You are what your feeds makes you: A study of user aggressive behaviour on Twitter

Swapnil Mane [1], Suman Kundu [2], and Rajesh Sharma [2]

[1]Indian Institute of Technology Jodhpur
[2]Affiliation not available

April 28, 2023

You are what your feeds make you: A study of user aggressive behavior on Twitter

Swapnil Mane[a,*], Suman Kundu[a,*] and Rajesh Sharma[b]

[a]Department of Computer Science and Engineering, Indian Institute of Technology, Jodhpur, Rajasthan, India
[b]Institute of Computer Science, University of Tartu, Estonia

ARTICLE INFO

Keywords:
Cyberbullying
Aggression
User behavior
Social Media Analysis

ABSTRACT

The widespread use of aggressive language on Twitter raises concerns about potential negative influences on user behavior. Despite previous research exploring aggression and negativity on the platform, the relationship between consuming aggressive content and users' aggressive behavior remains underexplored. This study investigates whether exposure to aggressive content on Twitter can lead users to behave more aggressively. Our methodological approach contains four stages: data collection and annotation, aggressive post detection, user aggression intensity metric, and user profiling. We proposed the English Twitter Aggression dataset (TAG) with substantial inter-annotator agreement (Krippendorff's alpha=0.78). Subsequently, we benchmark the aggression detection performance on TAG dataset (macro F1=0.92) by fine-tuning a pre-trained RoBERTa-large. We quantified user aggression with a proposed "user aggression intensity" metric based on their overall aggressive activity. Our analysis of 14M posts from 63K users revealed that aggressive Twitter feeds can influence users to behave more aggressively online. Furthermore, the study found that users tend to support and encourage aggressive content on social media, which can contribute to the proliferation of aggressive behavior.

WARNING: This paper includes examples of potentially harmful abusive language typically observed on Twitter.

1. Introduction

As society becomes increasingly interconnected through technology, social media platforms like Twitter (now re-branded as X)[1] have become a fundamental part of modern communication. According to Kwak et al. (2010), and Wu et al. (2011), platforms like Twitter provide valuable insights into user behavior and information dissemination. However, this platform has also led to the emergence of aggressive behavior among users. This can have serious consequences, including cyberbullying, hate speech, and even physical and mental harm or suicide (Hinduja and Patchin, 2010; Liu et al., 2021). Research has also shown that certain types of content, such as political content, can contribute to increased aggression and polarization among users (Bakshy et al., 2015; Bruns et al., 2017). In light of these challenges, it is crucial to understand the nature and causes of aggressive behavior on social media platforms.

Despite an upsurge in research on aggressive behavior on Twitter, the relationship between consuming aggressive content and user behavior remains unexplored. Previous research has primarily focused on aggression detection on social media platforms using various machine learning (Datta et al., 2020; Arroyo-Fernández et al., 2018; Gutiérrez-Esparza et al., 2019), deep learning models (Aroyehun and Gelbukh, 2018; Srivastava and Khurana, 2019; Kumari et al., 2021), as well as identifying aggression in multilingual (Sharif and Hoque, 2022; Kumari et al., 2021; Torregrosa et al., 2022) and multi-modal posts (Khandelwal and Kumar, 2020; Kumari and Singh, 2022) on social media. Apart from aggression detection, some sociological studies of aggressive behavior on social media exist (Vladimirou et al., 2021; Pascual-Ferrá et al., 2021). However, there are limited large-scale empirical works to understand various aspects of aggressive behavior.

This study aims to investigate whether exposure to aggressive content or event-specific content can increase an individual's aggressiveness. Firstly, we collected a large amount of data from January 1, 2022, to July 15, 2022. The crawling was initiated by using hashtags from three specific seed events. We collected all posts (irrespective of the seed events) from users who wrote at least once on the seed events during the same time frame. Further, we collected all posts from their followers. As a result, the majority of the collected posts contained information on various events that occurred during that period. Overall, we collected 14M English and corresponding 63K users. We proposed the English Twitter Aggression dataset (TAG) through human manual annotation with 0.78 substantial agreement. Subsequently, we developed an aggression detection model by fine-tuning a pre-trained transformer and established benchmark performance on our TAG dataset. The method achieved an average macro F1 of 0.92. We then used this model to predict the rest of the data and measure users' aggressive behavior over a specified period using a proposed metric "user aggression intensity". This metric has been used to profile users as aggressive and non-aggressive. Based on these computational techniques,

*Corresponding author.
mane.1@iitj.ac.in (S. Mane); suman@iitj.ac.in (S. Kundu); rajesh.sharma@ut.ee (R. Sharma)
ORCID(s): 0000-0002-8234-4557 (S. Mane); 0000-0002-7856-4768 (S. Kundu); 0000-0003-3581-1332 (R. Sharma)

[1]Since we performed our study on the data collected during July, 2022, we will refer to the platform as Twitter.

we studied the following three Research Questions (RQ) related to aggressive behavior on Twitter social media.

- **RQ1:** Do aggressive feeds can make someone aggressive?
  Our study found that the content of a user's feed does affect their likelihood of engaging in aggressive behavior. This finding highlights the importance of understanding the impact of the content that users are exposed to on their likelihood of engaging in aggressive behavior.
- **RQ2:** Does exposure to event-aggressive feeds increase the user's event-specific aggressive behavior?
  We observed a correlation between event-related content and an escalation in aggressive behavior among Twitter users. This finding is significant as it shows that certain types of content can lead to increased aggression among users.
- **RQ3:** Do users engage more with aggressive posts?
  Our analysis revealed that user engagement towards aggressive content is high. This finding suggests that users were more likely to engage with, support, and encourage aggressive content.

These questions investigate potential relationships between aggressive content exposure and user aggression levels on Twitter, specifically examining the associations between general feeds, event-specific feeds, and user engagement with aggressive posts. Our findings can facilitate the development of interventions aimed at reducing aggressive behavior on social media platforms. The rest of the paper is organized as follows: Section 2 provides a literature review on aggression detection and behavior analysis in social media. In Section 3, we present our methodology for analyzing aggressive behavior on Twitter. The qualitative results of our analysis of the research questions are discussed in Section 4. Finally, Section 6 summarizes our conclusions.

2. Related Work

The literature review is organized into two groups of work in the following sections.

2.1. Aggression detection

Studies have shown that aggressive behavior is prevalent on Twitter (Kwak et al., 2010; Wu et al., 2011). Several studies have created annotated datasets of aggression in various languages, such as English (Kumar et al., 2018b,a; Bhattacharya et al., 2020; Kumar et al., 2020; Ali et al., 2023), Hindi (Kumar et al., 2018b,a; Bhattacharya et al., 2020; Kumar et al., 2020), Bengali (Bhattacharya et al., 2020; Kumar et al., 2020; Sharif and Hoque, 2022), Italian (Gattulli et al., 2022), Spanish (Torregrosa et al., 2022), Turkish (Balci and Salah, 2015), and Russian (Gordeev, 2016). These datasets were curated from social media platforms such as Facebook, Twitter, YouTube, and others. The political Hindi-English code-mixed Twitter aggression dataset was introduced by Rawat et al. (2023), and its performance was benchmarked through the fine-tuning of pre-trained transformers. Various machine learning models such as Random Forest, VIMs, OneR (Arroyo-Fernández et al., 2018; Gutiérrez-Esparza et al., 2019), bagging XGBoost, Gradient Boosting Machine models (Datta et al., 2020), XGBoost classifiers (Tawalbeh et al., 2020) and weighted ensemble technique (Sharif and Hoque, 2022) have been utilized for aggression detection. Different deep-learning models for aggression detection have also been used. These include CNN (Agbaje and Afolabi, 2022), LSTM (Agbaje and Afolabi, 2022; Aroyehun and Gelbukh, 2018; Kumari et al., 2021; Pareek et al., 2022; Ali et al., 2023), BiLSTM (Srivastava and Khurana, 2019), and their combinations. Different features such as two-dimensional TF-IDF vectors (Chen et al., 2020), embedding from Convolutional Capsule Layer (Srivastava and Khurana, 2019) and embedding from FastText (Pareek et al., 2022), and sentiment analysis(Agbaje and Afolabi, 2022) have also been employed for aggression detection. Multilayer perceptron classifiers with TF-IDF of unigram and bigram features are effective in identifying aggression (Sadiq et al., 2021). Moreover, LSTM with GRU (Ali et al., 2023), and deep learning models with emotional features and word embeddings have shown better performance (Khan et al., 2022). An ensemble of multiple fine-tuned BERT models based on bootstrap aggregating for aggression detection was proposed by Risch and Krestel (2020), which performed best on the dataset developed by Kumar et al. (2020). Similarly, a multitask deep neural network model using attention on top of the BERT model to identify aggression and misogyny was proposed by Samghabadi et al. (2020). Shrivastava et al. (2021) developed an aggression detection model based on GPT-2 and data balancing techniques using an ensemble approach. Multitask learning (MTL) with transformer-based models (RoBERTa) (Ramiandrisoa, 2022) were utilized for hate and aggression detection. Furthermore, Kumari and Singh (2022) addressed multimodal posts having symbolic images and text using pre-trained VGG-16 models and three-layered CNN, respectively, and combined the features to create hybrid features. These features are optimized using binary particle swarm optimization (BPSO) and binary firefly optimization (BFFO) algorithms. Khandelwal and Kumar (2020) proposed a multimodal deep learning architecture with linguistic and psychological linguistic features for aggression detection in code-mixed conversations.

2.2. Aggression behavior

Apart from aggression detection, researchers have also analyzed aggressive behavior and its effects on individuals and society. Cairns et al. (1988) found that aggressive adolescents may be unpopular with peers but may still be accepted in certain subgroups. Several studies have proposed methodologies to identify and quantify the effect of aggressive behavior, including negative content, specific keywords, and the influence of network dynamics (Chatzakou et al., 2017; Terizi et al., 2021; Sengupta et al., 2022). Online aggression has been found to impact individuals'

well-being and mental health, particularly university students (Mishna et al., 2018). The General Aggression Model (GAM) developed by Allen et al. (2018) explores the factors that influence aggressive behavior. (Ali et al., 2023) has used the Girvan Newman community detection algorithm to detect aggressive communities of social media influencers. Henneberger et al. (2017) found that users can be influenced to act aggressively and bully others because of high toxicity and aggression in their social circle. Eraslan and Kukuoglu (2019) found that counter-comments about participants' values significantly affected participants' aggressive tendencies. Studies have also explored the relationship between social media use and aggressive behavior. Boadi and Kolog (2021) used contemporary deterrence theory to examine how religion influences an individual's daily life and its direct influence on online aggression behaviors. Wong et al. (2022) investigated how motivation to obtain rewards through aggressive behavior can lead to cyberbullying, which can cause addiction to social media. Adinugroho et al. (2022) identified moral emotions and the frequency of social media usage as predictors of cyber aggression. The potential impact of fake media on social media users has also been examined by Galyashina and Nikishin (2022), who proposed anti-aggression strategies based on linguistic knowledge to detect fake media narratives. They highlighted linguistic markers that can help identify forgery and problematized the interconnection of fakes and violent speech aggression. The use of aggressive language on social media has been studied in various contexts. The various types of aggressive behaviors directed toward robots with AI enhancements has studied by Oravec (2023) Torregrosa et al. (2022) found that ideologically extreme political parties tend to use more aggressive language. Pascual-Ferrá et al. (2021) explored how aggressive language is used in the online discourse around wearing face masks during the COVID-19 pandemic. Vladimirou et al. (2021) examined how complaining is expressed in social media, which can become more aggressive due to the features of complex participation and multimodality. Chatzakou et al. (2019) categorized a user as aggressive if they had at least one aggressive post without conducting an analysis of the overall aggressive activity across all posts. Karan and Kundu (2023) identified patterns in bullying behavior and victim profiles on Twitter, incorporating factors like follower count, following count, and tweet frequency. They designed a multilingual framework for detecting aggression using a Fasttext-LSTM model after identifying the tweet language. Users were classified as bullies if a specified threshold percentage of their last 100 tweets exhibited aggression. However, relying solely on percentages to identify bullies has limitations, underscoring the need for more robust and nuanced approaches.

Overall, research on aggression on social media has established that aggressive behavior is widespread on online social media platforms. However, most of the computational work focused mainly on the detection of aggression, and there has been a relatively limited study investigating user behavior. Our studies contribute by employing various computational techniques on large-scale data to understand the aggressiveness of social media users and the potential causes of such behavior in terms of feeds.

3. User Aggression in Twitter

In order to address the research questions of aggressive behavior, we followed four stages methodology. These stages are i) Data collection and annotation, ii) Aggressive post detection, iii) Proposed metric of user aggression: aggression intensity, and iv) User profiling. Each stage is described in detail in the following sections, including the specific methods and techniques used, as well as the reasoning behind their selection. Figure 1 shows the representation of the overall flow of the work and the connections between the different stages.

3.1. Data Collection and Annotation

Firstly, we collected and prepared data for our research from the posts of Twitter. Twitter is a widely-used microblogging platform that allows users to share their thoughts and ideas in 280 characters or less. Despite its popularity, the platform has been criticized for not effectively addressing aggressive behavior among its users. Twitter's official policy states that it promotes freedom of speech but not freedom of reach. Taking advantage of this, most users engage in aggressive behavior under the guise of free speech. Currently, the platform relies on a manual reporting system to identify and address aggressive content. While Twitter has mechanisms in place to detect certain types of sensitive content, such as graphic violence and hateful imagery, it does not explicitly consider the text of a post when identifying aggressive behavior. The platform is organized around the use of hashtags, which are used to categorize and find posts on specific topics. The existing literature lacks large, accessible time-series datasets for behavioral studies. Previous publically available datasets have sparse Twitter posts, e.g., Kumar et al. (2018b) contains only 1257 entities. Others, like Bhattacharya et al. (2020), focused on YouTube comments, and Kumar et al. (2022), were biased toward only a few hashtags. Recognizing this gap, we developed a dedicated English aggression dataset of Twitter, filling an important gap and advancing research in this domain.

3.1.1. Data Collection

In order to initiate the crawling, we utilized manually selected hashtags associated with three major events in India which attracted intense reactions. The hashtags serve as seeds for our crawling, detailed in Table 1. The rationale behind selecting these three seed events is to get a large number of initial users. The details of the events are as follows:

1. **Agneepath**: The youth took to the streets to express their displeasure over the Government's newly launched Agneepath scheme, which led to massive protests across India.

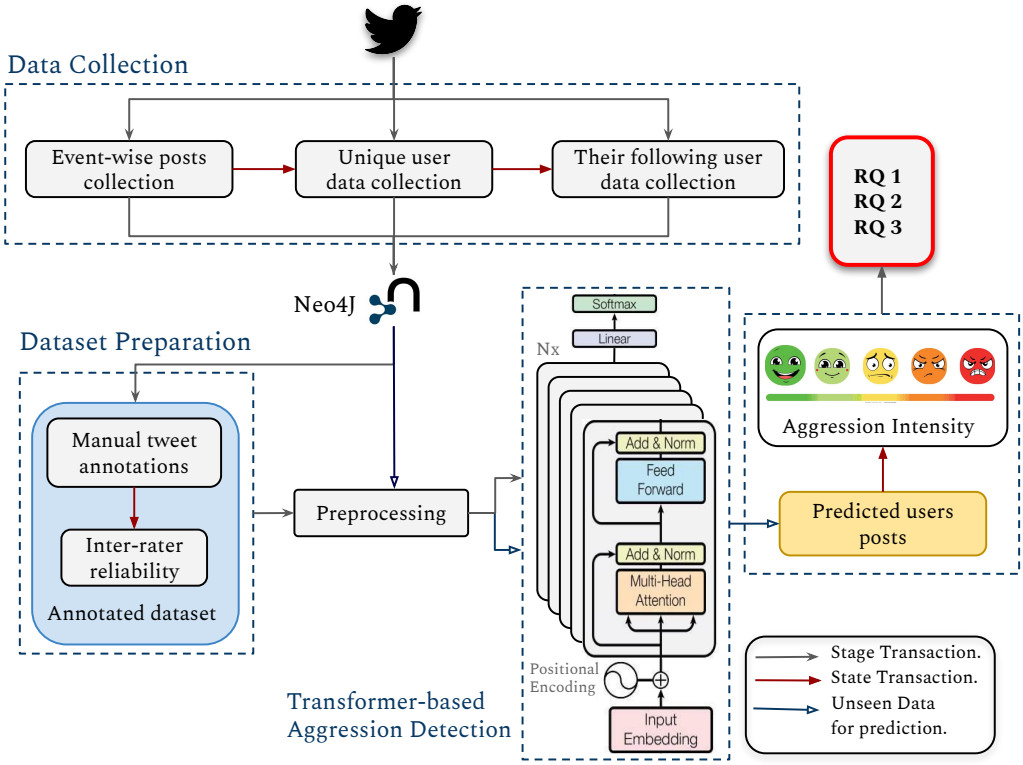


Figure 1: Flow of the methodology used to answer our research questions. The methodology includes four stages: data collection and annotation, aggressive post detection, user aggression intensity, and user profiling.

2. **KashmirFiles**: A movie was released, based on the exodus of Kashmiri Pandits from the Muslim-majority state of Jammu and Kashmir.
3. **ReligiousControversy**: A controversial statement by political leader Nupur Sharma led to Hindu-Muslim disputes, massive protests, and violence.

The extracted data spans six months from January 2022 to July 2022, encompassing periods before and after the events under consideration. We found 38K+ users participated in these events. In alignment with our research objective, we extended our data collection to include the neighbors (followings) of these users and their corresponding posts during the same timeframe. Consequently, our dataset includes not only content directly related to the seed events but also incorporates posts from unrelated events within the specified period (refer to Figure 2) and users unrelated to the seed events.

The data collection process was facilitated through the Twitter academic API, allowing comprehensive extraction of relevant data attributes, including text, language, engagement metrics, and user information for each post. Table 1 provides a summary of the initial event-related data extraction, presenting details such as the time period, event-specific hashtags, total extracted posts, the count of English-language posts, and the number of unique users associated with the data. To manage the vast amount of collected data effectively, we utilized the Neo4J native graph database,

Figure 2: Illustrates the events included in the collected data. It is important to note that these events are not limited to seed events, and their frequency of occurrence is not higher than other events.

renowned for its efficiency in handling large-scale real-time data. The heterogeneous nature of our graph $G = (v, e)$ is attributed to the diverse nodes, representing tweets, users, language, and hashtags ($v : \{v_t, v_u, v_l, v_h\}$), each linked by heterogeneous relationships ($e : \{e_f, e_l, e_m, e_p, e_q\}$), such as following_of, language, mentions, posts, and quoted. An illustrative representation of this heterogeneous graph is

Table 1
Description of the seed event-related data extracted from Twitter.

Time-period	1^{st} Jan 2022 to 15^{th} July 2022	
Seed Events hashtags	**Agneepath**	#agneepathprotest, #agneepathyojana, #AgnipathScheme, #agnipathschemeprotest, #agnipathschemeprotests, #agnipathprotest, #Agnipath, #AgnipathProtest, #AgnipathProtests,
	KashmirFiles	#StopPakSponsoredTerrorism, #KashmirAgainstTerrorism, #AakhirKabTak, #KashmiriPandits, #kashmirihindus
	ReligiousControversy	#MuslimsUnderAttackinIndia, #KanhaiyaLal, #HindusUnderAttack, #NupurSharma, #NupurSharmaControversy
Total posts	339,390	
English posts	175,606	
Unique users from English posts	38,680	

presented in Figure 3. The complete graph comprises extensive 43,668,610 (43M+) nodes and 134,740,385 (134M+) edges. For this study, we focused on a specific heterogeneous sub-graph (G_{sub}), which is connected to the English node, resulting in a graph with 15,113,304 nodes and 46,971,867 edges. Specifically, there are 14M English tweet nodes (v_t) originating from 63K users (v_u). The rest of the nodes correspond to the hashtags in the subgraph. This refined graph dataset serves as the foundation for our further analyses and conclusions.

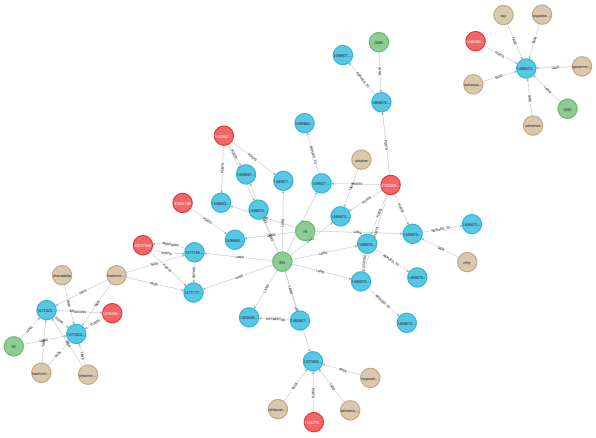


Figure 3: Representation of a heterogeneous graph stored in Neo4J. Nodes are categorized by color: blue for tweets, orange for users, green for language, and pale yellow for hashtags.

3.1.2. Data Annotation

For data annotation, we selected Tweet nodes (v_t) randomly from our heterogeneous sub-graph (G_{sub}), aiming for diversity. Annotators from diverse racial, regional, cognitive, cultural, and religious backgrounds were selected to ensure a generalized annotation. Manual annotation was conducted by a group of five cultural diverse annotators, consisting of two graduates, one post-graduate, one research scholar, and one academic expert. Table 2 outlines annotator attributes, including experience, expertise, and relevant demographics. Annotators are proficient in understanding English social media posts and have key characteristics such as ages ranging from 22 to 38 years, research expertise in NLP (Natural Language Processing) and CSS (Computational Social Science), and 1-4 years of experience. They had no religious extremes, no affiliation with political organizations, and were active on social media. Annotators adhere to the annotation guidelines and annotation process shown in Figure 4. We referred to the definition and dimensions analyzed by Mane et al. (2023). Cyber-aggression refers to harmful intentional online behavior, irrespective of whether it is overt or covert. It includes various dimensions such as the use of hostile, offensive, insults, threats, and abusive comments intended to cause discomfort, distress, or harm to individuals or communities. The objective of our manual annotation was to determine whether a tweet exhibits aggression or not. A non-aggressive tweet was labeled as NAG. However, in the case of an aggressive tweet, our focus was on identifying the reasons behind it. If the reasons did not relate to the stated dimensions of aggression, the text was flagged for further discussion. Otherwise, it was labeled as AG. Initially, annotators were given specific examples for each category, along with explanations that justified the labels assigned to it. Each tweet was annotated by two annotators, with disagreements resolved through expert discussion. Table 3 presents instances of annotated tweets. The final annotated dataset namely ***T**witter **Ag**gression dataset **(TAG)***, had a total of 7,812 posts, of which 3,416 were labeled as aggressive "AG" and 4,396 were labeled as non-aggressive "NAG". Also, the choice of annotating 7,812 posts was based on the time and resources available for the annotation task.

To ensure the consistency and reliability of the annotation process, we employed Krippendorff's Alpha evaluation metric (Hayes and Krippendorff (2007)), a widely recognized method for assessing inter-annotator agreement. Krippendorff's Alpha is calculated as follows:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

Where, D_o represents the observed disagreement and D_e is the expected disagreement. The numerator D_o is computed as the sum of the pairwise disagreements between annotators for each tweet in the annotated dataset. In our study, this

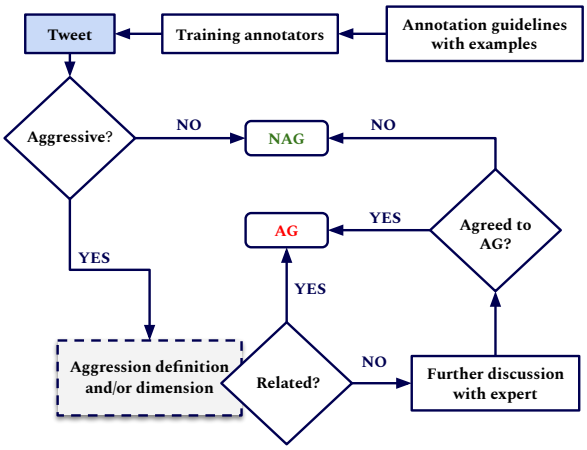


Figure 4: Depicts the annotation process utilized by annotators for effectively categorizing posts as either AG or NAG, resolving any conflicts that may arise during the process.

is the count of instances where annotators disagreed on the aggression label (AG or NAG).

$$D_o = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=j+1}^{A}\delta_{ijk}$$

Here, N is the number of tweets in the annotated dataset, and A is the number of annotators. δ_{ijk} is an indicator function that equals 1 if $(j \neq k)$ annotators j and k disagree on tweet i, and otherwise 0. The denominator D_e is the expected disagreement, calculated as the sum of the expected pairwise disagreements based on a chance agreement between annotators. For binary, the chance agreement is calculated as:

$$D_e = \frac{1}{2}\sum_{c=1}^{2}\sum_{k=1}^{A} p_{ck}(p_{ck}-1)$$

Where, p_{ck} is the proportion of tweets labeled as category c (AG and NAG) by annotator k. The resultant coefficient (α) ranges from -1 to 1, with a score of 1 indicating perfect agreement, 0 indicating no agreement, and negative values indicating conflicting agreement. The calculated α in our study was 0.7873, signifying a substantial level of agreement among annotators. The analysis of inter-annotator agreement is crucial for ensuring the robustness of our annotated dataset. Annotators encounter challenges in identifying covert aggression, potentially causing lower agreement.

3.2. Aggressive Post Detection

In this section, we present an aggression detection model. We formulate aggression detection for social media posts as a binary classification task. We fine-tunned the pre-trained RoBERTa-large model (Liu et al., 2019), which contains 24 layers, 1024 hidden units, 16 attention heads, and 355M parameters for the said task. The RoBERTa encoder uses a multi-head self-attention sub-layer to transform input representations $H \in \mathbb{R}^{n\times d}$ into contextual embeddings $Z \in \mathbb{R}^{n\times d}$. This projects queries Q, keys K, and values V to compute attention weights as:

$$\text{Attention}(H) = \sigma\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Where σ is the softmax function. This is followed by a feedforward sub-layer that applies two affine transformations with GeLU activation:

$$\text{FFN}(x) = \max(0, x\cdot W_1 + b_1)\cdot W_2 + b_2$$

For an input sequence $x = \{x_1, x_2, \ldots, x_n\}$ representing a tokenized Twitter post, contextual token embeddings $h = \{h_1, h_2, \ldots, h_n\}$ are generated by RoBERTa-large. A [CLS] token embedding $h_{\text{CLS}} \in \mathbb{R}^d$ aggregates the sequence where d is the hidden dimension. We add a classification layer with weights $W \in \mathbb{R}^{2\times d}$ and biases $b \in \mathbb{R}^2$ to predict aggression probabilities via the softmax function:

$$p(y=1|x) = \sigma(W\cdot h_{CLS} + b)$$

We optimize all parameters Θ using binary cross-entropy loss L:

$$L(\Theta) = -[y\log(p(y=1|x)) + (1-y)\log(1-p(y=1|x))]$$

A post is labeled aggressive if $p(y=1|x) > 0.5$. We set the maximum sequence length to 64 tokens based on TAG text lengths. Adam (Kingma and Ba, 2014) optimizer is used with a learning rate of 1×10^{-6}, weight decay of 0.1, dropout of 0.2, and a batch size of 32 for 100 epochs. The experiment stops when there is no validation loss improvement, however minimum 10 epochs where considered even if there is no change of validation loss. The resultant optimized model achieves a macro-F1 score of **0.92** on the TAG test set. We empirical this performance is found benchmark on this dataset. This achieves benchmark performance on the dataset. The detailed ablation study is presented in Section 5. We use this model to predict labels for unlabeled tweet nodes (v_i) within the heterogeneous sub-graph (G_{sub}) and use it further experiments.

3.3. Proposed Metric of User Aggression: Aggression Intensity (AI)

We proposed a metric for measuring the aggression intensity of users on Twitter. By using this metric, we are able to calculate the total aggressive behavioral activity of a user over a specific period of time, such as a week, day, or hour. This metric is important to measure the aggressive activity of users on social media platforms. In the user's aggression intensity metric, we first calculated the user aggressiveness aggregated score by the fraction of the total aggressive posts of a user in a given period AG_i^l and the total posts of that user in that period X_i^l. This score gives us the rate by which the user submits an aggressive post. However, this score alone may not be sufficient to accurately measure the aggression intensity of a user. Specifically,

Table 2
Provides information about important traits of the annotators, emphasizing the heterogeneity within the annotators.

	A-1	A-2	A-3	A-4	Expert
Research-status	Under-graduate	Under-graduate	Post-graduate	Doctoral student	Professor
Research-filed	NLP	NLP	CSS	NLP, CSS	NLP, CSS, AI, NS
Experience	1 year	1 year	2 year	4 year	20 year
Region	Maharashtra, India	Delhi, India	Rajasthan, India	Maharashtra, India	West Bengal, India
Age	22	23	28	26	38

Table 3
Presents examples from the manually annotated dataset. References to users have been anonymized by replacing them with the term *'anonymous'*.

Tweets	Label
Hey *@anonymous*, write something on bbc, shameless international media #*religon*UnderAttack	AG
It's me with my sister in tents — my childhood memories #happyMemories	NAG
Where is baba ki bulldozer This is new india. "Duplicacy is the Constant weapon of a rougue" RIP Equality *@anonymous* #AgnipathScheme	AG
@anonymous They just need a reason to show their street power and psycho warfare schemes. This time the protest is against #*anonymous*.	AG
@anonymous But excellent work By police team! Keep it up #jaiHind	NAG
Use their own laws against them and behead them. #*anonymous*. At times tit for tat is the best solution.	AG

consider two hypothetical users X and Y. User X has posted 2 aggressive posts out of a total of 4 posts. In contrast, user Y has made 10 aggressive posts, greater in absolute terms, but out of 30 total posts. Simplistically judging by aggressiveness aggregated score penalizes X more severely despite lower relative hostility levels. The former, who is less active, now has stronger aggression intensity, which is not desired. To counter this, we multiplied the aggressiveness aggregated score by the normalized value of the total number of posts as mentioned in Eq. 2 and called it the aggressive intensity score AI_i^l. Here, min^l and max^l are the minimum and maximum number of total posts by any users for that period, respectively. This normalization score ensures that all users are penalized appropriately, regardless of the total number of posts they make. This score ranges from 0 to 1, with 1 representing the highest aggressive behavior and 0 representing non-aggressive behavior.

$$AI_i^l = \frac{AG_i^l}{X_i^l} \times \frac{X_i^l - min^l}{max^l - min^l} \tag{2}$$

where

$$\begin{cases} 1 \leq min^l < max^l \\ min^l < max^l \leq n \end{cases} \tag{3}$$

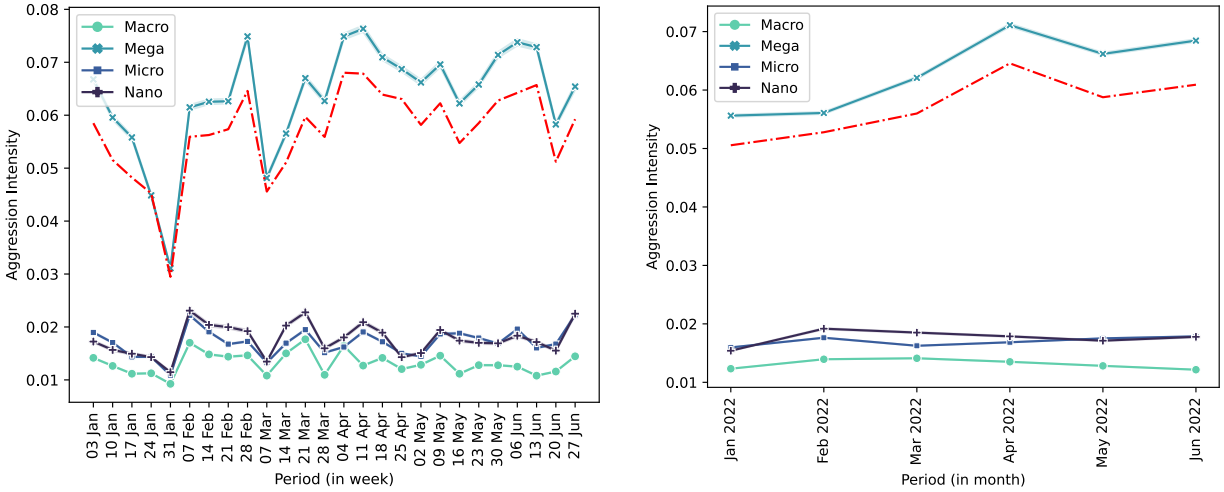
The aggression intensity metric is essential in our research as it allows us to accurately measure the aggressive behavior of users over time, and it is a key factor in addressing our research question.

Who is more aggressive?

In social media, users with a high number of followers have the potential to spread information and impact people quickly (Zhang et al., 2017; Wang et al., 2022). This is why understanding the behavior of influential users is crucial. In order to understand the aggression of influential users on Twitter, we analyzed the aggression intensity of users based on their level of influence. To achieve this, we have created four user buckets called cores based on the number of followers they have, which are defined as follows:

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, analysis reveals that users classified as Mega exhibit significantly higher aggression intensity compared to other users, over both weekly and monthly time periods ($p<0.05$). This effect is statistically significant based on a t-test comparing the aggression intensity distributions between Mega and non-Mega users for both time scales. The aggression intensity levels for Mega users were also found to be consistently above the threshold for aggressive behavior. The threshold we calculated by the addition of the mean and standard deviation of aggression intensity. The observation reveals that users with a high number of followers are more likely to engage in potentially aggressive behavior in this study. This may have a greater impact on their followers. This finding motivates us to do research on the relationship between the aggressive behavior of users and their following users on social media platforms like Twitter, which is addressed in section 4.1.



(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.

3.4. User Profiling

We profile the user based on their behavior intensity over time. We used the K-means clustering algorithm (Jain et al., 1999; MacQueen, 1967) to determine an appropriate aggressive level (e.g., low, high) of user behavior based on their aggression intensity score. This algorithm requires a value of K, which represents the number of clusters. Rather than assuming a fixed number of clusters, we used the K-means elbow method (Hardy, 1994) to select the optimal value of K. This method is commonly used to determine the number of clusters in a dataset by finding the point of inflection, or "elbow". Based on our analysis, the optimal value of K for all users aggression intensity was 2, indicating the presence of two clusters: low and high aggressive levels of users. Thus, we applied the K-means algorithm (with K=2) to the users aggression intensity scores across weeks. Finally, we profile the user based on the vector of user behavior over time. The low and high clusters are used to create a vector that represents the overall user's behavior. The vector encoding is a sequence of low and high levels, where low represents non-aggressive behavior, and high represents aggressive behavior. An example of an aggressive user vector is shown in upper Figure 6, in which each cell represents a week, and the length of the vector is the number of weeks considered for the user's profile. If the proportion of "High" levels is greater than the "Low" levels in the vector, we consider the user profile to be aggressive. Similarly, lower Figure 6 shows an example of a non-aggressive user profile.

Afterward, we analyzed the followers and following of aggressive and non-aggressive users (Figure 8). Additionally, we also analyzed the number of friends that aggressive and non-aggressive users have and found that aggressive users tend to have fewer friends. Furthermore, we observed that some of the top-targeted users are popular, but not all (Figure 8c, 8d). We identified top-targeted users that are

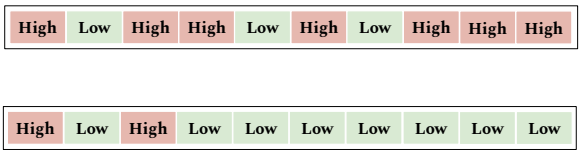


Figure 6: Vector representation of user behavior over a period. Upper for aggressive user profile, lower for non-aggressive user profile

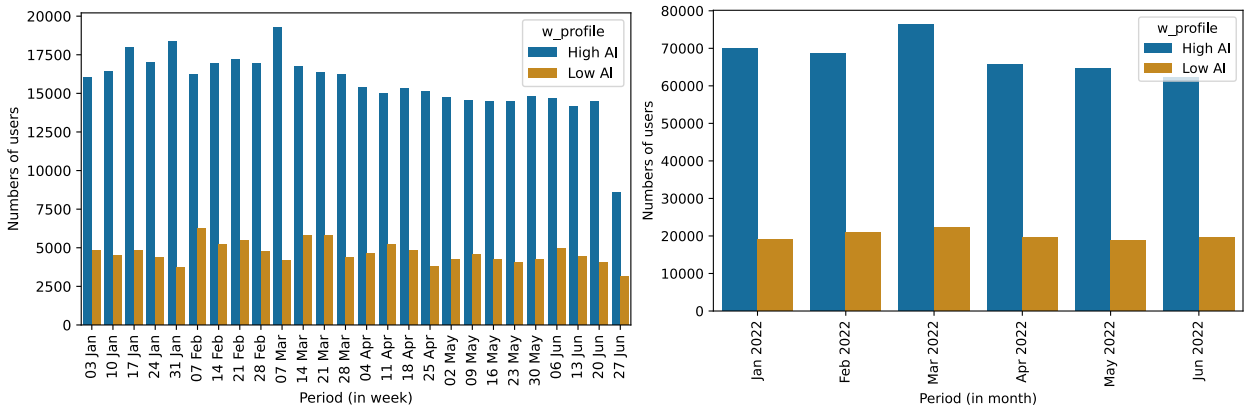
frequently mentioned in aggressive and non-aggressive posts (Figure 9). We found that there was no correlation between the number of followers and the number of mentions in both aggressive and non-aggressive posts. For example, the user having the highest mention of 8262 has only 769440 followers. Similarly, for non-aggressive, the user having the highest mention of 13086 has only 2963 followers.

4. Research Questions and Findings

In this section, we address our research questions to examine whether user behavior becomes aggressive after exposure to aggressive feeds and events, and we examine the relationship between user engagement and aggressive feeds.

4.1. RQ1: Do aggressive feeds can make someone aggressive?

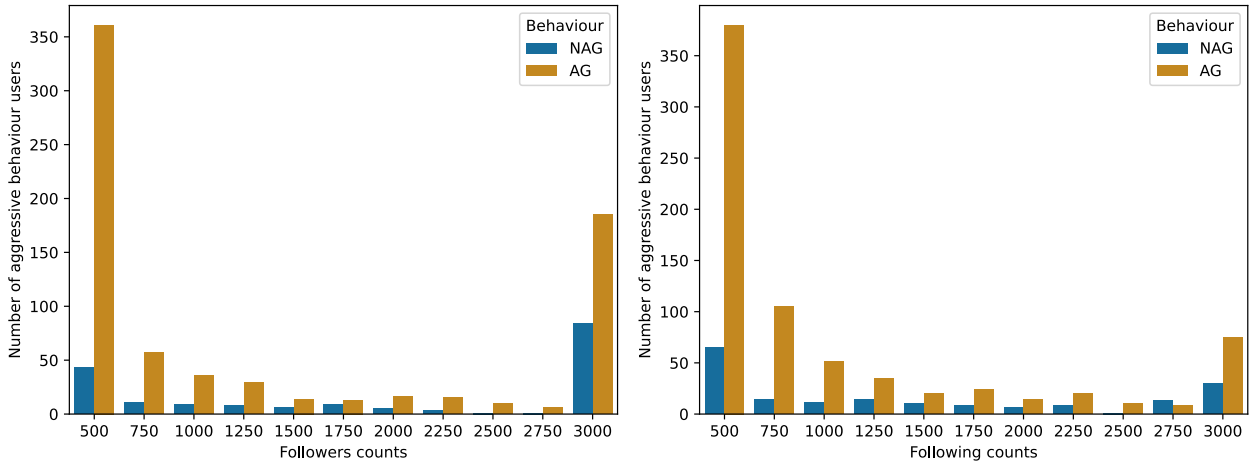
We answered this question by performing two distinct types of analysis. Firstly, we examined the level of aggression exhibited by users on a particular day, as well as the aggression level of their feeds (defined as posts from accounts they follow). Secondly, we investigate whether a user posted an aggressive post after being exposed to an aggressive feed.



(a) Week-wise

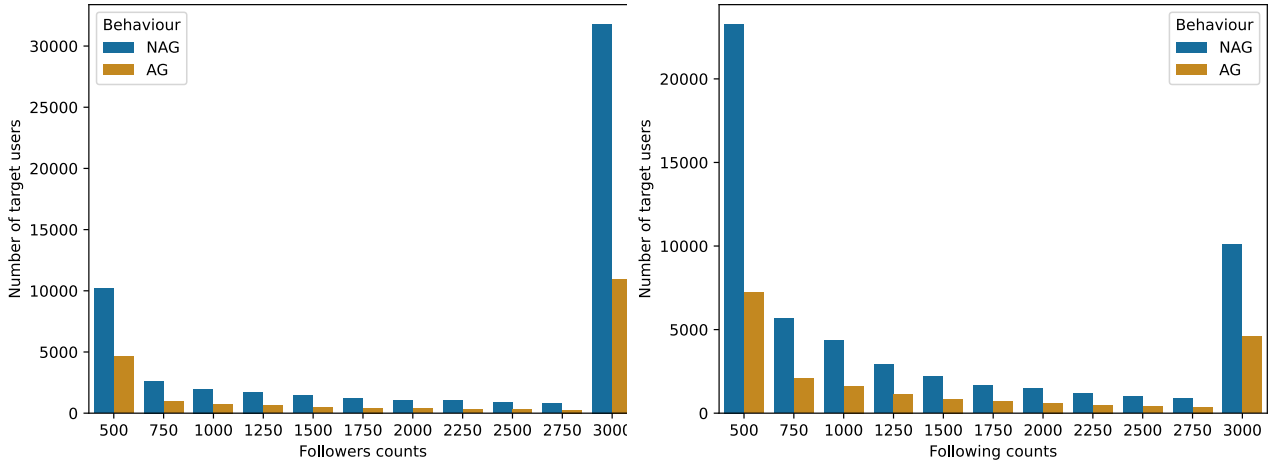
(b) Month-wise

Figure 7: Number of low and high aggressive intensity users.



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

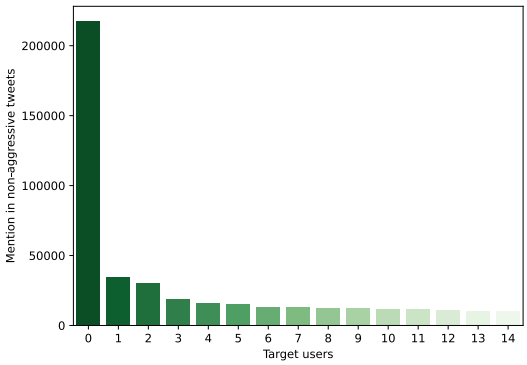
(b) Analysis of aggressive and non-aggressive user followings.



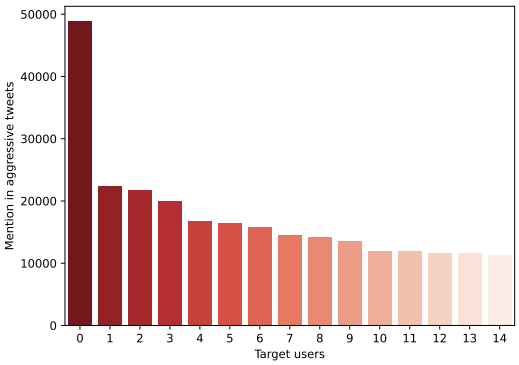
(c) Analysis of the followers of target users, which is mentioned by aggressive and non-aggressive users.

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

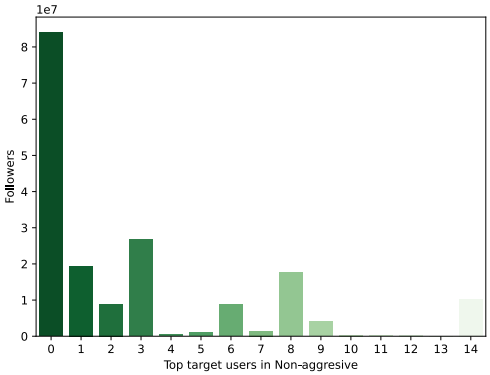
Figure 8: Followers and following analysis of potentially aggressive and non-aggressive users, and their target users.



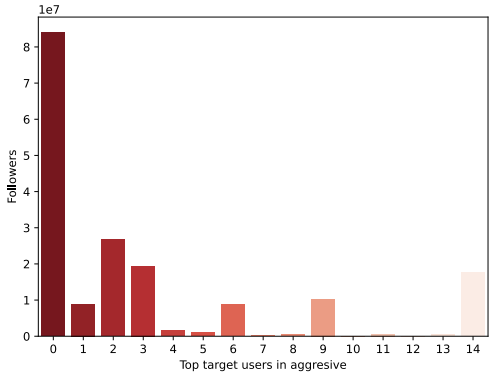
(a) Top target users mentioned in the non-aggressive tweets.



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



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



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

Figure 9: Analysis of top target users is mentioned in the aggressive and non-aggressive tweets.

For examining the level of aggression exhibited by users, we analyzed the data at the granularity of day, as hourly data was insufficient. We calculated the aggression intensity of the user (using Eq. 2) for the 24-hour period (l = day). The feed intensity of the user is the aggregation of the aggression intensity of the users who are followed by the user over the present and previous day (as per Eq. 4).

$$FU_i^l = \frac{\sum_{j=1}^{n} FAI_{ij}^{l} + \sum_{j=1}^{n} FAI_{ij}^{l-1}}{n} \tag{4}$$

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 user j who is followed by the user i on the period l = current day and $l-1$ = previous day.

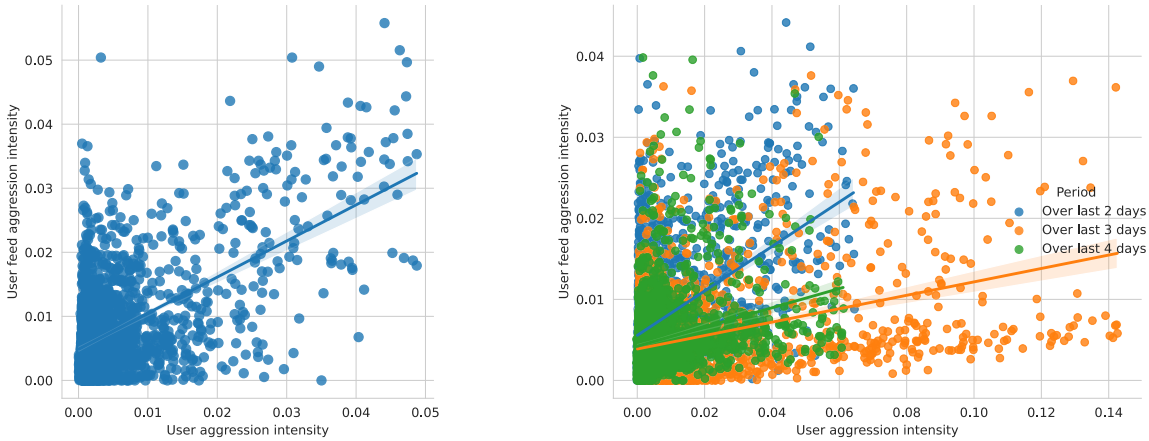
Following are the null and alternative hypotheses:

- H(0) = There is no relationship between the user and their feed aggression intensity.
- H(a) = There is a relationship between the user and their feed aggression intensity.

Further, we tested hypothesis (Fisher, 1992), using the Pearson correlation coefficient (Cohen et al., 2009) of user intensity and their feed intensity. The result showed a moderately strong positive correlation with a coefficient of 0.58085 ($p<0.05$). At the user level, we conducted an analysis of the correlation between individual users and their corresponding feeds over a defined time period. This analysis found that 90% of users exhibited a statistically significant positive correlation with their feeds ($p<0.05$). That is, only 10% of the users post aggressively even if their feed is non-aggressive and vice-versa. This positive correlation suggests that the aggression levels exhibited by users are associated with the aggression present in their feeds of present and previous day (Figure 10).

Additionally, we also calculate the correlation between present-day user intensity and feed intensity over present-day and "X" previous days, where X varied from 2 to 4. For X = last two days, we obtained a correlation of 0.4692, for the last three days 0.3695, and for four days correlation: 0.2974. The results showed that the correlation gradually decreases with an increase in previous feeds, implying that the impact of aggressive feeds is mostly short-term.

Finding 1. *A correlation exists between the intensity of users' aggressive behavior and the level of aggression in their feeds. Furthermore, the effects of exposure to aggressive feeds are generally short-term, indicating that it may not have long-lasting effects on users.*



(a) Relation between aggression intensity of user and their feed from the last day.

(b) Relation between aggression intensity of user and their feed over period wise.

Figure 10: Relation between aggression intensity of user and their feed. The relation gradually decreases with an increase of last days feeds.

Next, we identify whether an aggressive post is a result of aggressive feed or not. In order to do so, we analyzed the feed in the last 24 hrs before a particular post was created. The proportion of aggressive posts is considered the aggressiveness of the feed as shown in Eq. 5. We called this as feed intensity of the post.

$$FT_i = \frac{AGF_i}{n_i} \tag{5}$$

Where FT_i is the aggressive feed intensity of post i, AGF_i is the total aggressive feeds of post i, and n_i is the total feeds of post i.

We calculated the feed intensity of all the posts, irrespective of aggressive or non-aggressive, and compared them. Figure 11 illustrates the proportion of aggressive and non-aggressive posts in relation to the feed aggression intensity. The x-axis shows feed intensity, and the y-axis shows the ratio of aggressive and non-aggressive posts for each value of feed intensity. A value of a ratio greater than 1 indicates more aggressive posts, while a value less than 1 indicates more non-aggressive posts. It is visible that all the points except few above feed intensity 0.65 show a high post ratio > 1. This analysis reveals that as the feed aggression intensity of posts increases, the majority of users tend to post only aggressive posts, with a lower proportion of non-aggressive posts.

Afterward, we tested the hypothesis using a two-tailed student's t-test (Student, 1908), which shows that aggressive posts had a higher feed intensity compared to non-aggressive posts, with a $p-value$ of 1.37×10^{-13} ($p-value < 0.05$). This suggests that users are more likely to post aggressive content when their recent feeds have been aggressive in nature.

Finding 2. *Aggressive posts have been found to have more aggression in their feeds than non-aggressive posts. This suggests that exposure to aggressive content on Twitter may have a greater impact on users than exposure to non-aggressive content.*

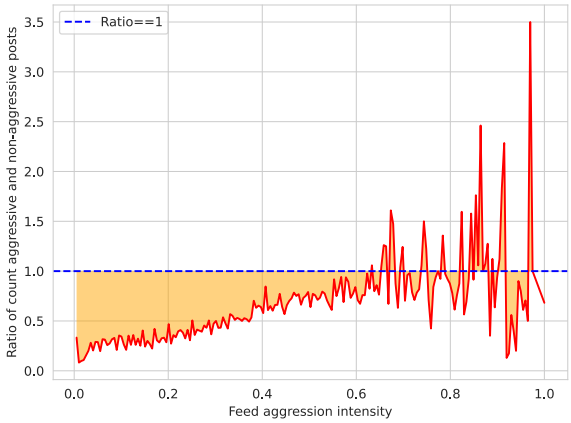


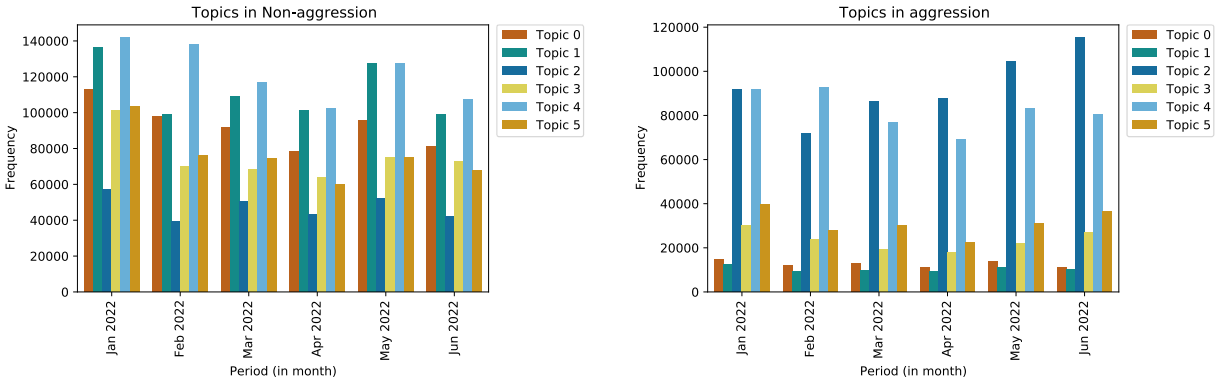
Figure 11: Feed aggression intensity wise ratio of counts aggressive and non-aggressive posts. The ratio increases along with feed intensity.

4.2. RQ2: Does exposure to event-aggressive feeds increase the user's event-specific aggressive behavior?

We addressed this research question by analyzing the effect of different topics and events on user aggression with consideration of user posts and their feeds. First, we identified the most frequently discussed topics using BERTopic (Grootendorst, 2022) to extract the top six topics and their associated keywords, as presented in Table 4. We further assigned potential topic names based on the context of the topic keywords. To examine the effect of topics on user aggression, we analyzed the frequency of topics discussed

Table 4
Top topics and their top keywords are discussed in collected data.

Topic 0: About Movie	Topic 1: Celebration	Topic 2: Religious Conflict	Topic 3: Indian Politics	Topic 4: Geopolitics	Topic 5: Spiritual leaders
congratulation	good	hindu	minister	india	bjp
time	happy	muslim	india	people	god
day	morning	kashmir	congress	ukraine	temple
true	thanks	terrorist	country	world	congress
movie	beautiful	police	student	russia	hindu
new	birthday	india	support	country	dharma
project	love	woman	agree	right	yogi
legend	best	killed	state	need	leader
film	like	pakistan	bjp	war	maharaj
madness	day	temple	new	china	truth



(a) Non-aggressive

(b) Aggressive

Figure 12: Month-wise analysis of topics in the aggressive and non-aggressive activity of users.

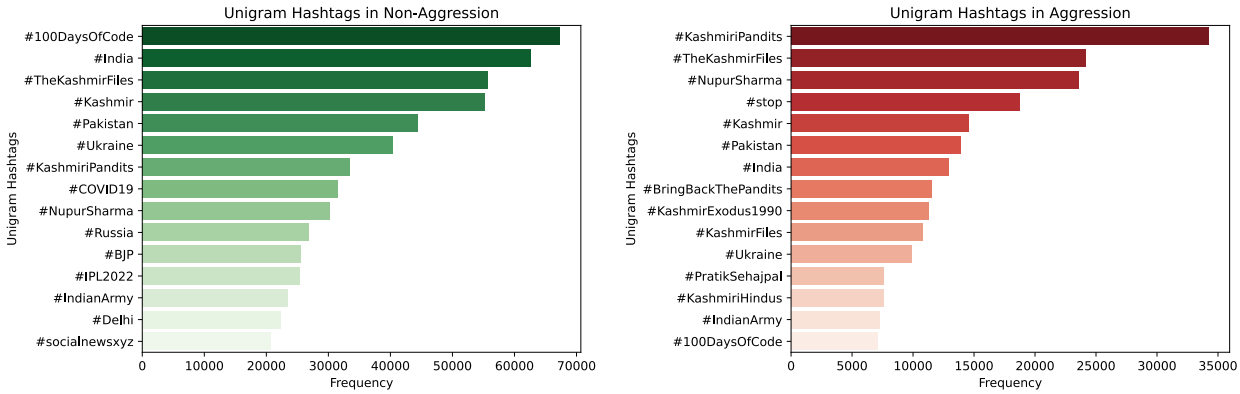
in aggressive and non-aggressive posts from January 2022 to June 2022, as shown in Figure 12. We observed that the order of frequency of topics discussed aggressively was consistent, with Topic 2: Religious Conflict being the most discussed, followed by Topic 4: Geopolitics, Topic 5: Spiritual Leaders, and Topic 3: Indian Politics. In contrast, the order of topics in non-aggressive discussions was not consistent, but all topics except Topic 2 were discussed the most. This finding suggests that there is a difference in the frequency of topics discussed between aggressive and non-aggressive posts. Moreover, we found that some topics, such as Topic 3: Religious Controversy, were more likely to trigger aggressive behavior among users. This conclusion was based on the consistently high frequency of this topic in aggression and consistently low frequency in non-aggression discussions. Thus, our study suggests that the topic of a post can influence user behavior toward aggression.

Next, we conducted an event analysis by considering the hashtags of each post as the event. Figure 13 illustrates the frequency of events involved in both aggressive and non-aggressive discussions. Our observations reveal that the most popular events in aggressive and non-aggressive discussions differ significantly. Specifically, we observed that several aggressive discussions were often related to events such as '#pakistan', '#BJP', '#NupurSharma', '#ImranKhan', '#Congress', '#JagoKashmir', and '#Hindus'. In contrast, these events were not as commonly discussed in non-aggressive discussions. This suggests that users are more likely to engage in aggressive behavior when discussing certain events.

To further explore the relationship between event-specific feeds and user aggression, we selected three seed events (as shown in Table 1) and calculated the aggressive intensity score for each event using Eq. 6. This analysis considered the collective posts of users as well as their feeds.

$$AIE_i = \frac{AGE_i}{XE_i} \times \frac{XE_i - minE}{maxE - minE} \tag{6}$$

Where AIE_i is the aggressive intensity of an event i, AGE_i is the total aggressive posts of an event i, XE_i is the total posts of the event i, $minE$ and $maxE$ are the minimum and maximum posts among all events, respectively. The results showed that the order of event aggressive intensity for user posts is $Event_3 > Event_1 > Event_2$, while the order of event aggressive intensity for user feeds is $Event_1 > Event_2 > Event_3$. Notably, the aggressive order of $Event_1$



(a) Non-aggressive activity

(b) Aggressive activity

Figure 13: Analysis of Events used in the aggressive and non-aggressive activity of users.

and $Event_2$ was consistent across user activity and their feeds. These findings suggest that event feeds can have a significant impact on the level of aggressiveness in users' posts.

To determine whether event-related feeds increase users' event-specific aggressive behavior, we calculated the event-specific feed aggression intensity of each user's event-specific post. We analyzed the respective event-specific posts of the feed that were created 24 hours before the particular post of the respective event. The proportion of event-specific aggressive posts was considered the aggressiveness of the event-specific feed. We called this the event-specific feed intensity of the post, as shown in Eq. 5. AGF_i represents the total aggressive feeds for a user i for a respective event E and n_i is the total posts of feeds of a user i. Our analysis aimed to explore the relationship between event-specific user posts and their aggressive feed intensity of the respective events. We found that the intensity of aggression in event-related feeds has an impact on the level of aggressiveness exhibited by users in their behavior towards that specific event. We demonstrated this through Figure 14, which illustrates the proportion of event-specific aggressive and non-aggressive posts in relation to the feed aggression intensity of the respective events. The x-axis shows event-specific feed intensity, and the y-axis shows the ratio of aggressive and non-aggressive posts for each value of feed intensity of the respective event. A value of a ratio greater than 1 indicates more event-specific aggressive posts, while a value less than 1 indicates more event-specific non-aggressive posts. It is visible that all the points except few above feed intensity 0.7 show a high post ratio $>$ 1. We observed that as the event-specific feed aggression intensity of posts increased, the majority of users tended to post only aggressive posts, with a lower proportion of non-aggressive posts related to the respective event. Our analysis revealed that the number of event-specific aggressive posts by users increased with the intensity of feed aggression of the respective events. To confirm this hypothesis, we conducted a two-tailed student's t-test, which yielded a $p-value$ of 4.237×10^{-14} ($p-value <$ 0.05) and revealed that aggressive posts had a significantly higher aggressive feed intensity for their respective events. This finding highlights the potential impact of aggression in event feeds on user behavior and attitudes toward specific events. Overall, these findings suggest that event feeds can significantly impact the level of aggressiveness in users' posts. Also, certain events tend to provoke more aggressive behavior compared to others.

Finding 3. *The event-related feeds can significantly influence the level of aggressiveness in users' event-specific posts, with event-related posts having a higher level of event-specific aggressive feed. This indicates that users are more likely to exhibit aggressive behavior when discussing certain events and topics on social media platforms.*

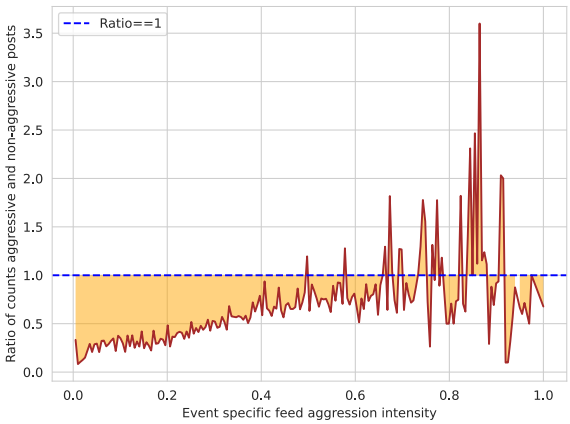


Figure 14: Event specific feed aggression intensity wise ratio of counts aggressive and non-aggressive posts. The ratio increases along with event specific feed intensity.

4.3. RQ3: Do users engage more with aggressive posts?

This research question has investigated whether users tend to engage more with aggressive posts on social media.

To assess this, we analyzed the engagement factors of users, such as likes, quotes, replies, and retweets, of approximately 15 million posts from November 28, 2021, to July 10, 2022. We normalized the count of each engagement factor from 0 to 1 and observed their inclination towards aggressive activity. Our analysis found that engagement factors for likes, quotes, and retweets towards aggressive posts were consistently higher than non-aggressive posts over weeks, except for replies. To further quantify the overall engagement level of each post, we calculated an overall engagement score by averaging the total number of quotes, likes, replies, and retweets received by each post and normalizing it. In Figure 15, we found that the overall engagement score is consistently higher for aggressive posts than non-aggressive ones over weeks. Our findings suggest that users tend to engage more with aggressive posts on Twitter. We confirmed this hypothesis through a two-tailed student's t-test, which yielded a $p - value < 0.05$ for overall engagement and engagement factors like likes $p - value$ is 3.7×10^{-4}, for quotes $p-values$ is 5×10^{-10}, and for retweets $p-values$ is 4.6×10^{-4}, indicating a significant difference in engagement between aggressive and non-aggressive posts. However, for the engagement factor replies, we found no significant difference between aggressive and non-aggressive posts ($p-value$ is 0.36), indicating that users engage similarly with both types of posts. Overall, our analysis highlights that users are more inclined to encourage and support aggressive content on social media.

Finding 4. *Aggressive posts on social media tend to have higher user engagement compared to non-aggressive posts. These findings suggest that aggressive content on social media may attract more attention and engagement from users, which could further amplify the spread of such content.*

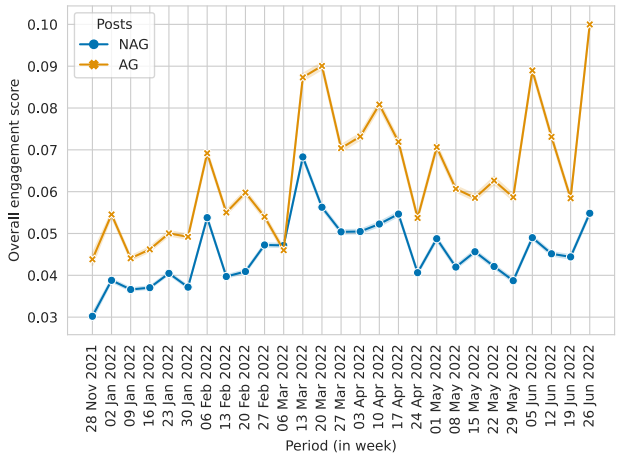


Figure 15: Week-wise overall user engagement with respective aggressive and non-aggressive posts. The user overall engagement of aggressive posts is consistently higher than non-aggressive posts

Table 5
Dataset Statistics for Aggression Detection, including Training, Validation, and Testing Sets. Here 'FB' denotes Facebook posts.

Datasets	Train		Val		Test		Total
	AG	NAG	AG	NAG	AG	NAG	
TRAC 1	6948(FB)	5051(FB)	1768(FB)	1233(FB)	774	483	16257
TAG (our)	2767	3560	307	396	342	440	7812

5. Ablation Study

In this section, we present an ablation study of aggression detection models used in our experiment. We explore different transformer-based model inplace of RoBERTa-large on proposed TAG dataset. The performance of models is also validated on publicly available TRAC 1 dataset (Kumar et al., 2018b). Key statistics of the datasets are shown in Table 5. Further, the effect of transfer learning is investigated. Finally, we evaluate the capability of the Large Language Model (LLM) for aggression detection using zero-shot learning.

5.1. Why RoBERTa-large?

Fine-tuning pre-trained language models has become a common technique for sequence classification in natural language processing (NLP) tasks and it provides state-of-the-art performance (Vaswani et al., 2017). We adopt this approach for aggression detection, evaluating several monolingual, multilingual and code-mixed transformer-based models: BERT (Devlin et al., 2018), RoBERTa, and XLNet (Yang et al., 2019) pre-trained on English; multilingual models XLM-RoBERTa (Bertin et al., 2019a) and mBERT (Bertin et al., 2019b); and Hindi-English models HingBERT, HingMBERT, and HingRoBERTA (Ravindran and Joshi, 2022). Additionally, we fine-tune twitter-roberta-base-offensive (Barbieri et al., 2020) and RoBERTa-hate-speech (Vidgen et al., 2021), leveraging large Twitter and Facebook hatred datasets. We include DistilBERT variants (Sanh et al., 2019), which reduce model size while retaining performance. By evaluating this diverse set of architectures, we aim to determine effective approaches for Twitter aggression detection. We fine-tuned each model on the TAG and TRAC 1 datasets and then evaluated their performance on the corresponding test set for each dataset. We utilize the TRAC 1 dataset by consolidating its overt and covert aggression labels into a unified aggressive class to align with our binary labeling. For this experiment, we utilized the parameters similar to those employed in Section 3.2. As shown in Table 6, RoBERTa-large obtained state-of-the-art performance with 0.92 macro F1 on TAG, outperforming other architectures. This strong generalization extended to the TRAC 1 dataset, with RoBERTa-large again achieving the top performance. The RoBERTa-large has additional pre-training techniques like dynamic masking and larger mini-batches that lead to more robust language representations. These modeling improvements enhance RoBERTa-large's generalization ability for aggression detection. In contrast, models specialized in offensive language or hate

Table 6
Evaluation of Fine-tuned pre-trained transformers on TAG and TRAC 1 using average macro-F1.

Models	TAG	TRAC1
$BERT_{base}$	0.8621	0.6863
$BERT_{large}$	0.8937	0.7075
$M\text{-}BERT_{base}$	0.8571	0.7596
$M\text{-}distillBERT_{base}$	0.8161	0.6860
$DistillBERT_{base}$	0.8541	0.7097
$DistillBERT_{base-squad}$	0.8639	0.7036
Hing-BERT	0.8698	0.7376
Hing-mBERT	0.8602	0.7714
Hing-RoBERTa	0.8718	0.7807
$RoBERTa_{base}$	0.9017	0.6996
$RoBERTa_{large}$	**0.9185**	**0.8052**
$Xlm\text{-}RoBERTa_{base}$	0.8722	0.7068
$Xlm\text{-}RoBERTa_{large}$	0.9030	0.6325
$Xlnet_{base}$	0.8930	0.6640
$Offensive\text{-}RoBERTa_{base}$	0.7633	0.7527
Hate-speech-RoBERTa	0.7428	0.7484

speech underperformed, underlining the nuanced distinction between these concepts and aggression. The aggression detection model proposed by Kumari et al. (2021) achieves strong performance on the TRAC-1, our experiments reveal that its effectiveness does not directly carry over to our English TAG dataset.

5.2. What is the adaptability of the model with the TAG dataset?

We explore the effectiveness of transfer learning for aggression detection by fine-tuning pre-trained transformers on two datasets: TRAC 1 and TAG. Our goal is to evaluate model adaptability across domains and analyze performance differences when applied to unseen test data. Transfer learning involves fine-tuning pre-trained transformers on the source dataset (TRAC 1 or TAG) and evaluating their performance on the target dataset. As shown in Table 7, the mBERT model tuned on TAG obtains the best macro F1-score of 0.77 on the code-mixed TRAC 1 corpus. mBERT's success on TRAC 1 can be linked to its multilingual capabilities, effectively capturing language nuances present in code-mixed posts. Conversely, offensive Twitter pre-trained transformer architectures optimized on the TRAC-1 source dataset perform exceptionally well (0.76 macro-F1) on the Twitter-centric TAG test set, emphasizing the importance of domain-specific pre-training.

A comprehensive comparative analysis is performed to evaluate models in both directions: TRAC 1 to TAG (TRAC 1→TAG) and TAG to TRAC 1 (TAG→TRAC 1). Notably, the transfer from TAG to TRAC 1 outperforms the reverse, indicating the adaptability of models fine-tuned on TAG data to the TRAC 1 domain. This suggests that models trained on TAG data acquire robustness to syntactic variances, allowing better adaptation. However, the challenges arise when attempting to merge the TRAC 1 and TAG datasets for combined fine-tuning. This approach results

Table 7
Performance evaluation of fine-tuned transformer models in transfer learning between TRAC 1 and TAG datasets that highlights model adaptability across domains. In particular, the mBERT and RoBERTa offensive Twitter models exhibit prominent results.

Models	TRAC 1→TAG	TAG→TRAC 1
$BERT_{base}$	0.7041	0.7017
$BERT_{large}$	0.7082	0.7412
$M\text{-}BERT_{base}$	0.6670	**0.7728**
$M\text{-}distillBERT_{base}$	0.5546	0.7263
$DistillBERT_{base}$	0.6739	0.7085
$DistillBERT_{base-squad}$	0.6842	0.7103
Hing-BERT	0.6850	0.7385
Hing-mBERT	0.6727	0.7471
Hing-RoBERTa	0.7262	0.7221
$RoBERTa_{base}$	0.6670	0.7110
$RoBERTa_{large}$	0.7350	0.7109
$Xlm\text{-}RoBERTa_{base}$	0.6424	0.6512
$Xlm\text{-}RoBERTa_{large}$	0.6677	0.6867
$Xlnet_{base}$	0.7365	0.7261
$Offensive\text{-}RoBERTa_{base}$	**0.7633**	0.7443
Hate-speech-RoBERTa	0.7428	0.7400

in diminished performance compared to individual dataset fine-tuning, primarily due to inherent linguistic and contextual differences between Facebook (TRAC 1) and Twitter (TAG) posts. The fusion of these disparate datasets risks conflicting linguistic patterns, hindering model generalization. Combining datasets will provide more training data, but risks the degrading of specialized knowledge that is important for generalization (Ruder et al., 2019).

5.3. Why not LLM?

In the field of NLP, language model (LLM) advancements have been leveraged for diverse downstream tasks, including text classification (Sun et al., 2023). Our study explores the applicability of the widely utilized LLM, ChatGPT 3.5, for aggression detection in text. We formulated a specific prompt for zero-shot learning, directing the model to analyze input tweets for aggressiveness:

Tweet Analysis for Aggressiveness Detection: Consider the following tweet for analysis to determine its level of aggressiveness: **Input:** <tweet>

Classify the aggressiveness of the tweet into one of the following categories:

1. **Aggressive**
2. **Non-aggressive**

We evaluated 500 test samples from the TAG dataset and achieved an average macro F1 score of **0.63768** compared to human annotations for detecting aggression. However, it's crucial to recognize that Language Models (LLMs) excel at zero-shot learning but might not perform optimally in specific tasks like aggression detection in text.

6. Conclusion

This study showed exposure to aggressive feeds significantly increases individual user aggression based on 14M posts from 63K user. We introduced an aggression intensity metric to quantify overall user aggression levels. Notably, 90% of users exhibited positive correlation between their feed and user aggression intensity. Findings also indicates that event feeds can significantly impact the level of aggressiveness in users' posts. Moreover, users tended to encourage aggression by supporting such content. Our methodology utilized the proposed Twitter Aggression Dataset (TAG) and fine-tuned RoBERTa-large aggression detection model which provide benchmark performance. However, as the dataset comprises only English Twitter, findings may not generalize broadly. Future work could expand multilingually and incorporate social graphs and linguistic context into the aggression metric.

CRediT authorship contribution statement

Swapnil Mane: Conceptualization of this study, Methodology, Data Curation, Formal analysis, Writing – Original Draft. **Suman Kundu:** Supervision, Conceptualization, Methodology, Validation, Writing - Review & Editing, Visualization. **Rajesh Sharma:** Supervision, Conceptualization, Methodology, Validation, Writing - Review & Editing.

Acknowledgement

This research was supported by the Prime Minister Research Fellowship funded by the Ministry of Education (MOE), India. Rajesh is supported by EU H2020 program under the SoBigData++ project (grant agreement No. 871042), by the CHIST-ERA grant No. CHIST-ERA-19-XAI-010 (ETAg grant No. SLTAT21096), and partially funded by CHIST-ERA project HAMISON. Ketevan Kvirikashvili and Suman Karan have supported the data annotation task.

References

Adinugroho, I., Kristiani, P., Nurrachman, N., 2022. Understanding aggression in digital environment: Relationship between shame and guilt and cyber aggression in online social network. Makara Human Behavior Studies in Asia 26, 105–113.

Agbaje, M., Afolabi, O., 2022. Neural network-based cyber-bullying and cyber-aggression detection using twitter text .

Ali, M., Hassan, M., Kifayat, K., Kim, J.Y., Hakak, S., Khan, M.K., 2023. Social media content classification and community detection using deep learning and graph analytics. Technological Forecasting and Social Change .

Allen, J.J., Anderson, C.A., Bushman, B.J., 2018. The general aggression model. Current opinion in psychology 19, 75–80.

Aroyehun, S.T., Gelbukh, A., 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 90–97.

Arroyo-Fernández, I., Forest, D., Torres-Moreno, J.M., Carrasco-Ruiz, M., Legeleux, T., Joannette, K., 2018. Cyberbullying detection task: the ebsi-lia-unam system (elu) at coling'18 trac-1, in: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp. 140–149.

Bakshy, E., Messing, S., Adamic, L.A., 2015. Exposure to ideologically diverse news and opinion on facebook. Science 348, 1130–1132.

Balci, K., Salah, A.A., 2015. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. Computers in Human Behavior 53, 517–526.

Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L., 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint arXiv:2010.12421 .

Bertin, A., Durrani, N., Pham, M., de Marneffe, M.C., Auli, T., 2019a. Cross-lingual language models for cloz test fine-tuning, in: EMNLP-IJCNLP 2019 Workshop on Multilingual and Cross-Lingual Processing.

Bertin, A., Pham, M., de Marneffe, M.C., Auli, T., 2019b. Multilingual bert: Harnessing multilingual language models for task-specific fine-tuning, in: ACL 2019.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., Ojha, A.K., 2020. Developing a multilingual annotated corpus of misogyny and aggression. arXiv preprint arXiv:2003.07428 .

Boadi, C., Kolog, E.A., 2021. Social media aggression: An assessment based on the contemporary deterrence theory. Americas Conference on Information Systems .

Bruns, A., Highfield, T., Lewis, S., 2017. Twitter and public communication: A microblogging platform for social movements?, in: Social movements and their technologies. Springer, Cham, pp. 22–44.

Cairns, R.B., Cairns, B.D., Neckerman, H.J., Gest, S.D., Gariepy, J.L., 1988. Social networks and aggressive behavior: Peer support or peer rejection? Developmental psychology 24, 815.

Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E.D., Stringhini, G., Vakali, A., 2017. Mean birds: Detecting aggression and bullying on twitter. Proceedings of the 2017 ACM on Web Science Conference .

Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E.D., Stringhini, G., Vakali, A., Kourtellis, N., 2019. Detecting cyberbullying and cyber-aggression in social media. ACM Transactions on the Web (TWEB) 13, 1–51.

Chen, J., Yan, S., Wong, K.C., 2020. Verbal aggression detection on twitter comments: convolutional neural network for short-text sentiment analysis. Neural Computing and Applications 32, 10809–10818.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. Noise reduction in speech processing , 1–4.

Datta, A., Si, S., Chakraborty, U., Naskar, S.K., 2020. Spyder: Aggression detection on multilingual tweets, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 87–92.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Eraslan, L., Kukuoglu, A., 2019. Social relations in virtual world and social media aggression. World Journal on Educational Technology: Current Issues 11, 1–11.

Fisher, R.A., 1992. Statistical methods for research workers. Springer.

Galyashina, E.I., Nikishin, V.D., 2022. Fake media products as speech aggression provokers in network communication. European Proceedings of Social and Behavioural Sciences , 205–211doi:10.15405/EPSBS.2022.03.26.

Gattulli, V., Impedovo, D., Pirlo, G., Sarcinella, L., 2022. Cyber aggression and cyberbullying identification on social networks, in: ICPRAM, Scitepress. pp. 644–651. doi:10.5220/0010877600003122.

Gordeev, D., 2016. Automatic detection of verbal aggression for russian and american imageboards. Procedia-Social and Behavioral Sciences 236, 71–75.

Grootendorst, M., 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 .

Gutiérrez-Esparza, G.O., Vallejo-Allende, M., Hernández-Torruco, J., 2019. Classification of cyber-aggression cases applying machine learning. Applied Sciences 9, 1828.

Hardy, A., 1994. An examination of procedures for determining the number of clusters in a data set, in: New approaches in classification and data analysis. Springer, pp. 178–185.

Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Communication methods and measures 1, 77–89.

Henneberger, A.K., Coffman, D.L., Gest, S.D., 2017. The effect of having aggressive friends on aggressive behavior in childhood: Using propensity scores to strengthen causal inference. Social Development 26, 295–309.

Hinduja, S., Patchin, J.W., 2010. Bullying, cyberbullying, and suicide. Archives of suicide research 14, 206–221.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM computing surveys (CSUR) 31, 264–323.

Karan, S., Kundu, S., 2023. Cyberbully: Aggressive tweets, bully and bully target profiling from multilingual indian tweets, in: International Conference on Pattern Recognition and Machine Intelligence, Springer. pp. 638–645.

Khan, U., Khan, S., Rizwan, A., Atteia, G., Jamjoom, M.M., Samee, N.A., 2022. Aggression detection in social media from textual data using deep learning models. Applied Sciences 12, 5083.

Khandelwal, A., Kumar, N., 2020. A unified system for aggression identification in english code-mixed and uni-lingual texts, in: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp. 55–64.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M., 2018a. Benchmarking aggression identification in social media, in: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp. 1–11.

Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M., 2020. Evaluating aggression identification in social media, in: Proceedings of the second workshop on trolling, aggression and cyberbullying, pp. 1–5.

Kumar, R., Ratan, S., Singh, S., Nandi, E., Devi, L.N., Bhagat, A., Dawer, Y., Lahiri, B., Bansal, A., Ojha, A.K., 2022. The ComMA dataset v0.2: Annotating aggression and bias in multilingual social media discourse, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 4149–4161. URL: https://aclanthology.org/2022.lrec-1.441.

Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T., 2018b. Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402 .

Kumari, K., Singh, J.P., 2022. Multi-modal cyber-aggression detection with feature optimization by firefly algorithm. Multimedia systems 28, 1951–1962.

Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P., 2021. Bilingual cyber-aggression detection on social media using lstm autoencoder. Soft Computing 25, 8999–9012.

Kwak, H., Lee, C., Park, H., Moon, S., 2010. What is twitter, a social network or a news media?, in: Proceedings of the 19th international conference on World wide web, pp. 591–600.

Liu, M., Xue, J., Zhao, N., Wang, X., Jiao, D., Zhu, T., 2021. Using social media to explore the consequences of domestic violence on mental health. Journal of interpersonal violence 36, NP1965–1985NP.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

MacQueen, J., 1967. Classification and analysis of multivariate observations, in: 5th Berkeley Symp. Math. Statist. Probability, pp. 281–297.

Mane, S., Kundu, S., Sharma, R., 2023. A survey on online user aggression: Content detection and behavioural analysis on social media platforms. arXiv preprint arXiv:2311.09367 .

Mishna, F., Regehr, C., Lacombe-Duncan, A., Daciuk, J., Fearing, G., Van Wert, M., 2018. Social media, cyber-aggression and student mental health on a university campus. Journal of mental health 27, 222–229.

Oravec, J.A., 2023. Rage against robots: Emotional and motivational dimensions of anti-robot attacks, robot sabotage, and robot bullying. Technological Forecasting and Social Change .

Pareek, K., Choudhary, A., Tripathi, A., Mishra, K., Mittal, N., 2022. Hate and aggression detection in social media over hindi english language. International Journal of Software Science and Computational Intelligence (IJSSCI) 14, 1–20.

Pascual-Ferrá, P., Alperstein, N., Barnett, D.J., Rimal, R.N., 2021. Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the covid-19 pandemic. Big Data & Society 8, 20539517211023533.

Ramiandrisoa, F., 2022. Multi-task learning for hate speech and aggression detection. Joint Conference of the Information Retrieval Communities in Europe .

Ravindran, Y., Joshi, R., 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models, in: LREC 2022 Workshops.

Rawat, A., Nafis, N., Bhadane, D., Kanojia, D., Murthy, R., 2023. Modelling political aggression on social media platforms, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pp. 497–510.

Risch, J., Krestel, R., 2020. Bagging bert models for robust aggression identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 55–61.

Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T., 2019. Transfer learning in natural language processing, in: Sarkar, A., Strube, M. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, Association for Computational Linguistics, Minneapolis, Minnesota. pp. 15–18. URL: https://aclanthology.org/N19-5004, doi:10.18653/v1/N19-5004.

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., On, B.W., 2021. Aggression detection through deep neural model on twitter. Future Generation Computer Systems 114, 120–129.

Samghabadi, N.S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., Solorio, T., 2020. Aggression and misogyny detection using bert: A multi-task approach, in: Proceedings of the second workshop on trolling, aggression and cyberbullying, pp. 126–131.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 .

Sengupta, A., Bhattacharjee, S.K., Akhtar, M.S., Chakraborty, T., 2022. Does aggression lead to hate? detecting and reasoning offensive traits in hinglish code-mixed texts. Neurocomputing 488, 598–617.

Sharif, O., Hoque, M.M., 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. Neurocomputing 490, 462–481.

Shrivastava, A., Pupale, R., Singh, P., 2021. Enhancing aggression detection using gpt-2 based data balancing technique, in: 2021 5th International Conference on intelligent computing and control systems (ICICCS), IEEE. pp. 1345–1350.

Srivastava, S., Khurana, P., 2019. Detecting aggression and toxicity using a multi dimension capsule network, Association for Computational Linguistics (ACL). pp. 157–162. URL: https://aclanthology.org/W19-3517, doi:10.18653/V1/W19-3517.

Student, 1908. The probable error of a mean. Biometrika 6, 1–25.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G., 2023. Text classification via large language models. arXiv preprint arXiv:2305.08377 .

Tawalbeh, S., Hammad, M., Mohammad, A.S., 2020. Saja at trac 2020 shared task: Transfer learning for aggressive identification with xgboost, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 99–105.

Terizi, C., Chatzakou, D., Pitoura, E., Tsaparas, P., Kourtellis, N., 2021. Modeling aggression propagation on social media. Online Social Networks and Media 24, 100137.

Torregrosa, J., D'Antonio-Maceiras, S., Villar-Rodríguez, G., Hussain, A., Cambria, E., Camacho, D., 2022. A mixed approach for aggressive political discourse analysis on twitter. Cognitive computation , 1–26.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Vidgen, B., Thrush, T., Waseem, Z., Kiela, D., 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection, in: ACL.

Vladimirou, D., House, J., Kádár, D.Z., 2021. Aggressive complaining on social media: the case of# muckymerton. Journal of Pragmatics 177, 51–64.

Wang, Y., Han, R., Lehman, T.S., Lv, Q., Mishra, S., 2022. Do twitter users change their behavior after exposure to misinformation? an in-depth analysis. Social Network Analysis and Mining 12, 1–16.

Wong, N., Yanagida, T., Spiel, C., Graf, D., 2022. The association between appetitive aggression and social media addiction mediated by cyberbullying: the moderating role of inclusive norms. International journal of environmental research and public health 19, 9956.

Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J., 2011. Who says what to whom on twitter, in: Proceedings of the 20th international conference on World wide web, pp. 705–714.

Yang, Z., Dai, Z., Yang, Y., Sun, X., Zhou, Q., Huang, Z., 2019. Xlnet: Generalized autoregressive pretraining for language understanding, in: NeurIPS 2019.

Zhang, Y., Moe, W.W., Schweidel, D.A., 2017. Modeling the role of message content and influencers in social media rebroadcasting. International Journal of Research in Marketing 34, 100–119.

SWAPNIL MANE is a Ph.D. Research Scholar in the Department of Computer Science and Engineering at the Indian Institute of Technology (IIT), Jodhpur. He received his B.Tech. degree in Computer Science and Engineering from the Rajarambapu Institute of Technology, Islampur, India, in 2019, and his M.Tech. degree in Computer Engineering from the College of Engineering Pune, India, in 2021. Swapnil also has three months of industrial experience as a Data Scientist Engineer with Keydabra Inc. in Atlanta, U.S. He was awarded an Indian prestigious Prime Minister Research Fellowship in 2022 for his exceptional research work. His research interests lie in the areas of Natural Language Processing, Social Network Analysis, Computational Social Science, Network Data Science, and Knowledge Graphs.

SUMAN KUNDU received the B.Tech. degree in information technology from the West Bengal University of Technology, Kolkata, India, in 2005, and the M.E. degree in software engineering from Jadavpur University, in 2009. His Ph.D. Research was with the Center for Soft Computing Research, Indian Statistical Institute, from 2010 to 2015. He visited the Engine Group for the Postdoctoral Research, Wroclaw University of Science and Technology, from June 2018 to April 2019. He has more than six years of industrial software development experience with ZINFI Software Systems Private Ltd., Kolkata. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur. He has published articles in social network analysis, granular computing, and soft computing. His research interests includes social network analysis, network data science, soft computing, crowd sourcing, fuzzy and rough set, and granular computing.

RAJESH SHARMA is presently working as associate professor and leads the computational social science group at the Institute of Computer Science at the University of Tartu, Estonia, since January 2021. Rajesh joined the University of Tartu in August 2017 and worked as a senior researcher (equivalent to Associate Professor) till December 2020. From Jan 2014 to July 2017, he has held Research Fellow and Postdoc positions at the University of Bristol, Queen's University, Belfast, UK and the University of Bologna, Italy. Prior to that, he completed his PhD from Nanyang Technological University, Singapore, in December 2013. He has also worked in the IT industry for about 2.5 years after completing his Master's from the Indian Institute of Technology (IIT), Roorkee, India. Rajesh's research interests lie in understanding users' behavior, especially using social media data. His group often applies techniques from AI, NLP, and most importantly, network science/social network analysis.