Detection and Analysis of Hate-Driven Violence on Social Networks

Yousef Abuhamda, Pedro García Teodoro
Network Engineering & Security Group
School of Computer Science and Telecommunication Engineering
University of Granada
18071 – Granada - Spain
yabuhamda@correo.ugr.es, pgteodor@ugr.es

Abstract - **In contemporary society, the widespread use of social networks such as Instagram, Facebook, TikTok and others facilitates quick access to information and services. Along with its benefits, social media carries negative effects, especially with regards to hate-driven violence. This phenomenon includes behaviors such as flaming, trolling, humiliating, masking, excluding, walking out, cyberbullying, bullying, and sexting, which pose significant concerns and have deep and lasting consequences, especially for vulnerable individuals such as children and adolescents. This paper deals with the detection of hate violence incidents by using some ML techniques. Moreover, a graphical analysis with Gephi is carried out for the datasets considered, concluding the necessity of getting better datasets for experimentation.**

Index Terms **– Hate-driven violence, Social media, Machine Learning, Visualization.**

I. Introduction

Today's society increasingly relies on new technologies and extensive use of social networks for quick access to information and services. Governments recognize the value of social networks for measuring public opinion on specific issues because of their effectiveness and ease of use. Some popular platforms include Instagram, Twitter, LinkedIn, YouTube, Pinterest, and Facebook [1]. Along with its benefits, the internet and social media also have adverse effects. Some works deal with them particularly those related to any violence or hate against persons.

Hate-driven violence (flaming, trolling, humiliating, masking, excluding, walking out, cyberbullying, bullying, and sexting) has become a significant concern with long-term consequences. Intentional violence, often targeting vulnerable individuals, leads to low self-esteem, depression, anxiety, learning difficulties, and even suicide. This issue affects many people worldwide, especially children and adolescents, and has lasting psychological effects into adulthood [2][3].

This way, hate-driven violence has emerged as a significant problem, it is affecting around 24% of teens who regularly use social networks. It tends to be an anonymous form of bullying through false information, making it difficult to identify perpetrators compared to traditional offline bullying [4].

This research makes several significant contributions. Firstly, introduces machine learning techniques aimed at enhancing the effectiveness of current detection mechanisms. Secondly, it provides a comprehensive review of existing literature, identifying key limitations, and proposing innovative solutions to address these challenges. Thirdly, the research presents empirical findings from the application of machine learning models to real-world social network datasets, offering insights into the efficacy of different detection approaches. Additionally, the study conducts graphical analysis using tools such as Gephi to visualize and interpret complex network structures, thereby deepening our understanding of hate-driven violence dynamics online. Finally, the research emphasizes the importance of obtaining better datasets for experimentation and future research endeavors, advocating for a comprehensive approach to combating hate-driven violence on social media platforms.

In that overall context, Section II conducts a thorough literature review, examining existing research on hate-driven violence detection and identifying key limitations and challenges for new contributions. Moving to Section III, we delve into the significance of social network datasets for experimentation, presenting selected datasets. Then, Section IV showcases the detection results achieved through the application of various machine learning models to these datasets, as well as the analysis. Finally, Section V succinctly concludes the paper, summarizing findings and suggesting future research directions.

II. Literature Review

Given the far-reaching consequences of hate-driven violence, like cyberbullying, it becomes imperative to foster the development of effective detection mechanisms for this menace. Various social networks now provide distinct automatic tools that empower users to control who can comment, view posts, or automatically establish connections.

Li *et al.* examine in [5] the effects of Internet use and cyberbullying on the psychological and behavioral well-being of Chinese adolescents. The study was conducted among 3378 middle school students aged 11–16 in different regions of China. The main findings showed that excessive Internet use (more than 3 hours per day) was associated with an increased risk of anxiety, depression, and mental health issues, such as stomach pain. Children are more likely to engage in online gaming, the game decline had a positive impact on well-being. Cyberbullying was common, with 37.5% of students admitting to having been involved, with those who had been bullies and victims being the most vulnerable to psychological mental, and physical health problems.

Alsubait and Alfageh review in [6] existing cyberbullying detection research, highlighting various approaches to

identify and classify cyberbullying behaviors. Multinomial Nave Bayes (MNB), Complement Nave Bayes (CNB), and Linear Regression (LR) are the three machine learning models considered by the authors. Furthermore, they use two feature extraction methods: Count Vectorizer and Tfidf Vectorizer.

Mahar uses various machine-learning approaches for detecting cyberbullying, including SVM, CNN, LSTM, naive Bayes, and logistic regression. The author concludes that LSTM has the best results, so they implemented their final approach using this kind of neural network [3].

Dadvar and Kai enhance in [7] the effectiveness of hate-driven violence detection, particularly in cyberbullying, by employing a Convolutional Neural Network (CNN). A ConvNet is a type of artificial neural network that employs perceptron, a machine learning algorithm designed to analyze data, inspired by studies of the mammalian central nervous system. Their proposed system is based on deep learning, which typically comprises three layers: the input layer, the hidden layer, and the output layer.

Authors in [8] present a novel Deep Learning approach which addresses the challenge of accurately assessing the severity of cyberbullying-related depression. Their technique replaces the challenging process of evidence extraction and selection with word vectors that capture the underlying semantics of words, achieved through the use of CNN. This approach proves to be more effective in characterizing tweets compared to conventional grouping computations.

Haidar *et al*. introduce a multi-level cyberbullying detection system that incorporates established Machine Learning (ML) and Natural Language Processing (NLP) techniques [9]. Their project focuses on identifying cyberbullying content in various languages, including Arabic, English, and texts written in Arabic with Latin letters.

Authors in [10] introduce an innovative neural convolutional system that incorporates theoretical features for cyberbullying detection. This study compares CNN and various classification models using two datasets, each characterized by distinct levels of sensitivity and class balance. The Estand method displays elitist behavior on the provided datasets. To address the class imbalance, three different systems were implemented and assessed. The findings revealed that the PCNN with labor cost adjustment emerges as the most effective solution.

Abouzaude and Savage highlight the ever-evolving nature of cyberbullying studies in [11]. Despite the significant progress in the field over the last two decades, substantial knowledge gaps persist, encompassing areas like the motivations behind "self-cyberbullying", the complexities of the "bully-victim phenomenon", the roles played by bystanders, shifts in cyberbullying prevalence among college students and adults, and culture-specific aspects. Additionally, there's an urgent need to explore cyberbullying subtypes and establish effective management practices, including mental health services and school interventions, all within a rapidly changing digital environment, requiring a more nuanced understanding of cyberbullying and how to effectively address it.

Idrizi and Hamiti tackle the pressing issue of cyberbullying [12], which has recently gained significant cultural relevance, causing psychological and emotional distress to victims of cyberbullying bites harmful types of electronic harassment. The authors distribute a variety of media (text, images, and audio) posted on social media, aiming to identify cases of cyberbullying, graph convolutional neural network, mail scale -analysis using advanced techniques such as filter banks, speech spectrograms, and other shows that audio post-processing mainly MFCCs and graph convolutional neural networks, provides accuracy for identifying cyberbullying in text, image, and video content to classify instances of bullying.

While first-hand hate-driven violence research such as cyberbullying is still evolving, impressive progress has been made in recent years. In this context, there are significant gaps in the existing literature so further research is needed to provide students, teachers, and stakeholders with effective cyberbullying prevention strategies. Systems for early detection of any kind of hate-driven cyberbullying activities early on social networking platforms play an important role in reducing and mitigating the negative impact on victims.

As a result, schools and organizations must take a proactive approach to the problem, as traditional measures such as cyber espionage tools, web filtering, and mobile phone censorship have proven to be inadequate in the fight against hate-driven violence.

III. Research Methodology

A. *Social Network Datasets*

Social network datasets are pivotal in understanding hate-driven violence, providing insights into discriminatory behaviors and violent expressions online. In this context, their significance is underscored, particularly concerning platforms like Instagram. These datasets grant researchers access to real-time interactions, aiding in pattern identification and trend analysis. Diverse datasets are essential for a nuanced understanding across demographics and cultures. Thus, this introduction highlights both the importance of social network datasets in detecting hate-driven violence and the need for comprehensive data sources to support effective research in this critical field.

For our research, we have selected two datasets to analyze and compare the detection of hate-driven violence.

- Dataset from [13], This first dataset is distributed into six cyberbullying classes (Fig. 1 and Fig. 3), ‘Age’, ‘Ethnicity’, ‘Gender’, ‘Religion’, ‘Other’, and ‘NotCb (not cyberbullying)’.

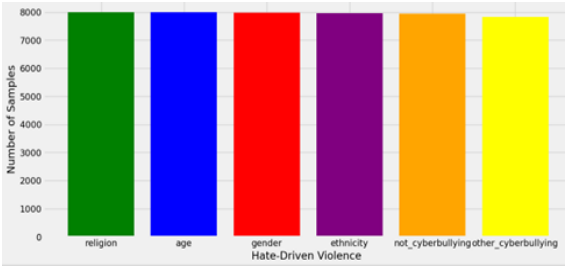

Fig. 1. Distribution of hate driven violence in the first dataset

tweet_text Text of the tweet	cyberbullying_type Type of cyberbullying harassment.
46017 unique values	religion 17% age 17% Other (31702) 66%
In other words #katandandre, your food was crapilicious! #mkr	not_cyberbullying

Fig. 3. Data classification

- Because Instagram's API access is heavily restricted, gathering data from this platform is quite difficult. As a result, researchers frequently turn to using data from Twitter instead, since its API access is more readily available. We obtained a new dataset consisting of real tweets from Kaggle (https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset), which has around 18000 rows. This dataset is distributed into two classes: class 1 represents instances containing violent words, while class 0 represents instances without violent words (Fig. 2).

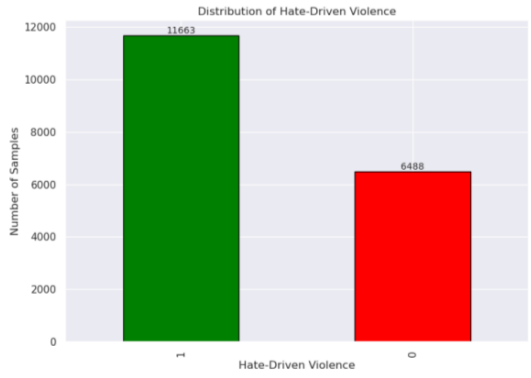

Fig. 2. Distribution of hate driven violence in the second dataset

B. Data Preprocessing

Before analyzing data, it is necessary to preprocess the information to:

- **Remove Patterns:** In this phase, specific patterns or substrings are removed from the text data to increase its purity and relevance for subsequent analysis This function targets features such as user no issue with other fixed observations that may introduce noise or distractions into the data set. By systematically identifying these patterns and using routines or similar techniques, the data structure is simplified and prepared for additional preprocessing steps without compromising its integrity hold or content is not corrupted.
- **Clean Text:** Data cleaning involves a series of tasks aimed at providing standardized and streamlined textual content to facilitate meaningful analysis. This phase often involves removing extra lines, punctuation marks, and other nonlinear markers that may interfere with comprehension or introduce bias in subsequent analysis Converting text to a consistent and uniform format reduces noise potential sources, making the dataset ideally suited for the construction of natural language processing techniques.
- **Tokenization:** This is the process of breaking up pure text into individual groups of tokens, usually words or subwords, to enable further analysis and processing This step involves partitioning text based on whitespace or alphanumeric characters boundaries to extract meaningful groups of information. Tokenization is an important preprocessing step in natural language processing tasks, and it provides a set of textual content that can be used for tasks such as feature extraction, sentiment analysis, and machine learning-based classification.
- **Lemmatization:** Lemmatization is a linguistic process aimed at reducing vocabulary to bases or elementary sets, known as lemmas, to generate appropriate vocabulary, and subsequently improve and interpret analyzes Lemmatizing text with language specificity rules use the generate to strengthen synonymous variables, reduce redundancy, and enhance the underlying natural language processing tasks.
- **Vectorization (TF-IDF):** Vectorization, specifically Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, is a crucial preprocessing step that transforms the cleaned and tokenized text data into numerical feature vectors suitable for machine learning algorithms. TF-IDF assigns weights to each term based on its frequency in a document relative to its occurrence across the entire corpus, thereby capturing the importance of terms in discriminating between documents. By representing textual data as TF-IDF vectors, complex linguistic information is encoded into a compact and interpretable format, facilitating the training and evaluation of machine learning models for tasks such as classification, clustering, and information retrieval.

At his point, the data set is divided into two distinct subsets: the training set and the test set. This separation is important for research on the efficiency and general applicability of machine learning models. A training set with most of the data is used to train models, allowing them to look for patterns and relationships in the data set in contrast, a test set isolated from training data is model independent used to evaluate the performance of models on unseen data.

C. Data Modelling

Various machine learning models are considered for hate-driven detection, which are implemented here using python machine learning packages. The models are chosen based on popularity, ease of use, training, and prediction time [14].

- **LinearSVC Model:** The Train LinearSVC Model phase involves fitting a Linear Support Vector Classification LinearSVC model to the training data, enabling it to identify patterns and relationships in the feature space and make predictions about new patterns. By optimizing linear decision constraints, LinearSVC achieves effective and efficient classification performance, especially at higher fractions. During training, the model adjusts its parameters to maximize the margin between instances of different classes to minimize the classification error. By training the LinearSVC model on the training data, we provide the knowledge needed to distinguish between

toxic and non-toxic read samples, enabling it to make more accurate predictions in real-world situations.

- **Random Forest Model:** Training the Random Forest Model uses training data to cluster decision trees and aggregate their predictions for collective decision-making Random Forest is a cluster learning method that builds multiple decision trees during training and combines their results for the prediction accuracy and strength improve. Each decision tree in the random forest is trained on a random subset of features and data samples, reducing the risk of overfitting, and increasing model generalizability. By training a Random Forest model on training data, we sub diversity and collective intelligence of multiple decision trees organized.
- **Logistic Regression Model:** Training a Logistic Regression Model estimates the parameters of a logistic regression function using training data to model the probability of a binary outcome (toxic or non-toxic) based on input features knowing the weights, thus probability able to predict, and instances are classified into appropriate classes. By training the Logistic Regression model on the training data, we obtain a well-calibrated classifier capable of estimating the likelihood of cyberbullying based on the extracted features, facilitating informed decision-making in cyberbullying detection scenarios.
- **KNN Model:** Training the KNN Model involves storing the entire training dataset and making predictions for new instances based on their proximity to the nearest neighbors in the feature space. KNN is a non-parametric classification algorithm that assigns the class label of most of its k nearest neighbors to a new instance, making it particularly suitable for locally smooth decision boundaries and diverse data distributions. During training, the KNN model memorizes the training instances and their corresponding class labels, enabling it to classify new instances based on their similarity to the existing data points. By training the KNN model on the training data, we create a flexible and adaptive classifier capable of accurately classifying cyberbullying instances based on their proximity to similar instances in the feature space.

D. *Graphical Data Analysis Tool*

In order to gain deeper insights into the data considered in our experimentation, we utilized a graphical tool known as Gephi (https://gephi.org/) is emerging as an important tool in the network analysis landscape, providing a versatile platform for visualizing and analyzing complex network structures. Through user-friendly interfaces for its robust features, Gephi further facilitates examination of social network datasets. Researchers are able to reveal complex patterns and relationships in digital environments through open source, and available for Windows, Linux, and even Mac.

IV. Experimental Results

Performance detection is assessed by counting True Negatives (TN), False Positives (FP), False Negatives (FN) and True Positives (TP). These four numbers can be represented as a confusion matrix. Different performance metrics are used to evaluate the performance of the constructed classifiers. In text classification, some common performance measurement functions are examined to determine the following metrics:

- **Precision:** Precision is also known as the positive predicted value. It is the proportion of predictive positives which are positive.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** Recall is the proportion of actual positives which are predicted positive.

$$Recall = \frac{TP}{TP + FN}$$

- **F-Measure:** F-Measure is the harmonic mean of precision and recall. The standard F-measure (F1) gives equal importance to precision and recall.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- **Accuracy:** Accuracy is the number of correctly classified instances (true positives and true negatives).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The specific detection results for the datasets considered are shown in Table II, Table I, Table IV, and Table III. They are as follows:

- First dataset:

Table II

CLASSIFICATION SUMMARY ALGORITHMS – FIRST DATASET

Algorithm	Accuracy	Precision	Recall	F1Score
LogisticRegression	80.46%	81.47%	80.46%	80.86%
RandomForest	79.63%	80.83%	79.63%	80.05%
LinearSVC	79.47%	79.66%	79.47%	79.47%
KNeighbors	30.50%	70.26%	30.50%	29.45%

Table I

CONFUSION MATRIX – FIRST DATASET

Algorithm	TN	FP	FN	TP
LogisticRegression	1473	2	628	1919
RandomForest	1491	0	602	1929
LinearSVC	1493	2	692	1835
KNeighbors	110	78	3126	1708

Linear Support Vector Classifier demonstrates a balanced performance with a relatively high accuracy of 79.47%. It achieves this by correctly predicting a substantial number of cyberbullying instances while minimizing false positives, as indicated by the low count of 2 in the false positive category. However, it still exhibits a notable number of false negatives 692, indicating instances of cyberbullying that were not identified by the model. The precision, recall, and F1 score are also around 79.5%, indicating a good balance between correctly classified instances and avoiding false positives.

RandomForest classifier achieves a commendable accuracy of 79.63%, performing slightly better than LinearSVC. Notably, it achieves this without any false positives, implying robustness in classification. However, it still faces a considerable number of false negatives 602, suggesting instances of cyberbullying that were missed by the model. Despite this, RandomForest maintains a good balance between precision, recall, and F1 score, with precision slightly higher than LinearSVC at 80.83%. This indicates a

high proportion of correctly identified cyberbullying instances out of the total predicted positives.

Logistic Regression emerges as the top-performing model among the evaluated algorithms, boasting the highest accuracy of 80.46%. It effectively predicts cyberbullying instances with a relatively low count of false positives 2 and false negatives 628. This model achieves the highest precision 81.47% among all algorithms, indicating a high proportion of correctly identified cyberbullying instances out of the total predicted positives. The recall and F1 score are also notably high, both around 80.5%, indicating a good balance between precision and recall.

KNeighbors classifier exhibits the lowest performance among all models, with an accuracy of only 30.50%. It struggles to effectively classify cyberbullying instances, as evidenced by the high count of false negatives 3126 and relatively low count of true positives 1708. The precision, recall, and F1 score are also notably lower compared to other models, further indicating its inadequacy for this task. The precision and recall are particularly low, at around 70% and 30%, respectively, leading to a low F1 score of 29.45%.

- Second dataset:

Table IV

CLASSIFICATION SUMMARY ALGORITHMS – SECOND DATASET

Algorithm	Accuracy	Precision	Recall	F1Score
LogisticRegression	92.87%	94.98%	93.71%	94.34%
RandomForest	92.81%	96.28%	92.23%	94.21%
LinearSVC	92.98%	95.03%	93.84%	94.43%
KNeighbors	49.55%	96.46%	21.27%	34.85%

Table III

CONFUSION MATRIX – SECOND DATASET

Algorithm	TN	FP	FN	TP
LogisticRegression	1213	114	145	2159
RandomForest	1245	82	179	2125
LinearSVC	1214	113	142	2162
KNeighbors	1309	18	1814	490

The Linear Support Vector Classifier exhibited strong performance with an accuracy of 92.98%. The confusion matrix reveals that out of 2631 test instances, it correctly classified 1214 instances as negative (True Negatives) and 2162 instances as positive (True Positives). However, it misclassified 113 instances as positive (False Positives) and 142 instances as negative (False Negatives). This model demonstrates a precision of 95.03%, indicating a high proportion of correctly predicted positive instances out of all instances predicted as positive. The recall of 93.84% signifies the model's ability to correctly identify positive instances out of all actual positive instances. The F1 score of 94.43% reflects a balanced performance between precision and recall.

The Random Forest Classifier achieved an accuracy of 92.81%. The confusion matrix indicates that out of 2631 test instances, it correctly classified 1245 instances as negative and 2125 instances as positive. However, it misclassified 82 instances as positive and 179 instances as negative. With a precision of 96.28%, it demonstrates a high proportion of correctly predicted positive instances out of all instances predicted as positive. The recall of 92.23% highlights the model's ability to correctly identify positive instances out of all actual positive instances. The F1 score of 94.21% reflects a balanced performance between precision and recall.

Logistic Regression performed well with an accuracy of 92.87%. The confusion matrix demonstrates that out of 2631 test instances, it correctly classified 1213 instances as negative and 2159 instances as positive. However, it misclassified 114 instances as positive and 145 instances as negative. With a precision of 94.98%, it shows a high proportion of correctly predicted positive instances out of all instances predicted as positive. The recall of 93.71% indicates the model's ability to correctly identify positive instances out of all actual positive instances. The F1 score of 94.34% reflects a balanced performance between precision and recall.

K-Nearest Neighbors achieved an accuracy of 49.55%, significantly lower than the other models. The confusion matrix reveals that out of 2631 test instances, it correctly classified 1309 instances as negative and 490 instances as positive. However, it misclassified 18 instances as positive and 1814 instances as negative. Despite having a high precision of 96.46%, its low recall of 21.27% highlights its inability to correctly identify positive instances out of all actual positive instances. The F1 score of 34.85% reflects the overall performance of the model, considering both precision and recall.

A. Graphical Data Analysis

Beyond the above results, Gephi analysis has revealed several significant points:

- **Firstly,** it's crucial to remember that node size correlates with entry degree, representing the number of retweets received. This ensures that profiles with greater impact are visually prominent.
- **Secondly,** the layout algorithm employed in our visualization operates by attracting nodes with more common connections while repelling those with fewer. Consequently, the distance between nodes becomes inversely proportional to their common connections (Fig. 4). Thus, nodes that are farther apart in the graph have fewer connections between them. In the context of segregating opinions into communities, this implies that distant nodes represent divergent opinions.

Fig. 4. Graph distributed by communities

The "comet tail," indicating numerous users who have retweeted but lack connections with each other (Fig. 5). This scenario corresponds to a clustering coefficient of C=0, where C represents the ratio of connections between a node's neighbors to the total possible connections.

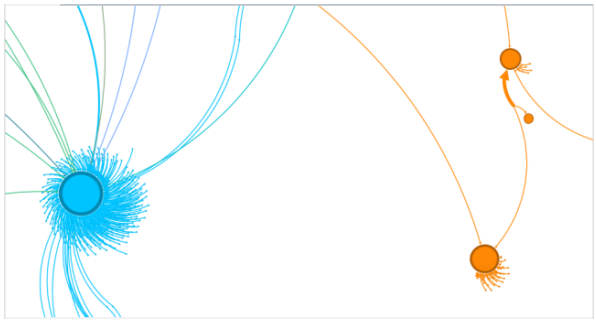

Fig. 5. Comet tails

When a user communicate content from multiple communities, the edges representing these interactions will be depicted in different colors (Fig. 6).

Fig. 6. Different color for multiple communities

When a community lacks a "dense" color and instead exhibits a connection structure characterized by several highly retweeted nodes linked together by sparse connections (Fig. 7), it indicates a group that is not deeply engaged in the analyzed topic.

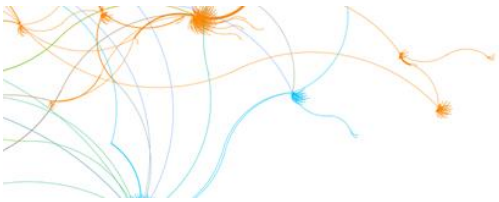

Fig. 7. Less densely connected communities

When profiles within a community frequently retweet each other, resulting in a higher clustering coefficient, densely connected areas emerge. These visually manifest as regions of "dense" color (Fig. 8).

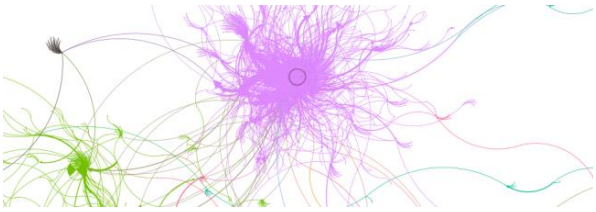

Fig. 8. Densely connected communities

After analyzing a dataset with Gephi, several key insights emerged about the network's structure and dynamics. The observation of divergent communities, the presence of comet tails indicating separate users, and highly interconnected communities reveal the presence of interconnections in the network. Despite examination, no clusters indicative of cyberbullying events was found. Instead, the clusters identified corresponded to various topics of discussion and communities within the network. This study not only confirms the absence of incidents of cyberbullying in the context studied, but also highlights the richness and complexity of current interactions.

However, to strengthen the validity and scope of our findings, acquiring new datasets becomes imperative. By incorporating additional data, such as diverse timeframes and social media platforms, we can deepen our insights, validate observed patterns, and enhance predictive capabilities, thereby ensuring a comprehensive understanding of the network dynamics.

V. Conclusions

This work deals with cyberbullying detection by applying machine learning algorithms to early detect hate-motivated violence. Despite the good results obtained, it is imperative to develop a larger dataset to comprehensively study people's violent behavior. Specifically, such a dataset should include.

Firstly, it should cover a broad spectrum of textual content, ranging from comments and posts to messages and captions, to capture the various manifestations of hate speech and abusive language prevalent on social networks. Additionally, incorporating demographic information such as age, gender, ethnicity, religion, and geographical location would facilitate analyses across different demographic groups and geographic regions. Behavioral patterns exhibited by both perpetrators and victims, including engagement frequency, posting behaviors, interaction networks, and content sharing activities, are essential components. Furthermore, integrating multimedia content like images, videos, and audio recordings would offer a more nuanced understanding of how hate speech is disseminated across different media formats. Temporal dynamics should also be considered, with timestamps for each interaction or post enabling the analysis of the evolution of hate-driven behaviors over time and identification of emerging trends.

Moreover, the stringent restrictions on Instagram's API access presented significant challenges in acquiring data from this platform. As a result, researchers often turn to utilizing data from Twitter due to its more accessible API. In our case, we acquired a new dataset comprising real tweets from Kaggle to supplement our analysis and visualization using the Gephi tool. However, this doesn't negate the necessity of developing alternative datasets.

Acknowledgement

This work has been developed within the "Plan de Recuperación, Transformación y Resiliencia", project C025/24 INCIBE-UGR, funded by the European Union (Next Generation).

References

[1] Global Social Media Statistics Research Summary 2023. Retrieved from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

[2] López-Vizcaíno M. F., Nóvoa F. J., Carneiro V., Cacheda F. (2021). Early detection of cyberbullying on social media networks. Volume 118, May 2021, Pages 219-229. DOI: 10.1016/j.future.2021.01.006

[3] Mahat M. (2021). Detecting Cyberbullying across Multiple Social Media Platforms Using Deep Learning. International Conference on

Advance Computing and Innovative Technologies in Engineering (ICACITE). DOI: 10.1109/ICACITE51222.2021.9404736

[4] Gorro K. D., Sabellano M. J. G., Maderazo C., Capao K. (2018). Classification of Cyberbullying in Facebook Using Selenium and SVM. In 2018 3rd International Conference on Computer and Communication Systems, ICCCS 2018 (pp. 233–238). DOI: 10.1109/CCOMS.2018.8463326

[5] Li J., Wu Y., Hesketh T. (2023). Internet Use and Cyberbullying: Impacts on Psychosocial and Psychosomatic Wellbeing among Chinese Adolescents. Volume 138, 107461. DOI: 10.1016/j.chb.2022.107461

[6] Alsubait T., Alfageh D. (2021). Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments. IJCSNS International Journal of Computer Science and Network Security, Volume 21, pages 1-5. Retrieved from http://paper.ijcsns.org/07_book/202101/20210101.pdf

[7] Dadvar M., Kai E. (2020). Cyberbullying Detection in Social Networks Using Deep Learning Based Models. In Big Data Analytics and Knowledge Discovery, pp. 245–255.DOI: 10.1007/978-3-030-59065-9_20

[8] Yin D., Xue Z., Hong L., Davison B. D., Kontostathis A., Edwards L. (2009). Detection of Harassment on Web 2.0. Proceedings of the Content Analysis in the WEB, 2(January), 1–7.

[9] Haidar B, Maroun C, Serhrouchni A. (2017). A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. Advances in Science Technology and Engineering Systems Journal 2(6):275-284. DOI: 10.25046/aj020634

[10] Zhang X., Tong J., Vishwamitra N., Whittaker E., Mazer J.P., Kowalski R., Hu H., Luo F., Macbeth J., Dillon E. (2016). Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. 15th IEEE International Conference on Machine Learning and Applications (ICMLA). DOI: 10.1109/ICMLA.2016.0132

[11] Aboujaoude E., Savage M. W. (2023). Cyberbullying: Next-Generation Research. World Psychiatry, Volume 22, Issue 1, February 2023, Pages 45-46. DOI: 10.1002/wps.21040

[12] Idrizi, E., & Hamiti, M. (2023). Classification of Text, Image, and Audio Messages Used for Cyberbullying on Social Media. In 2023 46th MIPRO ICT and Electronics Convention (MIPRO) (pp. 1-6). DOI: 10.23919/MIPRO57284.2023.10159835

[13] Wang J., Fu K., Lu C.T. (2020). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. Proceedings of the 2020 IEEE International Conference on Big Data. (pp. 1-10). DOI: 10.1109/BigData50022.2020.9378065. Dataset available at: [https://drive.google.com/drive/folders/1oB2fan6GVGG83Eog66Ad4wK2ZoOjwu3F?usp=sharing]

[14] Hadiya M. (2022). Cyberbulling Detection in Twitter using Machine Learning Alorithms. International Journal of Advances in Engineering and Management. Volume 4, Issue 8, pp. 1172-1184, 2022.