Volume 3, Issue , 2024
PP 165-175

International Journal of Futuristic Innovation in Arts, Humanities and Management (IJFIAHM)

ANALYSIS OF CYBER WARFARE ON SOCIAL MEDIA

1st **Mr. Awadh Sao**
awadh@duck.com
Student
Dept. of CSE
Shri Sai College of Engineering and Technology
Chandrapur, India

2nd **Mr. Lowlesh Yadav**
lowlesh.yadav@gmail.com
Assistant Professor
Dept. of CSE
Shri Sai College of Engineering and Technology
Chandrapur, India

3rd **Mrs. Pushpa Tandekar**
p.tandekar@yahoo.in
Assistant Professor
Dept. of CSE
Shri Sai College of Engineering and Technology
Chandrapur, India

ABSTRACT

Antisocial behavior becomes more common because many online platforms encourage people to interact with one another. Lately, there has been a significant increase in aggressive behavior on social media, leading to various negative effects, including mental health issues and controversies. To address this, I conducted the very first analysis of aggressive user behavior on Twitter, a microblogging platform that doesn't have strict rules against aggressive behavior.

This analysis process involves three main steps: first, I collect data from Twitter; then, I identify instances of aggression in user interactions; finally, I create profiles of users based on their online behavior. In this study, I took a close look at how users exhibit aggressive behavior by examining their aggressive posts and the events they engage in. Interestingly, our findings show that users tend to be more engaged with aggressive content on the platform. This research sheds light on the relationship between user behavior and the prevalence of aggressive posts on Twitter.

Keywords: Aggression Detection, Aggression Behavior Analysis, Online Social Platform.

INTRODUCTION:

Social media is now an integral part of our lives, serving as a primary source of global information. The content shared on social platforms can have a profound impact on individuals' personal lives. As technology continues to advance, the popularity of social media continues to surge, raising concerns about the potential dissemination of harmful messages or information.

Social media is best described as a vast virtual space where users can access a wealth of information spanning various topics, engage in discussions and debates, or seek assistance

[1]. Social networks, also known as Social Network Sites (SNS), have significantly simplified the lives of people across various professions. These platforms have provided numerous benefits, including the ease of sharing information, facilitating communication, and supporting educational endeavors. Notably, social media platforms have piqued the interest of scientists due to the wealth of data they offer, which can be harnessed for diverse research purposes, enabling the prediction of various outcomes through effective data utilization.

However, the unrestricted ability to express one's opinions on social media has given rise to several challenges. With minimal oversight, virtually anyone with internet access can publish or post content as they see fit, leading to issues such as the expression of aggression. In many cases, individuals infringe upon the freedoms and fundamental rights of others by concealing their identity and creating fake accounts to vent their aggression towards others. Furthermore, automated social network accounts, oftenreferred to as "bots" and possessing characteristics similar to genuine user profiles, exert a substantial influence on individuals [2]. These bots can incite aggression by repeatedly posting the same content, misleading users, or subjecting them to psychological pressures. Notably, aggression on social media is not confined to the general public; even politicians engage in aggressive behavior, and their words and the actions of their supporters can pose societal threats, particularly when political candidates employ platforms like Twitter for political discourse.

The detection of aggression has gained significant attention in recent years, emerging as a crucial research area within natural language processing. Numerous researchers have dedicated their efforts to identifying and addressing harmful posts or comments on social media, striving to develop methods for detecting abusive behaviors such as cyberbullying, hate speech, and aggressive attitudes towards individuals [3]. However, it's important to note that the concept of aggression can be understood in various ways by different researchers and philosophers. There isn't a single, universally accepted definition. Some researchers, argue that aggression may not always be negative and harmful but can also serve as a form of defensive behavior. Sigmund Freud and Konrad Lorenz viewed aggression as an inherent genetic impulse rather than a negative act. Moreover, a study by [4] indicates that aggression may not always be overt and can be concealed. Researchers categorize it into three classes:

Overtly Aggressive, Covertly Aggressive, and Non-aggressive. Identifying and differentiating these categories is not always a straightforward task [4].

METHODOLOGY OF DETECTION:

Researchers have tried different ways to find mean stuff online over the years, even making websites to check for it. They used fancy methods like BERT, deep learning, and other things. One group came up with a way to find mean things on Twitter in real-time. They used computer tricks like Hoeffding Trees and stuff [5]. Some folks looked into lots of ways to find mean behavior, and one called Naive Bayes did really well, with 92% accuracy and 95% recall. Another bunch used a similar trick called Random Forest to spot hate speech.
Finding mean stuff online can be tricky, especially in different languages. Some words that are nice in one language are mean in another. One group made a computer program to find mean comments, and it worked in both English and Hindi. They also checked English, Hindi, and Bangla and found that Bert was good at finding mean words [6].
Some social media places, like Instagram, can't stop mean comments, so one group used Naïve Bayes to sort out the comments there.

Another bunch collected a bunch of tweets and used network, user, and text tricks to see if they were mean. They found that network tricks worked best for spotting mean stuff.
Another study used deep learning and BERT to find sarcasm on social media. They looked at Twitter and Reddit and saw that BERT did better, even on small or messy datasets.
Sometimes, it's hard for computers to catch sarcasm in tweets. One team used deep learning and different methods to find sarcasm in a mix of Hindi and English tweets, and the Bi-directional LSTM did the best, with 78.49% accuracy. Another study found that logistic regression was great for finding sarcasm with deep learning tricks [7].

DATASET:

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.
Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a

paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. DATA COLLECTION

I gathered numerous aggressive tweets from January 1, 2022, to July 15, 2022. During this time frame, several notable events occurred that generated a significant number of aggressive tweets related to these specific events.

B. EVENTS DISCUSSED

- Young people hit the streets to show they're not happy with the Government's new Agneepath scheme. This caused big and angry protests all over India, especially in Bihar and Uttar Pradesh. The Agneepath scheme is about hiring soldiers for the Army, Navy, and Air Force on a short-term basis to control salary and pension costs. Possible hashtags: #agneepathprotest, #agneepathyojana, #agnipath, #AgnipathScheme, #agnipathschemeprotest, #agnipathschemeprotests, #AgnipathProtests, #AgnipathProtest.

- A statement from a political leader named Nupur Sharma caused arguments between Hindus and Muslims, big protests, and fights.Possible Hashtags: #HindusUnderAttack, #NupurSharma, #NupurSharmaControversy, #KanhaiyaLal, #MuslimsUnderAttackinIndia.

- Many Kashmiri Pandits left the mostly Muslim area of Jammu and Kashmir because Kashmiri Hindus were getting killed and hurt a lot.Possible Hashtags: #KashmirAgainstTerrorism, #StopPakSponsoredTerrorism, #KashmiriPandits, #AakhirKabTak, #kashmirihindus.

C. ANNOTATION

This part talks about how I marked social media messages. I looked at Twitter posts in English and put them in two categories: aggressive and non-aggressive. The people marking the posts thought that aggression on social media could be both direct and indirect. They meant being aggressive, not just by using mean words, but also by using nice words or not swearing. Aggression could be aimed at different things, like threats of violence, sexual threats, or threats based on things like gender, where

someone is from, their politics, ethnicity, community, or race. Sometimes, aggression and abuse went together, but I only marked it as aggression when it was more than just friendly teasing. I followed what others had done in their work.

To make sure the marking was good, I had four people do it. They were students studying computer science and engineering, and there were two guys who were undergrads and one guy and one girl who were postgrads. Each person marked the same set of messages, and if they weren't sure about a message, they marked it as NaN. I checked how well they agreed with each other, and I used something called "Feiss kappa" to do that. This Feiss kappa thing is a way to see how much people agree when there are many people making judgments. The scores go from 0 (no agreement) to 1 (perfect agreement). For our marked data, the Feiss kappa score was 0.7873, which is pretty good agreement.

I used these marked messages to teach and test our model for spotting aggression. The model worked just as well whether I used 5,000 or 6,000 marked messages, so I stopped at 6,000. Out of these, 2,602 were aggressive and 3,398 were non-aggressive.

AGGRESSION DETECTION:

I figured out how strong the aggression was in a user's behavior by looking at their posts during a specific time. I made a timeline for each user based on their aggressive actions. I used these aggression intensity scores to create profiles for users in a vector format.

A. DATA PRE-PROCESSING

Before using the data in our model, I prepared it. I got rid of punctuation marks, numbers, and URLs because they don't really matter. I also made all the English letters lowercase. I took out common English stop words, spaces, and new lines. Sometimes,tweets mention users, even several of them. While these names can help identify potentially vulnerable users, they don't really help us find aggression. So, I got rid of them for the aggression detection model. After that, I did word lemmatization, which changes inflected words to their original form and keeps things simple.

B. LSTM AND BIDIRECTIONAL LSTM

As I covered in section 1, aggression relies on the whole meaning of a sentence rather than specific words. To understand sentence context with long-term connections, I used a type of structure called LSTM (Long Short-Term Memory). Figure 1 illustrates the basic design of an LSTM cell. LSTM works kind of like how the human brain handles things it learned in the past. It uses its memory to hold onto the important parts of a sentence and discard the less important parts through the initial layer called the forget gate (ft). The forget gate uses a sigmoid function (σ(.)), and the output is 0 for things to forget and 1 for things to remember (Eq. 1).

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

The second layer, known as the input layer (it), updates the information held in memory from before (Eq. 2). The outcome of the forget gate is multiplied by the previous LSTM's cell state (Ct−1). This product is then combined with the result of the input gate's output, along with the outcome of the tanh function applied to the previous hidden state (ht−1), resulting in the cell state at timestamp t (Ct) (Eq. 4). After passing through the tanh function, the cell state (Ct) is modified by multiplying it with the output gate (Ot), which leads to the hidden state at timestamp t (ht). Consequently, the last layer, denoted as Crt, encompasses the amalgamation of previous and current cell states, representing the memory of the timestamp from 0 to t, and this information is carried forward to the next LSTM state

$$i_t = \sigma(W_i.[h_{t-1}, X_t] + b_i) \quad (2)$$

$$C_t^r = \tanh (W_c.[h_{t-1}, X_t] + b_{tc}) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t + C_t^r \quad (4)$$

$$o_t = \sigma(W_o.[h_{t-1}, X_t] + b_o) \quad (5)$$

$$h_t = \tanh (C_t) * o_t \quad (6)$$

Here, ht−1 stands for the prior output, and xt represents the current input of the LSTM unit. Wx and bx are the weights and biases for the corresponding layer state, denoted by x. In the standard LSTM, significant information is retained and passed from state 0 to state 1, then to state 2, and so on, making it a forward-oriented process. The conventional LSTM is typically used in this forward direction. However, LSTM can also work in reverse, memorizing important details as it moves from state n to state n-1,and further to state n-2, ultimately reaching state 0. By combining both forward and

backward LSTM methods, I created a bidirectional LSTM (BiLSTM). This BiLSTM comprehensively understands the context of a sentence. To detect aggression effectively, the BiLSTM learns the context of a sequence of words twice: once by moving forward and once by moving backward. This architecture involves multiple cell states and the acquisition of crucial features, which enhances the strength of our model.

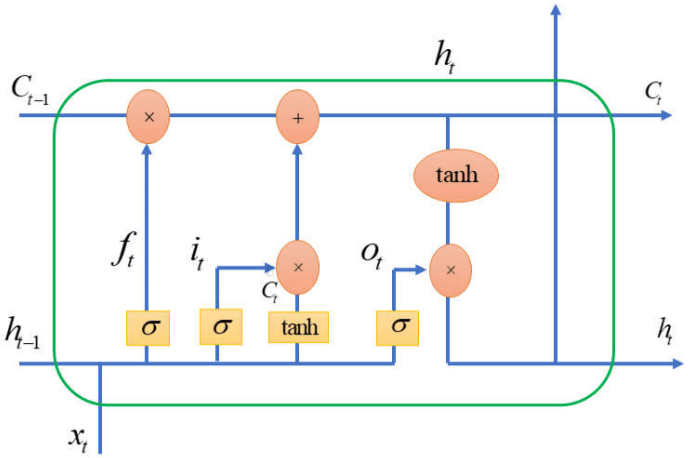

Figure 1: Performance Comparison with proposed methodology

I examined data broken down by time periods (such as weeks, days, or hours) to gauge how intense aggression was for specific users. The intensity of a user's aggression shows how much aggressive behavior they displayed within a set time frame. If a user has a high Aggressive Intensity score, it suggests they may have acted aggressively, while a low score indicates non-aggressive behavior. These Aggression Intensity scores range from 0 to 1, where 1 means the user showed the highest level of aggression, and 0 means they exhibited non-aggressive behavior.

To calculate the Aggression Intensity, I used posts that were labeled as either aggressive or non-aggressive for each user. The Aggression Intensity of user i during period l (AIl i) is determined by multiplying two factors: the user's aggregated aggressiveness score and the normalization score of their total posts. The user's aggressiveness score is a fraction of their total aggressive posts AGl i) over the total posts Xl i) they made in that period. The normalization score of total posts considers the difference between Xl i and the minimum number of posts within that period for all users (minl) and the difference between (minl) and the maximum number of posts within that period for all users (maxl).

DISCUSSION :

In recent years, technology has advanced significantly, and social media's popularity has surged, making it easier for people to share information. However, this rise in technology and social media usage has also led to an increase in the expression of aggression. Many researchers have delved into this issue, developing various models to detect aggression. Yet, identifying aggression on social media remains a complex and time-consuming task, posing numerous challenges for researchers. Surprisingly, not many of them have explored this problem from a behavioral perspective.

Our analysis in this paper reveals that individuals with a larger following tend to exhibit higher levels of aggression. This suggests that influential figures, as previously noted [8], may have a disproportionate impact on disseminating information or even inciting actions among their followers. Our study underscores the idea that the content in our feeds can influence our own level of aggression.

EVALUATION :

We did a special test using different combinations of features with our LSTM and BiLSTM models. The goal was to find an effective way to detect aggressive content. We used things like emotional cues and specific word meanings to do this. We also looked at how well our models performed using different metrics like accuracy and recall.

In our test, we found that the BiLSTM model using FastText word meanings performed the best out of all the models we tried. All of the BiLSTM models were more accurate than the LSTM ones. This tells us that the BiLSTM model is better at spotting aggression. What's interesting is that both the BiLSTM and LSTM models performed the best when using FastText word meanings, along with the emotional cues.

In conclusion, our experiment suggests that using the BiLSTM model with FastText word meanings is an effective way to detect aggression in tweets. This approach is better suited for understanding the context of tweets. We also have a graph that shows how our BiLSTM model with FastText performed during training, and it didn't underperform or overperform.

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 1. Performance comparison between the suggested model and the traditional, conventional model.

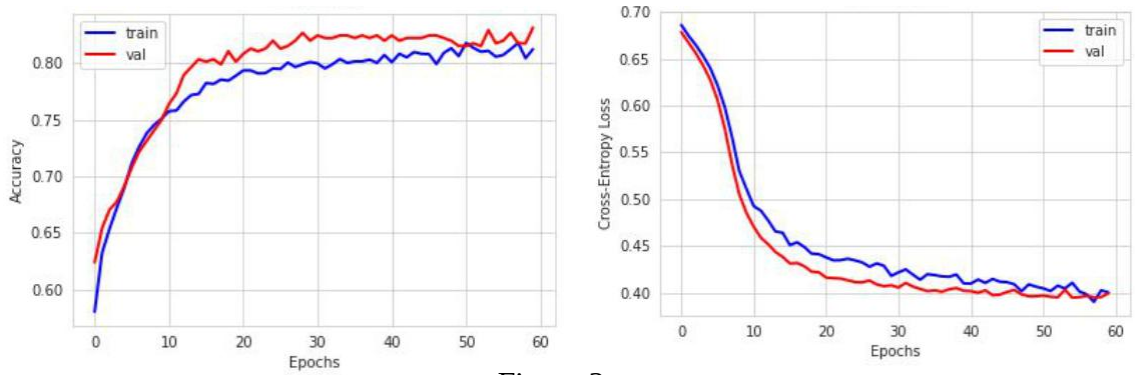

Figure 2.

a) Accuracy recorded at different training stages and during validation for the BiLSTM model with FastText embedding.

b) Loss recorded at different training stages and during validation for the BiLSTM model with FastText embedding.

CONCLUSION AND AND FUTURE WORK

In this research, I looked at how people act on social media based on what they post and what's happening. I created a model to find aggressive behavior, and it helped us understand

how users behave. I found that people are more involved with aggressive posts, and their actions depend on the events and content they see.

Our goal with this study was to help society by identifying aggressive users early and predicting their behavior to prevent problems. But there are some limitations to our analysis. I only looked at text on Twitter, so I didn't consider images or emojis. In the future, I can expand this study and develop a model that can detect aggression using various types of content.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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. IEEE Transl. J. Magn. Japan, vol. 2, pp.

[7] 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.

[8] 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.

[9] Lowlesh Nandkishor Yadav, "Predictive Acknowledgement using TRE System to reduce cost and Bandwidth" IJRECE VOL. 7 ISSUE 1 (JANUARY- MARCH 2019) pg no 275-278.