5}$$
$$h_t = \tanh\left(C_t\right) * o_t \tag{6}$$

Where $h_{t-1}$ and $x_t$ are the outputs of the previous and inputs of the current LSTM unit, and $W_x$, $b_x$ are the weights and biases of the corresponding x layer state.

LSTM remembered important information and passed from state 0 to state 1 to state 2 and further to state n, i.e. LSTM is forward oriented. Conventional LSTM is used in the forward direction method. LSTM can also work in the backward direction, which remembers important information and goes from state n to state n-1 to state n-2 and further to state 0. With a combination of both forward and backward LSTM methods, we constructed bidirectional LSTM (BiLSTM). BiLSTM learns the exact context of a sentence. To detect aggression, BiLSTM learns the context of a word sequence twice, once going forward and once going backward. This architecture introduced several cell states and learned important features that strengthened our model.

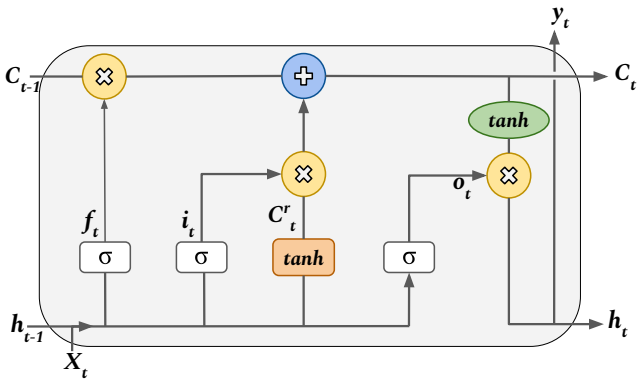



Figure 3. The representation of LSTM cell.

## 5.3 Experimental Setup

For this experiment, we used two popular word embedding techniques for text representation: Glove [PSM14], and FastText [JGB+16] in both classification models (section 5.2). The Glove and FastText word embeddings are adequate representations of the text [KSDR21b]. The Twitter corpus was utilized to obtain pre-trained Glove embeddings. It has 1.2 million words in its vocabulary and 27 billion tokens. We used 300-dimensional Glove and FastText word embedding for word representation. For regularization, we used a Spatial dropout [TGJ+15] of 0.2 after the embedding layer for LSTM and BiLSTM (in section 5.2). For both LSTM and BiLSTM models, we used 200 hidden units with 0.2 dropout and recurrent dropout rates, and the maximum sequence length is 64. In each model, the classification layer contains the fully connected layer with a non-linear sigmoid activation function. The output size of the model is the one that gives the probability of an aggressive class. A tweet is aggressive if the aggressive class probability is more significant than 0.5.

The models are trained on 32 batches of data using an Adam optimizer [KB14] with a learning rate of 1e-5, and a binary cross-entropy loss function [C$^+$18] is used to calculate the loss between predicted and actual values. We trained our model on the proposed aggressive dataset for 100 epochs using train and validation data with a 10% random split from the shuffled dataset to preserve the best model.

## 5.4 Evaluation

We conducted a separate experiment with different feature combinations using the proposed LSTM and BiLSTM classification models (in Sections 5.2 and 5.3). In this experiment, we used emotional features [KKR$^+$22] with specific word embeddings to detect aggression effectively. We have also used the word embedding of the embedding layer (Keras) [Ket17] along with Glove and FastText. Model performance is evaluated using five metrics precision, accuracy, recall, weighted F1 score and AUC (area under the ROC curve) [cite].

The classification performance of LSTM and BiLSTM models which is trained using word embeddings and emotional features is presented in Table 2. The proposed BiLSTM model with FastText word embedding reported the highest evaluation metrics compared to all other models. All BiLSTMs have higher accuracy than all other LSTM models. This observation highlights the effectiveness of the BiLSTM model over the LSTM model for aggression detection. Nevertheless, the performance of both the BiLSTM and LSTM models is better for FastText word embedding than all other word embeddings and combinations of emotional features. These observed results extrapolate that the proposed BiLSTM with FastText Word Embbeidng is effective for tweet aggression detection. Therefore tweet representation for aggression detection using FastText word embeddings is more contextually effective. Figure 4 shows the epoch-wise loss and accuracy of training and validation of the BiLSTM with the FastText model, which is neither underfitted nor overfitted.

| Models | Accuracy | Precision | Recall | weighted-F1-score | AUC |
|---|---|---|---|---|---|
| Embedding layer (Keras) - LSTM | 0.7552 | 0.7046 | 0.7351 | 0.7557 | 0.8084 |
| Embedding layer (Keras) - BiLSTM | 0.7615 | 0.6891 | 0.7556 | 0.7629 | 0.8106 |
| Emotions + Embedding layer (Keras) - BiLSTM | 0.7760 | 0.7200 | 0.6625 | 0.7739 | - |
| Glove - LSTM | 0.7983 | 0.7631 | 0.8055 | 0.7986 | 0.8705 |
| Glove - BiLSTM | 0.8004 | 0.7688 | 0.8009 | 0.8006 | 0.8782 |
| Emotions+Glove - BiLSTM | 0.7983 | 0.7752 | 0.7824 | 0.7983 | 0.8787 |
| FastText- LSTM | 0.8004 | 0.7641 | 0.8101 | 0.80 50 | 0.8745 |
| **FastText- BiLSTM** | **0.8151** | **0.7758** | **0.8333** | **0.8154** | **0.8818** |

Table 2. Comparison of the performance of the proposed model and the traditional popular model.

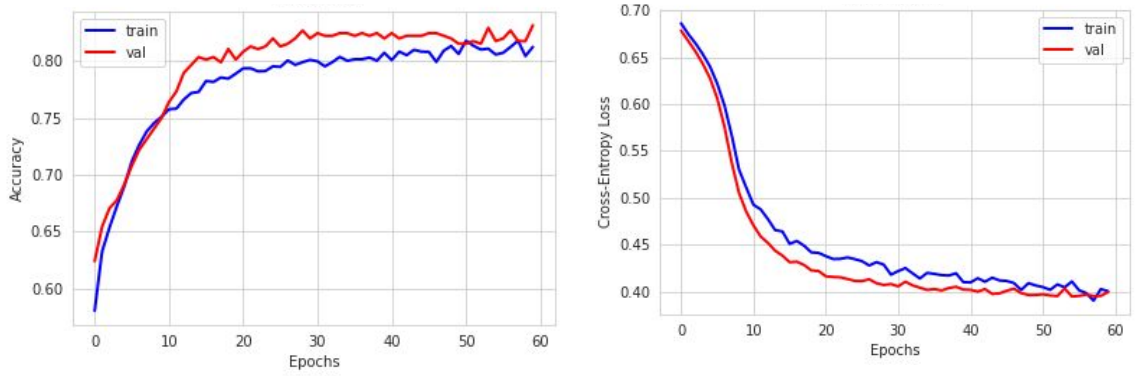



Figure 4. a) Epoch-wise accuracy during training and validation of the BiLSTM model with FastText embedding. b) Epoch-wise loss during training and validation of the BiLSTM model with FastText embedding.

# 6 Aggression Intensity Calculation

We used data disaggregated by time period (i.e., week / day / hour) to measure aggression intensity with respective users. A user's intensity of aggression represents the total aggressive behavioral activity over a specific period of time. A higher Aggressive Intensity score of a user indicates that the user has potentially aggressive behavior compared to a lower Aggressive Intensity score of the user, which indicates non-aggressive behavior of the user. Aggression intensity scores range from 0 to 1. 1 represents the user with the highest aggression activity behavior, and 0 represents the user with non-aggressive behavior.
For the calculation of aggression intensity, we used user's labeled (aggressive/non-aggressive) posts. The aggression Intensity of user i in period l ($AI_i^l$) (in Eq 7) is a multiplication of user aggressiveness aggregated score and the normalization score of the total posts. The user aggressiveness aggregated score is a fraction of the total aggressive posts of user i in period l ($AG_i^l$), and the total posts of user i in period l ($X_i^l$). The normalization score of the total posts is the fraction of the difference between $X_i^l$ and the minimum number of posts in period l from all users ($min^l$) and the difference between $min^l$ and the maximum number of posts in period l from all users ($max^l$).

$$AI_i^l = \frac{AG_i^l}{X_i^l} * \frac{X_i^l - min^l}{max^l - min^l} \tag{7}$$

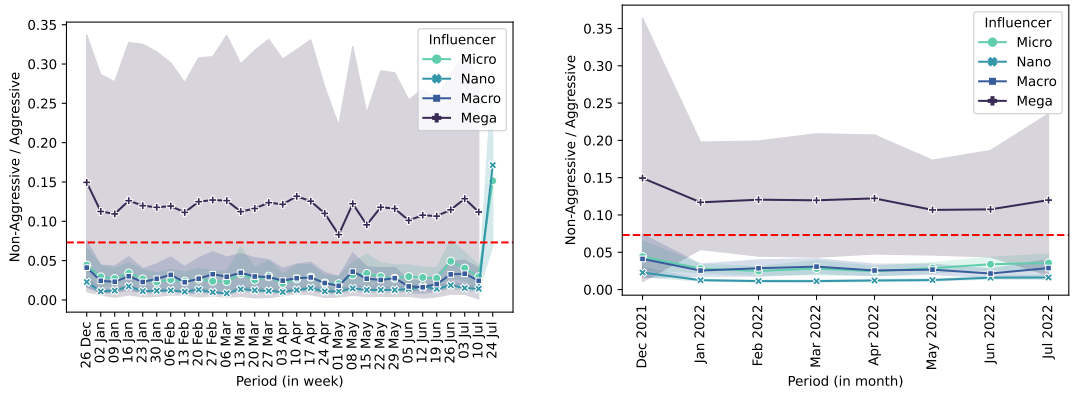
where

$$\begin{cases} 1 \leq min^l < max^l \\ min^l < max^l \leq n \end{cases} \tag{8}$$

In this experiment, we calculated a week-wise aggression intensity score, which represents the user's overall aggressive activity in a given week.
**Influencer's aggression behavior analysis**: We created user buckets based on the number of followers they had. Buckets are nothing but cores of influencers. Following are the four influential core of users based on the number of their followers.

1. **Nano**: User has less than 10,000 followers.
2. **Micro**: User has greater than 10,000 and less than 100,000 followers.
3. **Macro**: User has greater than 100,000 and less than 1 million followers.
4. **Mega**: User has greater than 1 million followers.

As shown in Figure 5, aggression intensity scores of the mega influence core are higher than all other cores. The users in the mega influence core are the most influential.




(a) Week-wise aggressive behavior of Influencer cores

(b) Month-wise aggressive behavior of Influencer cores

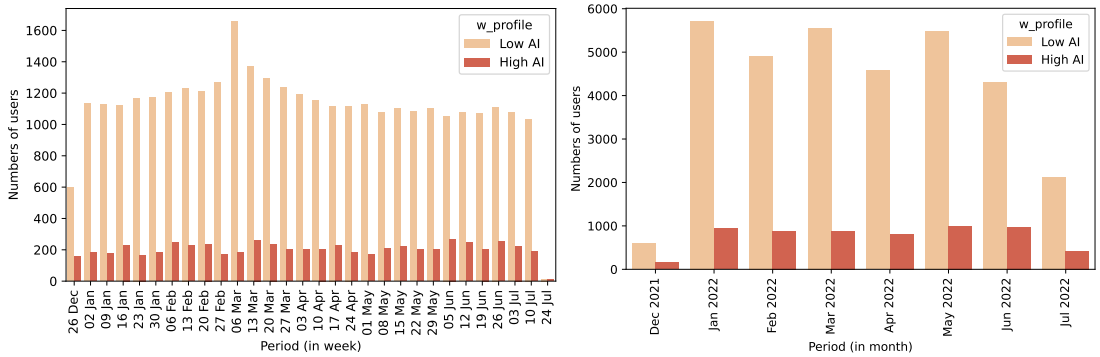
Figure 5. Aggressive behavior of Influencer cores in week and month respectively. The Mega Influencer Core has consistently high Aggression Intensity scores over both weekly and monthly periods.

This observation shows that the most influential users portray more aggressive behavior to the greater number of followers on Twitter. So there is a chance that followers may start to behave aggressively after going through their aggressive feeds. This analysis encourages us to understand "whether user's aggressive feeds (following user's posts) make them aggressive?", which is analyzed in section 8.

## 6.1 Threshold Selection

We calculated each user's aggression intensity score from 0 to 1. An intensity score of 1 indicates that the user's behavior is aggressive, and on the other hand, 0 indicates that the user's behavior is non-aggressive. We need a threshold in the intensity of aggression to make two partitions, one into aggressive behavior users and the other into non-aggressive behavior users. If the intensity score is greater than the threshold, the user's behavior is aggressive, and if it is less, the user's behavior is non-aggressive.

To select a threshold value, we used a K-means clustering algorithm ([JMF99], [Mac67]) on the aggression intensity score. K-means requires the value of k, which represents the number of clusters. It is not correct to directly take the k value of 2, high and low aggressiveness clusters, because there is also the possibility of having more than two clusters, for example, if k is 3, low, medium and high clusters are possible. So to select the k value, we used the K-means elbow method [Har94]. For our scalar aggression intensity, we obtained k = 2 using the k-means elbow method. We applied K-means to all users' aggression intensity scores across weeks. Figure 6 shows low and high aggression intensity users by month and week.




(a) Week-wise user aggressiveness. (b) Month-wise user aggressiveness.

Figure 6. Number of low and high aggressive intensity users per week and month, respectively.

# 7 User Profiling

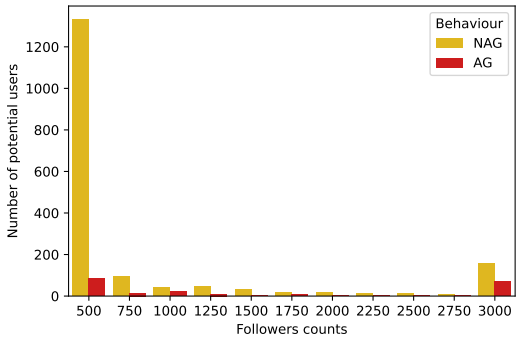
We used the low and high clusters (Section 6.1) to label each user's week, which represents the user's vector over the entire week period. A vector is a sequence of low and high encoding. This method enables to obtain of the perspective of the users over a period of time.

An aggressive user profile is identified if the user's vector has a high encoding of 75% and above. A non-aggressive user profile is identified if the user vector has less encoding than 75% and above.

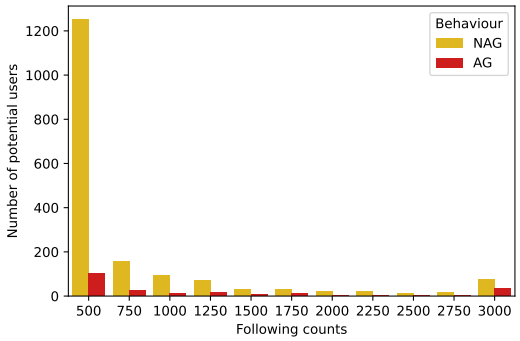
For our study, we mainly considered only two user profiles aggressive and non-aggressive. We would like to mention that different types of profiles are also possible such that aggression intensity scores can range from low to high and high to low in several times, but we did not consider these cases.

We analyzed the number of followers and following of the aggressive and non-aggressive users, and also for target users mentioned by aggressive and non-aggressive users (in Figure 7). For this analysis, we categorized the number of followers and following in an interval of 250. Figure 7a presents the number of followers of aggressive and non-aggressive users, in which some potentially aggressive users are influencers. Similarly, Figure 7b presents the number of followings that are nothing but friends of aggressive and non-aggressive users, in which some potentially aggressive users have fewer numbers of friends. Several targeted users are popular users (i.e., celebrities, politicians, etc.) because mostly those users have a higher number of followers and a lower number of followings reported in Figures 7c and 7d.

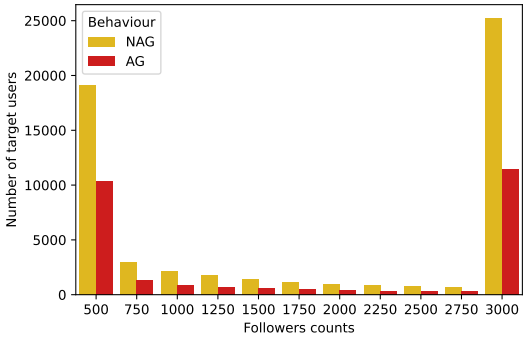
We identified top target users in aggressive and non-aggressive domains, which is highly mentioned in the respective tweets (in Figure 8). These top target users are the most vulnerable, particularly mentioned in aggressive tweets (in Figures 11a and 11b).



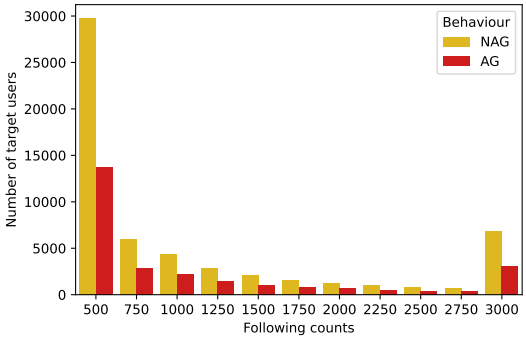

(a) Analysis of aggressive and non-ggressive user profile followers.




(b) Analysis of aggressive and non-aggressive user profile followings.

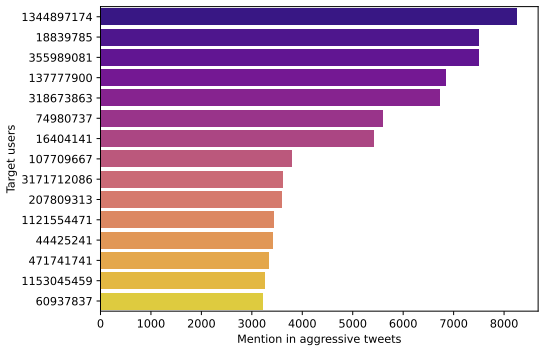



(c) Analysis of the followers of target users, which is mentioned by aggressive and non-aggressive user profiles.

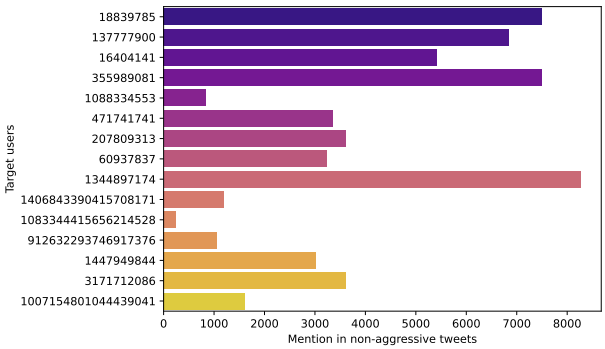



(d) Analysis of the following of target users, which is mentioned by aggressive and non-aggressive user profiles.

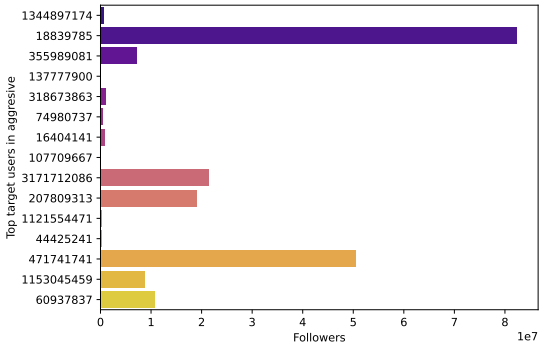
Figure 7. Followers and following analysis of potentially aggressive and non-aggressive users, and their target users.



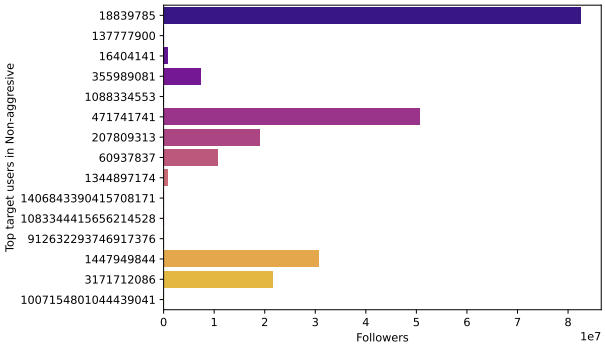

(a) Top target users mentioned in the aggressive tweets.




(b) Top target users mentioned in the non-aggressive tweets.




(c) Followers of top target users are mentioned in the aggressive tweets.




(d) Followers of top target users are mentioned in the non-aggressive tweets.

Figure 8. Analysis of top target users is mentioned in the aggressive and non-aggressive tweets, respectively.

Further, we analyzed some top target users are popular, but not all are reported in Figures 11c and 11d.

# 8 RQ1: Do my feeds makes me aggressive?

## 8.1 Aggressive intensity of users and their feeds are correlated

In this section, we investigate the user aggression intensity in the period of the day and the aggression intensity of the respective user's feeds which is their following user's posts. For this investigation, we considered the period granular level as day instead of week or month, in an hour period data is insignificant. For the day period (l=day), we calculated the aggression intensity of the user as discussed in section 6. Further, calculated the feed intensity of the user for that we considered the feeds were the following user's posts, so we figured the aggression intensity of all following users using equation 7 in the current and previous daytime period. Then we aggregated the aggression intensity of all the following users of the respective user. The following equation is used to calculate the feed aggression intensity considering both the current and previous daytime periods for respective user.

$$FU_i^l = \frac{\sum_{j=1}^{n} FAI_{ij}^l + \sum_{j=1}^{n} FAI_{ij}^{l-1}}{n} \tag{9}$$

Where $FU_i^l$ is the feed intensity of user (i) on the period (l=day) and $FAI_{ij}^l$ is the aggression intensity of the following user (j) on the period (l=current day and l-1 = previous day of user i).

The aggressive intensity of users has a positive correlation with the aggressive intensity of feeds (current and previous day) at 0.48 ($P < 0.05; P = 2.16e - 69$), which indicates that aggressive feeds indeed make users aggressive. Similarly, we investigate correlation of user intensity and their feed intensity ($P < 0.05$) for last 2 day (correlation: 0.1722), for last 3 day (correlation: 0.1712) and for last 4 day (correlation: 0.17038) using eq. 9. This analysis shows that the correlation gradually decreases with the growth of previous feeds, which implies that the feed's effect is mostly short-term for aggression. This experiment is conducted on 2 million of users and their 16 million following users.

## 8.2 Aggressive post has a higher feed intensity

To analyze the aggression intensity of individual post feeds, we considered feed posts in the last 24 hours since the post was created. So, we calculated the aggressive feed intensity using posts of feeds instead of users (Eq. 10), which is the likelihood of feed aggressiveness for a respective post. For all aggressive and non-aggressive posts, we assessed feed intensity.

$$FT_i = \frac{AGF_i}{n_i} \tag{10}$$

Where FTi is the aggressive feed intensity of post (i), AGFi is the total aggressive feeds of post (i) and ni is the total feeds of post (i). For hypothesis testing used two tailed student t-test to Further, we analyzed the aggressive post has higher aggressive feed intensity than the non-aggressive post using two-tailed student's t-test, which is 7.4070 ($P < 0.05; P = 1.37e - 13$). This investigation revealed that most aggressive posts are posted when the user's feed has been aggressive in the past few hours.

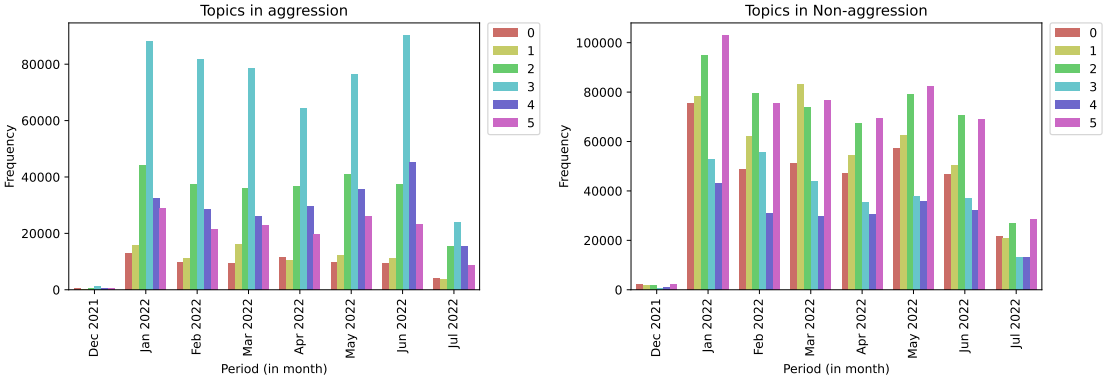
# 9 RQ2: Do event feeds makes event specific aggressive?

We obeyed every topic has positive as well negative activity of users in our case, those are aggressive and non-aggressive. On another side some topics are aggressive, i.e., users are most aggressive in them, for example, the Hindu-Muslim controversy in India. So, we discover the topics of user posts using Linear Discriminant Analysis (LDA) [BG98] which has aggressive and non-aggressive activity. Figure 9 shows that most topics discussed aggressively are different from non-aggressive topics. Specifically, topic 3 has high aggressive activity, but it has not that much non-aggressive activity. Topic 3 related to the Hindu-Muslim people in India and Pakistan, which we analyzed had highly aggressive user activity on Twitter social media from 1st Jan 2022 to 15th July 2022. So, the insight of this analysis is most of the time, the aggressive topic makes the user aggressive.

For event analysis, we considered the event as the hashtag of the tweet. Figure 10 presents the event analysis of users' aggressive and non-aggressive activity. The most popular events in aggressive and non-aggressive are different. For a detailed analysis of events, we considered the three events discussed in section 3.2. To investigate whether event feeds makes event specific aggressive, we considered posts of only those three events. Then we calculated the post-wise feed aggressive intensity of the respective event (like section 8.2). In this experiment, AGFi represents the total aggressive feeds of the user (i) for an individual event (E), and ni is the total feeds of the user (i). The feed intensity of the post represents the probability of aggressiveness in the respective event. Finally, we analyzed aggressive post has higher aggressive feed intensity of individual post events than non-aggressive posts. We tested the significance of the result using a two-tailed student's t-test, which is 7.5618 ($P < 0.05; P = 4.237e - 14$). To further understand whether the order of aggressiveness of events in users' posts is similar to their feed's post. To analyze, we calculated the aggressive intensity score of each event using user's posts and their feeds, respectively(Eq 11).

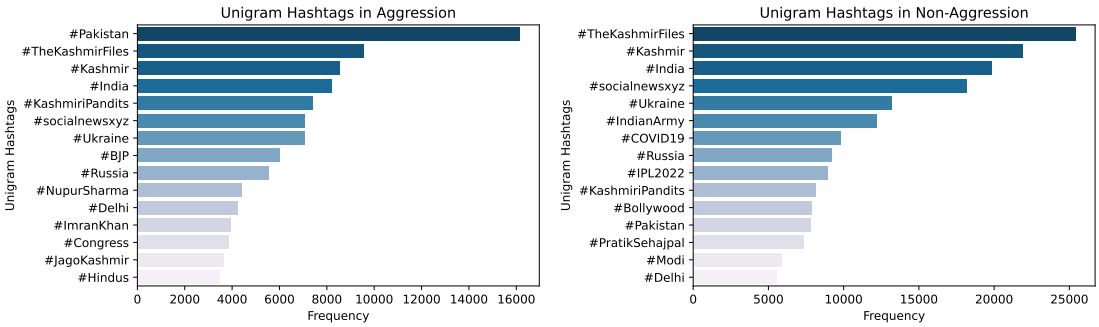
$$AIE_i = \frac{AGE_i}{XE_i} * \frac{XE_i - minE}{maxE - minE} \tag{11}$$

Where AIEi is the aggressive intensity of an event (i), AGEi is the total aggressive posts of an event (i), XEi is the total posts of event (i), minE and maxE are the minimum




(a) Topics in the aggressive activity of users. (b) Topics in the non-aggressive activity of users.

Figure 9. Month-wise Analysis of Topics in the aggressive and non-aggressive activity of users.




(a) Events used in the aggressive activity of users.

(b) Events used in the non-aggressive activity of users.

Figure 10. Analysis of Events used in the aggressive and non-aggressive activity of users.

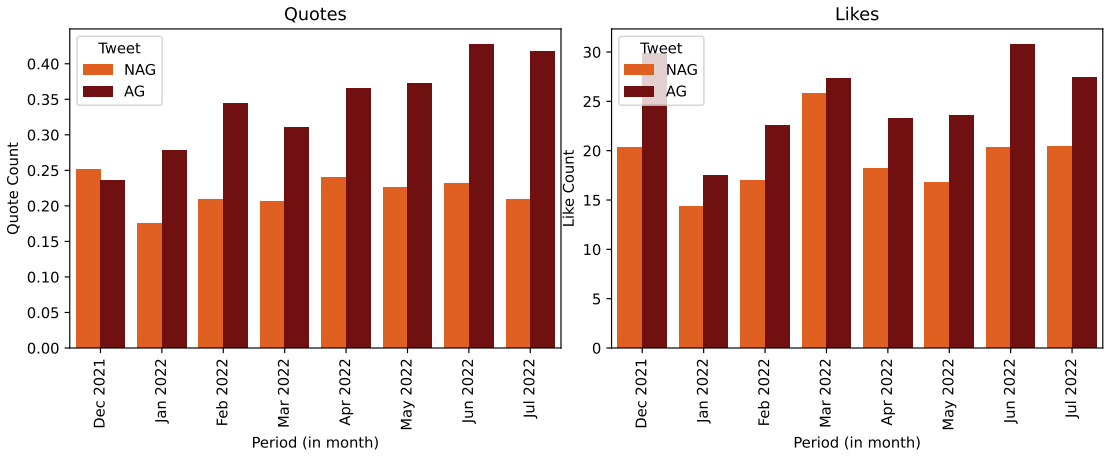
and maximum posts among all events, respectively. After event intensity calculation, the order of events in the user's posts is $Event_3 > Event_1 > Event_2$, and the order of events in their feeds is $Event_1 > Event_2 > Event_3$. The aggressive order of event 1 and event 2 is similar in user's activity and their feeds.

# 10 RQ3: Do user engagement is more toward aggressive posts?

In Twitter, users respond to posts in different ways, such as to quote, like, reply, and retweet or re-post on the post, which we consider as user engagement on a particular post.

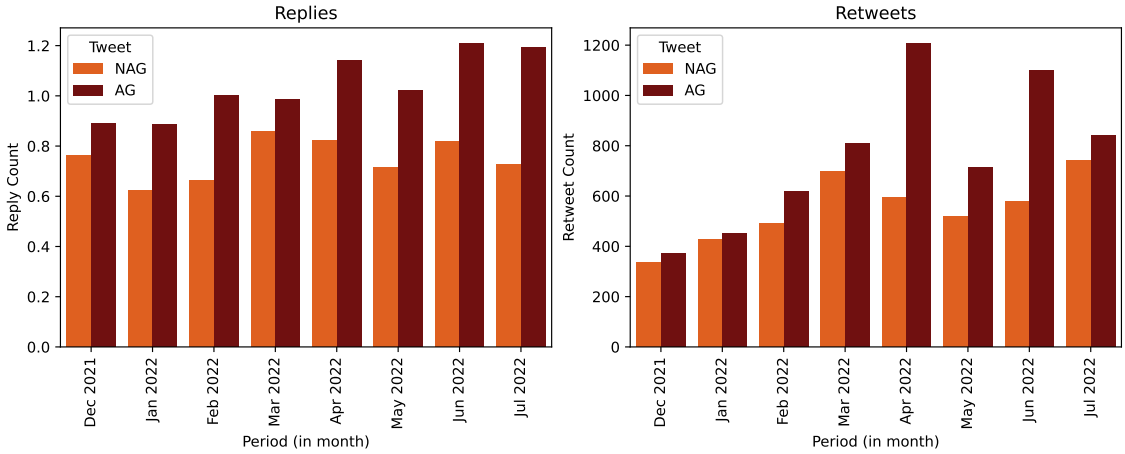
Figure 11 shows that aggressive posts have consistently higher quotes, likes, replies, and retweets than non-aggressive posts. We encounter this research question after observation of the user's engagement on aggressive and non-aggressive posts.

To analyze this research question, we calculated each post's engagement score, which is the average of the total quotes, likes, replies, and retweets of a particular post. Further, we analyzed using a two-tailed Student's t-test that aggressive posts had a higher engagement score than non-aggressive posts, which is 46.98 ($P < 0.05; P = 0.0$. This investigation revealed that most users engage more with aggressive posts.




(a) The number of users quotes for aggressive and non-aggressive posts.

(b) The number of users like reaction for aggressive and non-aggressive posts.




(c) The number of users replies for aggressive and non-aggressive posts.

(d) The number of users retweets for aggressive and non-aggressive posts.

Figure 11. Analysis of users engagement for aggressive and non-aggressive posts, respectively.

# 11 Discussion

Although the technological capabilities increased a lot during the previous years and the popularity of social media grew, and it became easier to exchange information with each other,at the same time the expression of aggression level also saw an uptick. Many researchers investigated this problem, and they made different kinds of aggression detection models. However, it can be challenging to identify aggression on social media, and the work that needs to be done is a bit time-consuming. The researchers face many challenges while doing it. Not many of them tried to explore this issue from the behavioral perspective.

As we analyzed in this paper, people with more followers have more outstanding aggression intensity scores. So as it is described, famous people could have a disproportionate impact on spreading the news or be the main reason causing some action from their followers [BHMW11].Our study reinforces this idea that my feed could make me aggressive.

# 12 Conclusion and and Future work

In this paper, we have studied the aggressive behavior of social network users based on their feeds and events. After building an aggression detection model, we were able to investigate users' aggressiveness properly. As we saw, people engage more in aggressive posts, and their behavior depends on the specific events and feed. Finally, the goal of this research was fulfilled. The work done to identify social media users' behavior and interactions could be vital for society. Detecting aggressive users in an early stage and predicting their behavior may prevent some of the problems that currently exist. Our analysis also has some limitations. Our research included analyzing only the Twitter text, and we haven't used any images, emojis, etc. In the future, this study can be extended, and we can work on a multimodel aggression detection model.

# Bibliography


[AB18] Craig A. Anderson and Brad J. Bushman. Media violence and the general aggression model. *Journal of Social Issues*, 2018.

[Ame15] Jacob Amedie. The impact of social media on society. 2015.

[AWCM20] Akshita Aggarwal, Anshul Wadhawan, Anshima Chaudhary, and Kavita Maurya. "did you really mean what you said?" : Sarcasm detection in hindi-english code-mixed data using bilingual word embeddings. *ArXiv*, abs/2010.00310, 2020.

[BDBD20] Arup Baruah, K Amar Das, Ferdous Ahmed Barbhuiya, and Kuntal Dey. Aggression identification in english, hindi and bangla text using bert, roberta and svm. In *Workshop on Trolling, Aggression and Cyberbullying*, 2020.

[BG98] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

[BHMW11] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.

[BSK+20] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*, 2020.

[C+18] François Chollet et al. Keras: The python deep learning library. *Astrophysics source code library*, pages ascl–1806, 2018.

[CKB+17] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. *Proceedings of the 2017 ACM on Web Science Conference*, 2017.

[DAB11] C. Nathan DeWall, Craig A. Anderson, and Brad J. Bushman. The general aggression model: Theoretical extensions to violence. *Psychology of Violence*, 1:245–258, 2011.

[DFW+20] Yongfeng Dong, Yu Fu, Liqin Wang, Yunliang Chen, Yao Dong, and Jianxin Li. A sentiment analysis method of capsule network based on bilstm. *IEEE Access*, 8:37014–37020, 2020.

[DS21] Maibam Debina and Navanath Saharia. Delab@iiitsm at icon-2021 shared task: Identification of aggression and biasness using decision tree. In *ICON*, 2021.

[EK19a] Levent Eraslan and Ahmet Kukuoglu. Social relations in virtual world and social media aggression. *World Journal on Educational Technology: Current Issues*, 11(2):1–11, 2019.

[EK19b] Levent Eraslan and Ahmet Kukuoğlu. Social relations in virtual world and social media aggression. *World Journal on Educational Technology: Current Issues*, 2019.

[FD12] Christopher J. Ferguson and Dominic Dyck. Paradigm change in aggression research: The time has come to retire the general aggression model. *Aggression and Violent Behavior*, 17:220–228, 2012.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[GIPS22] Vincenzo Gattulli, Donato Impedovo, Giuseppe Pirlo, and Lucia Sarcinella. Cyber aggression and cyberbullying identification on social networks. In *ICPRAM*, 2022.

[GP20] N.A. Gunathillake and Hasuli Kumarika Perera. Association among depression, social anxiety, and aggression caused by cyberbullying on facebook among sri lankan adults. *SLIIT Journal of Humanities and Sciences*, 2020.

[Har94] André Hardy. An examination of procedures for determining the number of clusters in a data set. In *New approaches in classification and data analysis*, pages 178–185. Springer, 1994.

[HCK20] Herodotos Herodotou, Despoina Chatzakou, and Nicolas Kourtellis. A streaming machine learning framework for online aggression detection on twitter. *2020 IEEE International Conference on Big Data (Big Data)*, pages 5056–5067, 2020.

[HCR+20] Fattah Hanurawanb, Tutut Chusniyahc, Hetti Rahmawatid, Ifdil Ifdile, and Mario Pratamaf. Cyber aggression of students: The role and intensity of the use of social media and cyber wellness. 2020.

[HG14] Jack Hollingdale and Tobias Greitemeyer. The effect of online violent video games on levels of aggression. *PLoS ONE*, 9, 2014.

[HWGD+21] McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H. A. Schwartz, David H. Epstein, Lorenzo Leggio, and Brenda L Curtis. Bots and misinformation spread on social media: Implications for covid-19. *Journal of Medical Internet Research*, 23, 2021.

[HZZL21] Jinyu Huang, Zhaohao Zhong, Haoyuan Zhang, and Liping Li. Cyberbullying in social media and online games among chinese college students and its associated factors. *International Journal of Environmental Research and Public Health*, 18, 2021.

[JGB+16] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[JMF99] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Ket17] Nikhil Ketkar. Introduction to keras. In *Deep learning with Python*, pages 97–111. Springer, 2017.

[KKR+22] Umair Khan, Salabat Khan, Atif Rizwan, Ghada Atteia, Mona M Jamjoom, and Nagwan Abdel Samee. Aggression detection in social media from textual data using deep learning models. *Applied Sciences*, 12(10):5083, 2022.

[KRBM18] Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*, 2018.

[KSDR21a] Kirti Kumari, Jyoti Prakash Singh, Yogesh K. Dwivedi, and Nripendra P. Rana. Bilingual cyber-aggression detection on social media using lstm autoencoder. *Soft Computing*, 25:8999 – 9012, 2021.

[KSDR21b] Kirti Kumari, Jyoti Prakash Singh, Yogesh Kumar Dwivedi, and Nripendra Pratap Rana. Bilingual cyber-aggression detection on social media using lstm autoencoder. *Soft Computing*, 25(14):8999–9012, 2021.

[KT20] A. Kalaivani and Durairaj Thenmozhi. Sarcasm identification and detection in conversion context using bert. In *Fig-Lang@ACL*, 2020.

[Mac67] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

[MD18] Sreekanth Madisetty and Maunendra Sankar Desarkar. Aggression detection in social media using deep neural networks. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 120–127, 2018.

[MHM22] Alexandra Maftei, Andrei Corneliu Holman, and Ioan-Alex Merlici. Using fake news as means of cyber-bullying: The link with compulsive internet use and online moral disengagement. *Comput. Hum. Behav.*, 127:107032, 2022.

[NBL+19] Muhammad Zidny Naf'an, Alhamda Adisoka Bimantara, Afiatari Larasati, Ezar Mega Risondang, and Novanda Alim Setya Nugraha. Sentiment analysis of cyberbullying on instagram user comments. *Journal of Data Science and Its Applications*, 2019.

[NNF+19] Kristiawan Nugroho, Edy Noersasongko, Ahmad Zainul Fanani, Ruri Suko Basuki, et al. Improving random forest method to detect hatespeech and offensive word. In *2019 International Conference on Information and Communications Technology (ICOIACT)*, pages 514–518. IEEE, 2019.

[OVM21] Xavier Oriol, Jorge J. Varela, and Rafael Miranda. Gratitude as a protective factor for cyberbullying victims: Conditional effects on school and life satisfaction. *International Journal of Environmental Research and Public Health*, 18, 2021.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[RHY+21] Md Saifullah Razali, Alfian Abdul Halin, Lei Ye, Shyamala Doraisamy, and Noris Mohd Norowi. Sarcasm detection using deep learning with contextual features. *IEEE Access*, 9:68609–68618, 2021.

[SCA20] Gabriel Araújo De Souza and Márjory Da Costa-Abreu. Automatic offensive language detection from twitter data using machine learning and

feature selection of metadata. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2020.

[SH21] Omar Sharif and Mohammed Moshiul Hoque. Identification and classification of textual aggression in social media: Resource creation and evaluation. In *CONSTRAINT@AAAI*, 2021.

[SSR21] Francesc Sidera, Elisabet Serrat, and Carles Rostan. Effects of cybervictimization on the mental health of primary school students. *Frontiers in Public Health*, 9, 2021.

[TGJ+15] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[WHL+22] Yichen Wang, Richard O. Han, Tamara Lehman, Qin Lv, and Shivakant Mishra. Do twitter users change their behavior after exposure to misinformation? an in-depth analysis. *Social Network Analysis and Mining*, 12, 2022.

[ZHB20] Xi Zhang, Ziqiang Han, and Zhanlong Ba. Cyberbullying involvement and psychological distress among chinese adolescents: The moderating effects of family cohesion and school cohesion. *International Journal of Environmental Research and Public Health*, 17, 2020.

# Appendix

# I. Glossary

AG - Aggressive
NAG - Non-Aggressive
SNS - Social network sites
OAG- Overtly Aggressive
CAG - Covertly Aggressive
RQ - Research Question
GAM - General aggression model
BERT - Bidirectional Encoder Representations from Transformers
LSTM - Long short-term memory
RNN - Recurrent Neural Network
BiLSTM - bidirectional Long short-term memory
CNN - Convolutional neural network
URL - Uniform Resource Locator
API - Application programming interface
TP - True Positive
FP - False Positive
TN - True Negative
FN - False Negative
AUC- Area under the ROC Curve

# II. Licence

## Non-exclusive licence to reproduce thesis and make thesis public

I, **Ketevani Kvirikashvili**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   **Study of cyber-aggressive behavior on social media**,
   (title of thesis)

   supervised by Swapnil Mane, Rajesh Sharma and Suman Kundu.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ketevani Kvirikashvili
***05/01/2023***