Cyberbullying Detection using LLMs and sentiment analysis

1st Khaddar Hela
Dept of computer engineering and mathematics
INSAT, Tunisia
Sherbrooke, Canada

2nd Dr Yasir Malik
dept of computer science
Bishop's university
Sherbrooke, Canada

3rd Dr Dorsaf Sebai
dept of computer engineering and mathematics
INSAT, Tunisia
Tunis, Tunisia

Abstract - Social media platforms like Facebook and Twitter have become integral to daily life, enabling users to easily share messages, images, and videos. While these platforms offer many benefits, they also present serious concerns, including the rise of harmful behavior such as cyberbullying. Cyberbullying manifests in various forms, from damaging images and videos to text-based harassment. This research focuses on the detection of cyberbullying within textual comments. Despite extensive efforts using techniques such as n-grams, convolutional neural networks (CNNs), gated recurrent units (GRUs), and long short-term memory networks (LSTMs), understanding the context of these comments remains challenging. In this study, we fine-tuned a pre-trained model, DistilBERT, by incorporating sentiment features and applied it to a cyberbullying dataset from Kaggle. We then evaluated our approach on an unseen dataset, consisting of text extracted from Twitter and Facebook. The model achieved an F1 score of 89.72% and an accuracy of 89.47% on the test set. However, when applied to the unseen dataset, which had a different class distribution, the model's performance dropped, achieving an accuracy of 64%. The model was notably more effective at detecting non-cyberbullying content, with an F1 score of 75%, compared to a 35% F1 score for cyberbullying content.

***Index Terms*—Cyberbullying Detection; Text classification; Feature extraction; Natural Language Processing (NLP); Machine Learning; Deep Learning; Sentiment Analysis; Transformers, Large Language models(LLMs); BERT Model**

I. Introduction

Social media platforms have become an integral part of our daily lives, connecting people across the globe and providing a wide array of services. Each type caters to different user needs and interests, attracting a diverse audience ranging from hobbyists and professionals to activists and entertainers. [1]

While social networking sites like Facebook and Twitter facilitate personal and professional connections, media sharing platforms such as YouTube, Instagram, and TikTok enable widespread multimedia sharing. However, the rapid increase in user-generated content has outpaced the platforms' ability to monitor and manage it effectively. This oversight has led to the spread of disinformation, increased polarization, and significant psychological harm. [2]

The challenge of identifying and removing offensive content is pressing. Platforms like YouTube and Facebook have implemented measures, such as YouTube's removal of inappropriate videos and Facebook's 'protective detection' algorithm, which flags posts indicating suicidal thoughts, self-harm, or hostility.

The objective of this research is to develop and evaluate a model for detecting cyberbullying that not only identifies harmful text but also takes its context and sentiment into consideration. We begin by analyzing various methods used in machine learning, deep learning, and transformer-based models. Our approach leverages sentiment features with DistilBERT to evaluate our method on a dataset from Twitter and Facebook.

The rest of the paper is organized as follows: Section 2 discusses the background of cyberbullying and its types. In Section 3, we review related advanced works on detecting cyberbullying. The proposed methodology is detailed in Section 4. Section 5 presents the results of our evaluation, and the final section outlines potential future work.

II. Background

A. Bullying and Cyberbullying

Bullying is longstanding violence, physical or psychological, conducted by an individual or a group directed against an individual or group who is not able to defend himself in the actual situation. [3] While, traditional bullying is limited to a specific time and place, cyberbullying can occur at anytime and from any location on earth.

Cyberbullying is *intentional* and *repeated harm* inflicted through digital devices [4]. In fact, it involves deliberate actions and patterns of behavior, rather than isolated incidents. The victim must feel the harm inflicted for it to be considered cyberbullying. Additionally, a single harmful message can be shared and seen by many, amplifying its impact even without repeated actions by the original bully.

It can occur through social media (Facebook, instagram, tiktok, snapchat), text messaging, in online forums, chat rooms, message boards, Email and even through online gaming communities [5]. It can include sending posts or sharing negative, harmful, false or mean content about an individual. (insults, threats, profane language and swear words.) It can be

targeted insult or untargeted (post containing general profanity) , sharing personal or private information about someone causing embarrassment or humiliation [2].

B. Types of Cyberbullying

There are many terms related to cyberbullying such as flaming, outing, rumor, fake news spreading, doxing, trolling... In this section, we describe these terms according to how they occur. [6]

- *Harassment:* is done by sending persistent and hurtful messages (threats, insults,..) repeatedly.
- *Doxing:* It occurs when someone maliciously shares personal data about an individual such as home address, social security member,..
- *Flaming:* is the act of posting about or mailing insults and vulgarity to a victim. A cyberbully fires a victim in the hopes of provoking an online brawl. It's the most severe one because if online fight occur between internet's users, it could be difficult to recognize the cyberbully and the victim.
- *Cyberstalking:* It's when a cyberbully monitors a victim's online presence closely, makes a false accusations and threats against the victim and their loved ones. It's the less severe because it's easy to detect the cyberbully.
- *Trolling:* is the act of posting derogatory comments about someone in the hopes of hurting these individuals.
- *Rumors:* It is an unverified claim which is made by users on social media platforms and can spread beyond their accounts. The methodology to work on rumors is first to do the detection, if a tweet is a rumor or not then do the rumor veracity detection which is to see if it's a true, false, unverified rumors [7].
- *Fake news:* is a form of bullying on a larger scale where many people are involved. Techniques include creating fake profiles, posting false content, creating destructive websites about others, or using VPNs/servers located outside the country to protect their identity, making tracking difficult. [8] The goal of fake news is to manipulate the minds and actions of users, making everyone a potential victim.

C. Features of cyberbullying

This section specifically focuses on the features that can be incorporated in in detecting cyberbullying. It has five dimensional feature set cited in this paper [9], which are the user-based, content-based, network-based , episode-based and others dimensions.

- *User-based:* This dimension focus on the characteristics of individual users, including age, gender, popularity and activeness of the owner. It involves understanding the demographic and psychological traits of both potential bullies and victims to see the trait of the imbalance of power.
- *Content-based:* This dimension studies the linguistic features and semantic content of user-shared text, photos, and videos. Sentiment analysis, abusive language use, and contextual awareness are all important factors to consider. It aims to identify the trait of harmful, repeated content that may be intended to harass others.
- *Network-based:* This dimension analyzes the relationships and interactions between users on social media platforms. It includes social connections, frequency of interactions,.. It touches the traits of aggressive behavior and the imbalance of power. By understanding how users interact within their networks, this dimension helps in detecting coordinated or repeated bullying efforts. [9]
- *Episode-based:* This dimension examines users' activity over time, measuring factors such as posting frequency, time of posting, etc. It addresses aspects of user behavior and psychology, assisting in the identification of unusual or worrying trends that may indicate a move toward hostile or bullying conduct. It is critical to understanding the temporal dynamics of cyberbullying.

Each of these factors is critical for a complete approach to recognizing cyberbullying. In our work, we focused on the content-based dimension, combining sentiment analysis with lexical features to improve detection. This combination enabled us to capture both the emotive tone and specific language patterns that are frequently associated with cyberbullying.

III. Related Work

This section presents the related work in the area of detecting cyberbullying on social media, from using machine learning techniques , to deep learning techniques to the use of large language models.

A. The use of ML/DL methods

In their study [10], Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia developed a cyberbullying detection model using machine learning techniques combined with psychological features of Twitter users, such as personality traits, sentiment, and emotion. The model consisted of three main stages: data collection from Twitter using the hashtag #GamerGate, feature extraction (including content-based, network-based, and user-based features), and cyberbullying detection and classification. Personality traits were analyzed using IBM Watson's Personality Insight API, mapping the Big Five traits to the Dark Triad, while sentiment and emotion were assessed using the Indico API. They compared three machine learning algorithms—Random Forest, Naïve Bayes, and J48—using 10-fold cross-validation. The model categorized tweets into four roles: bully, aggressor, spammer, or normal. Their findings indicated that the J48 algorithm performed better when personality and sentiment features were included, while emotions did not significantly enhance the detection of bullying patterns.

Mohamad Ahmadinejad, Nashid Shahriar, and Lisa Fan [11] proposed a self-training data annotation method designed to label unlabeled content with high prediction confidence. This approach augments a labeled dataset, reducing imbalance and increasing diversity. To ensure reliability, they employed six different models, only adding samples to the labeled set if all

six classifiers agreed. The classifiers used were Random Forest with TF-IDF, Decision Tree with TF-IDF, XGBoost with TF-IDF, BERT with its own tokenizer, LSTM with Keras Embedding, and BiLSTM with Keras Embedding. Their classification process involved two phases: first, a binary classification to separate cyberbullying from non-cyberbullying content, followed by multi-class classification within the cyberbullying category. Their findings showed that even with an imbalanced test set (10% cyberbullying and 90% non-cyberbullying), the models, particularly BERT and XGBoost + TF-IDF, performed effectively.

The researches has been limited to binary/multi-class text classification due to the lack of comprehensive datasets for training and evaluation, in this work [12] they worked on multi-labeled classification to accurately identify various types of cyberbullying in text. So, first they created a dataset by collecting the texts from various social media (Kaggle, Twitter, WikiTalk, Youtube), merging four-binary labeled dataset to get 95608 samples of texts categorized into five classes (aggression, attack, toxicity, seximn acceptable) and 0 and 1 to indicates the presence and abscence of each class. After, they did the data preprocessing that includes cleaning the data, stemming with Porter's stemmer , lemmatization with Wordnet lemmatizer and spell correction with Symspellpy. Afterwards, they extracted the feature using TF-IDF. For the classification, they used machine learning algorithms : random forest, stochastic gradient descent (SGD), logistic regression (LR), multinomial naive bayes and also Deep learning models like BiLSTM, LSTM, BiGRU, C-LSTM (CNN combined with LSTM). The results indicate the CLSTM model had the highest accuracy, precision, and recall scores of 87.8%, 88.2%, and 88.5%, respectively.

Another study focused on using an unique hybrid Random Forest based CNN model for text classification for real-time datasets [13]. The authors collected the data from Twitter and Instagram via a snowballing technique. They did much preprocessing, this involved cleaning and filtering the social media content to remove noise, irrelevant information, and duplicate posts. Additionally, they conducted text normalization, stemming, and removal of stop words. To balance the data , they used the SMOTE technique. For the vectorization , they used GloVe word embedding. All general metrics showed that the CNN-RF model performed the best compared to Random Forest, SVM, Naive Bayes, RNN and CNN. There is also some studies [14], that used the attention mechanism with some deep learning models. Like this work [15], they proposed a complete model combining the bidirectional gated recurrent unit (Bi-GRU) and the self-attention mechanism. They used three datasets in their experiment, including two Twitter datasets and one Wikipedia dataset. They concluded that the self-attention layer significantly helps the model improve the performance of the class with fewer samples on all three unbalanced datasets.

While traditional machine learning models have achieved good results, they often struggle with understanding context, sarcasm, and irony, which can lead to misclassifications. These models also rely on manual feature extraction, making it easy to miss important nuances in the data. Deep learning models have improved upon this by automatically learning features, capturing more subtle details in the text. However, they too have difficulty with deep contextual understanding and subtle forms of cyberbullying. The use of sentiment analysis helps by identifying the emotional tone of messages but alone may not be a reliable indicator of cyberbullying, missing nuanced or context-specific cases, focusing on positive, negative sentiments which might not directly correspond to cyberbullying behavior. So it comes the use of Large Language models (LLM), offer a significant step forward, excelling at understanding the context, distinguishing between harmless and genuine cyberbullying. It can detect subtle differences in sentiment, identifying not just P/N tones but also underlying hostility or aggression. But, it requires significant computational resources and careful handling of ethical concerns

B. The use of Large language models and sentiment analysis

Many studies have used large language models like BERT, its variants [16], or even GPT-3 to detect cyberbullying. These two works [17] [18] used GPT3 as an Large Language Model to detect cyberbullying and they found that GPT3 can be on par with BERT in some cases but it still has some shortcomings like the model may become biased towards some types of cyberbullying. Dan Ottosson, fine-tuned a GPT3 Large Language model with a multi-labelled dataset from Kaggle (toxic comments classification challenge) to reduce the gap to platform moderation and to enhance the research of cyberbullying detection. This study [17] used more specifically a GPT-3 Ada model, they got an overall performance where 90% of all classified comments are correct. He concluded that GPT-3 can be on par with BERT. and that LLMs are slower in classifying a comment than ML models for their use in the content moderation. Sayanta Paul, Sriparna Saha [19] used the knowledge distillation method of BERT to minimize the cost. During fine-tuning on , they add it a fully connected layer over the hidden states and a softmax classifier to optimize it. During distillation, knowledge is transferred to the distilled model by training it on a transfer set and using a soft target distribution for each case in the transfer set that is produced by using the BERTLarge. It achieved the state-of-the-art results across three real-world corpora: Formspring 92% F1-score, Twitter 94% F1-score and Wikipedia 91% F1-score. But, their proposed model misclassified certain posts or comments which contain comments that are likely to appear in other categories. This study [20] fine-tuned BERT model to detect Toxic comments using theToxic Comments Classification Challlenge dataset from Kaggle with its default configuration by adding a drop out and classification layer and then tested it on real-world data, which are two different tweets datasets, collected in two different periods based on a case study of the UK Brexit. They compared the results with some variants of

BERT (DistilBERT, RoBERTa, BERT, Multilingual BERT), and it showed that BERT base had the best results (98,56% accuracy). This study [21] proposed a semi-supervised approach in detecting cyberbullying. In fact, they fine-tuned a BERT model on a Twitter Datasetto extract the sentimental features of the cyberbullying texts, by adding a classifier layer to determine the sentiment. The result shows better accuracy when using the BERT model sentiment analysis on the Twitter dataset, it gaves a better accuracy of 91.90%.

These various methods and models address multiple challenges in detecting cyberbullying, leveraging machine learning, deep learning, and large language models. In our study, we began by extracting sentiment features using Transformers. We then fine-tuned a DistilBERT model on our Kaggle dataset, incorporating the sentiment features and adding a classification layer. Finally, we evaluated our model on an unseen dataset with data from Facebook and Twitter to see its performance on different social media platforms.

IV. Proposed Methodology

In this section, the proposed work is discussed in detail. The steps involved are shown in Figure 1 and are summarized as follows:

- Data Preparation is the collection of our data where we merged the datasets extracted from Twitter, Kaggle and Wikipedia
- Data cleaning and preprocessing is performed on the dataset by removing punctuation marks, stopwords, lemmatization, etc,..
- Feature Extraction is performed to extract the sentiment features using Transformers
- Fine-tuning the Distilbert by adding the sentiment feature extracted, and adding a fully connected layer over the final hidden state that corresponds to the [CLS] input token.
- Various evaluation metrics are used to evaluate the performance of the model not only on the test set but on another dataset (unseeen) where the data is extracted from Facebook comments and Twitter.

A. Data Preparation

1) Dataset: The dataset we used is publicly available and is a combination of three merged datasets. The primary dataset is a cyberbullying dataset from Kaggle [1], which includes data from various sources focused on the automatic detection of cyberbullying. We specifically merged the data from Kaggle and Twitter, which includes different forms of cyberbullying such as hate speech, aggression, insults, and toxicity. We got an imbalance class distribution as shown in the figure 2. To enrich the cyberbullying class and help the model better understand its various aspects, we merged an additional dataset exclusively from Twitter. This Twitter dataset [2], containing over 47,000 tweets labeled by cyberbullying type, was incorporated into our existing data. We combined all types of cyberbullying into a single class, the cyberbullying class. After cleaning the data by removing duplicates and null values, we achieved a stronger representation of the cyberbullying class. However, this improvement came at the cost of underrepresenting the non-cyberbullying class as shown in the figure 3. So, we addressed the class imbalance by applying an oversampling technique and using focal loss as our loss function.

2) Sentiment Feature Extraction: We chose to extract sentiment features before applying any preprocessing to the text. This approach allows the sentiment analysis model to capture the full sentimental context, as preprocessing steps like stopword removal and lemmatization can strip away important nuances and alter the original meaning of the text. Sentiment analysis models, particularly those that rely on context, perform more accurately on raw text because they can fully leverage the subtleties that might be lost during preprocessing. [22]

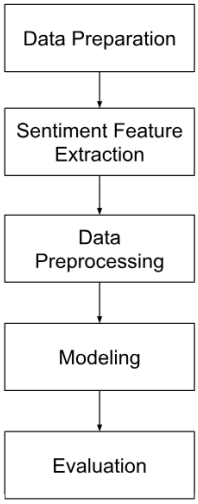

Fig. 1. Methodology for cyberbullying classification

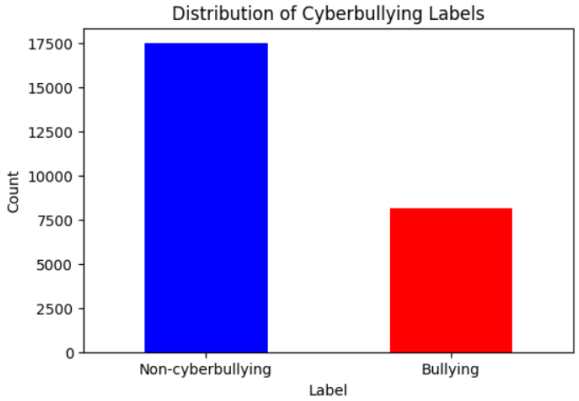

Fig. 2. Class Distribution of cyberbullying dataset

[1]https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset
[2]https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

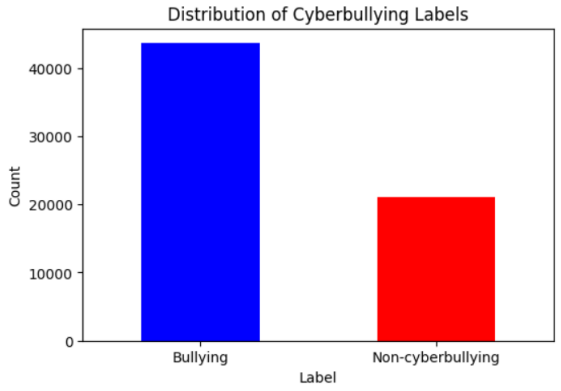

Fig. 3. Class distribution of the FINAL dataset

	Text	label	Sentiment	cleaned_text
0	@halalflaws @biebervalue @greenlinerzjm I read...	0	0	read contextno change meaning history islamic ...
1	@ShreyaBafna3 Now you idiots claim that people...	0	-1	idiot claim people tried stop becoming terrori...
2	RT @Mooseoftorment Call me sexist, but when I ...	1	0	call sexist go auto place would rather talk guy
3	@g0ssipsquirrelx Wrong, ISIS follows the examp...	1	-1	wrong isi follows example mohammed quran exactly
4	#mkr No No No No No No	0	0	

Fig. 4. Overview of the first 5 rows of the our final dataset after cleaning

There are various techniques and tools available for sentiment analysis, ranging from document-based to aspect-based approaches [23].

We opted to use a pre-trained sentiment model from Hugging Face's Transformers library. This model classified the text into positive, neutral, or negative sentiment before any modifications were made to the data. Using transformers for sentiment analysis offers significant advantages, such as a deep understanding of context, the ability to handle complex sentence structures. [22] This ensures that the sentiment features reflect the true emotional tone of the content, providing a more reliable foundation for the next steps in our study.

B. Data cleaning and preprocessing

Before building the model, we applied some light preprocessing steps to clean the data. We removed emojis, special characters, numbers, punctuation, links, URLs, hashtags, user mentions, and non-English words. We also eliminated retweet indicators, repeated words, and common stopwords like prepositions (in, on, at), conjunctions (and, thus, too), and articles (a, an, the). Finally, we used the WordNetLemmatizer to perform lemmatization, ensuring that words were reduced to their base forms for more effective analysis. The first 5 rows in our final dataset are shown in the figure 4 after cleaning and preprocessing the dataset.

For the tokenization, we used the pre-trained tokenizer of "distilbert base uncased" as its our pretrained model that we will fine-tune.

C. Data analysis

In this crucial phase, we started by exploring the relationship between sentiment and cyberbullying labels. Our first step was

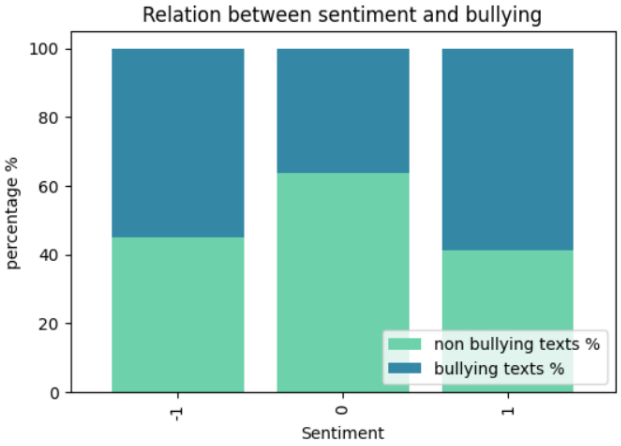

Fig. 5. Relationship between sentiment and cyberbullying

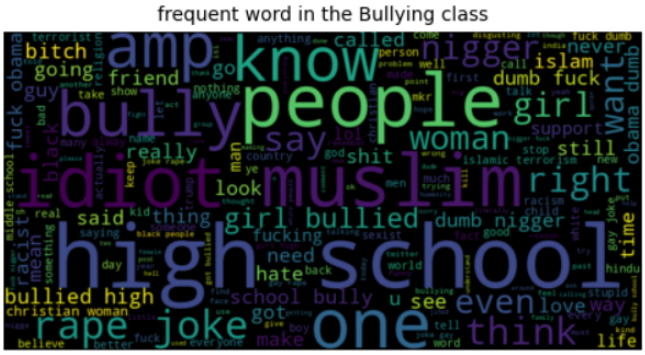

Fig. 6. Frequent Words on the cyberbullying class

to look at how the sentiment of the texts—whether positive, neutral, or negative— correlates with the likelihood of those tweets being labeled as cyberbullying. This analysis helps us uncover patterns in the data and provides valuable insights as we move forward with building our model.

As we can see in the figure 5, the chart reveals that negative sentiments are more likely to be associated with cyberbullying, while neutral and positive sentiments still show a significant presence of harmful behavior. This indicates that cyberbullying can occur across all sentiment types, making it important to consider these nuances in developing accurate detection models.

After analyzing the relationship between sentiment and cyberbullying, we created word clouds to visualize the most frequent words in both the cyberbullying and non-cyberbullying classes. This step helps to identify common language patterns and key terms associated with each class, providing further insights into how certain words or phrases are more likely to appear in cyberbullying content. Figure 6 shows the frequent word on the cyberbullying class and figure 7 shows the frequent on the non-cyberbullying class.

The word clouds provide a contrast between the language

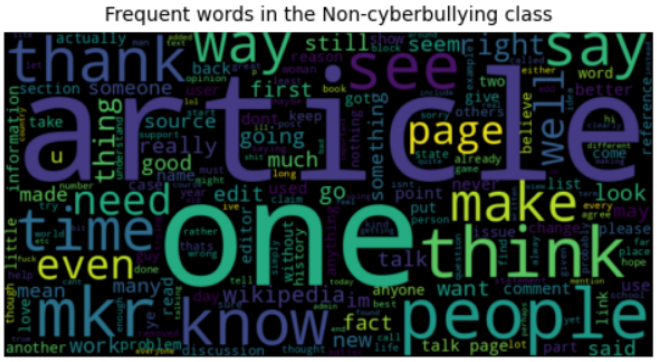

Fig. 7. Frequent Words on the non-cyberbullying class

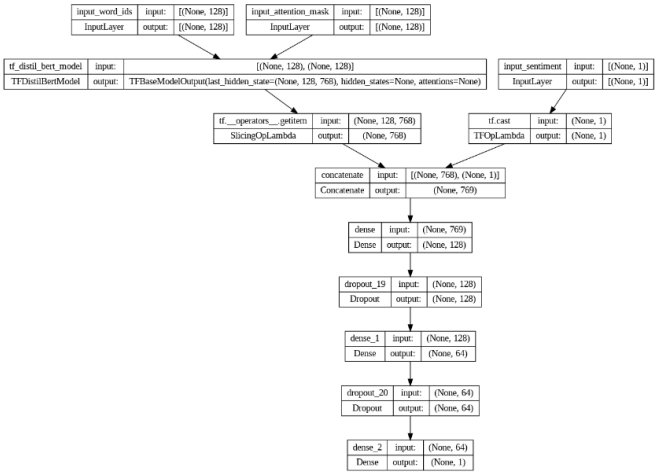

Fig. 8. Model's architecture

used in non-cyberbullying versus cyberbullying content. In the non-cyberbullying class, the most frequent words are neutral or constructive. Conversely, the cyberbullying class is dominated by aggressive and harmful language, with words like "bully," "idiot," and various slurs standing out.
This visualization highlights the distinct vocabulary associated with cyberbullying, proving the importance of detecting such patterns in text.

D. Modeling

We proposed a cyberbullying detection model based on transformers to overcome the limitations of RNNs. RNNs struggle to retain information when sentences are long and are slow in parallel processing [24]. Transformers, on the other hand, utilize a self-attention mechanism [25] that evaluates the importance of each word in a sentence in relation to all others. This allows the model to access and consider information from the entire input sequence at every step, making it more effective and efficient in processing and understanding complex language patterns.

A recent improvement on the natural language task introduced the BERT [26]. BERT stands for Transformers' Bidirectional Encoder is a language model that has been trained on two tasks, which are masked tokens prediction (Masked LM (MLM)) and next sentence prediction (NSP). It is a bidirectional model that has been pre-trained on unlabeled texts in both the left and right directions to understand the meaning of both contexts. BERT has several variants, including multilingual BERT, RoBERTa, and DistilBERT. DistilBERT [27] is a distilled version of Google's BERT, designed through a process called distillation, which reduces the number of parameters by 40% while maintaining nearly the same performance level as the original BERT. It also offers the advantage of faster processing, making it an ideal choice for our task.

In this study, we fine-tuned a distilBERT model to process and understand text inputs. The model takes the tokenized text and attention masks, which help it focus on the most relevant parts of each sentence. It uses a default configuration with 12 encoder blocks, 768 hidden dimensions, and 12 attention heads [28]. We then integrated the sentiment features with distilBERT's output for the [CLS] token, adding then several dense layers, two dropout layers to prevent overfitting. The final output is provided by a classifcation layer (a simple feed-forward layer with standard Sigmoid), based on both the deep text features and sentiment analysis, making the model effective. The model architecture is represented in the figure 8.

In the fine-tuning, most hyper-parameters stay the same as in DistilBERT's original training. Hence, a DistilBERT compatible tokenization model was used, which contains around 30,000 tokens. During this stage, the classifier and the pre-trained model weights are trained jointly. For finetuning, we use a maximum sequence size of 128 tokens, a batch size of 16, and we set the learning rate to 10-5.

1) Optimization technique used:

- **Adam optimizer:** Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data [29]. It offers adaptive learning rates, which dynamically adjust the learning rate for each parameter based on past gradients. This adaptivity provides faster convergence and improves performance across various data and model architectures.
- **Reduce learning rate on Plateau:** ReduceLROnPlateau technique is an adaptive approach helps to fine-tune the model more effectively by lowering the learning rate when progress stalls, allowing for more precise convergence [30]. Initially set at 1×10-5, the learning rate is reduced by a factor of 0.5 if the validation loss plateaus for 2 consecutive epochs. The learning rate is capped at a minimum value of 1×10-7 to ensure stability and prevent excessively small updates. This method not only enhances the training process but also optimizes performance by adapting the learning rate in response to the model's performance on the validation set.
- **Early stopping technique** The early stopping technique

is used to reduce overfitting without compromising on the model's accuracy. It consists of stopping training the model when the validation loss begins to increase while the training loss continues to decrease [31].

- **Regulization L2** L2 regularization, also known as weight decay, penalizes large weights by adding a term to the loss function proportional to the sum of the squared weights. This encourages the model to maintain smaller, more manageable weights, which helps in preventing it from fitting noise in the training data.
- **Focal loss** Focal loss [32] is designed to give more emphasis to hard-to-classify examples, which is particularly useful when dealing with imbalanced data where certain classes are underrepresented. Given the class imbalance present in our dataset, we utilized focal loss as our loss function to address this issue effectively. The focal loss function is parameterized with a gamma of 2.0 and an alpha of 0.25. The alpha parameter helps to balance the importance between positive and negative classes, while the gamma parameter adjusts the rate at which easy examples are down-weighted. Specifically, focal loss reduces the relative loss for well-classified examples and focuses more on the challenging, misclassified examples. It improves the model's ability to learn from the minority class and enhances its performance.

V. Experiment Results

This section contains a comprehensive analysis of the results of the research. We used a lot of evaluation metrics including accuracy, recall, precision, F1-score and AUC to judge how well our suggested model worked. And we saw also the confusion matrix in order to see if the model had the capicity to identify cyberbullying, and this analysis provides helpful insights into both the model's advantages and potential weaknesses.

A. Results of the study on the train and validation set

1) Training and validation loss: The figure 9 shows the training and validation loss during the fine-tuning process, where the model reaches its lowest value 0.0236 in the training set and 0.0486 in the valdation set. It demonstrates the model's optimization progress during training. The decreasing loss values indicate that the model's predictions closely match the true labels. This reflects improved predictive capabilities and parameter adjustments throughout the training process.

*2) **Confusion matrix**:* Additionally, we explored the confusion matrix for each class to gain more insights about the classes of misclassified instances. The figure 10 describes the performance of our classification model on the validation set. The matrix shows that out of the true "Not Cyberbullying" instances, 4,636 were correctly identified, but 418 were misclassified as "Cyberbullying." On the other hand, out of the true "Cyberbullying" cases, the model correctly identified 4,974 instances, but 693 were incorrectly labeled as "Not Cyberbullying."

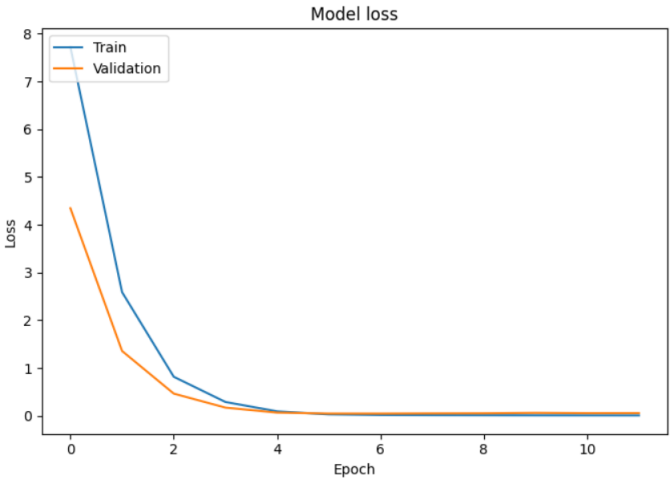

Fig. 9. Training Loss vs Validation Loss during fine-tuning

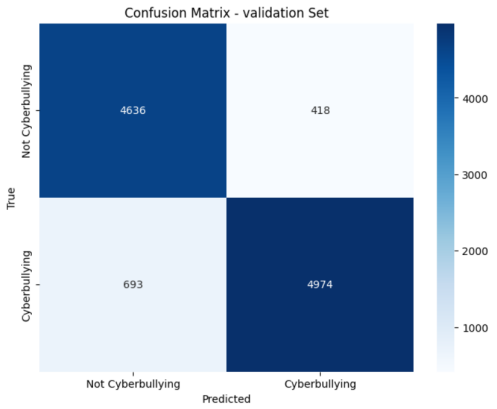

Fig. 10. Confusion Matrix on the validation set

*3) **Accuracy**:* To know how well a model performs in terms of correctly classifying instances into their respective categories, the accuracy metric is used. In fact,it is calculated as the ratio of correctly predicted instances (true positives and true negatives) to the total number of instances in the dataset. Mathematically, accuracy can be expressed as (1) :

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

or as (2) :

$$Accuracy = \frac{\text{TP + TN}}{\text{TP + TN + FN + FP}} \quad (2)$$

Our model achieved on the training set **94,56%** accuracy and on the validation set **90%** accuracy.

However, it's important to note that accuracy might not be sufficient to evaluate the performance of our model as in our case the classes are imbalanced. So we used additional metrics, such as precision, recall, and F1-score, to provide a more nuanced view of a model's performance.

*4) **Precision, Recall, F1-score**:* To assess the performance of classification models, especially in scenarios where class imbalances, we use *precision*, *recall* and *F1-score* metrics.

- **Precision:**
  Precision measures the accuracy of positive predictions made by the model. In other words, it indicates how many of the instances predicted as positive by the model are actually true positives. It can be expressed as (3):

$$Precision = \frac{\text{TP}}{\text{TP + FP}} \tag{3}$$

- **Recall:**
  Recall indicates how well the model captures all instances of a particular class. It measures the ability of the model to correctly identify all relevant instances from the dataset.

$$Recall = \frac{\text{TP}}{\text{TP + FN}} \tag{4}$$

- **F1-score:**
  The F1-score is the harmonic mean of precision and recall. It provides a balance between these two metrics and is particularly useful when there is a trade-off between precision and recall. It is expressed as (5). The F1-score is useful for evaluating the overall effectiveness of a model's classification performance, especially in scenarios where both false positives and false negatives need to be minimized.

$$F1score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision + Recall}} \tag{5}$$

The table 1 shows the results during training. In fact, the model achieved an accuracy of 94.56%, precision of 94.81%, recall of 93.31%, and an F1-Score of 94.56%. These results indicate that the model performed well on the training data, with a solid balance between precision and recall, suggesting it effectively learned the patterns in the data.

TABLE I
TRAINING RESULTS

Metrics	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Values	94.56%	94.81%	93.31%	94.56%

The table 2 shows the results on the validation set. In fact, the model's performance dropped slightly, with an accuracy of 89.63%, precision of 92.24%, recall of 87.77%, and an F1-Score of 89.95%. These results suggest that while the model is robust, there is still room for improvement, especially in enhancing its ability to generalize to new data.

TABLE II
VALIDATION RESULTS

Metrics	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Values	89.63%	92.24%	87,77%	89,95%

	precision	recall	f1-score	support
Not Cyberbullying	0.86	0.92	0.89	5054
Cyberbullying	0.93	0.87	0.90	5666
accuracy			0.89	10720
macro avg	0.90	0.90	0.89	10720
weighted avg	0.90	0.89	0.89	10720

Fig. 11. Classification report on the test set

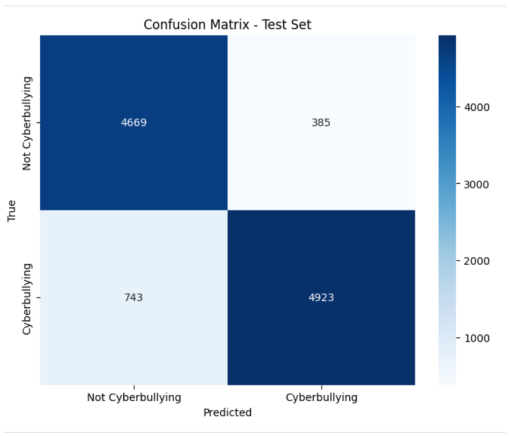

Fig. 12. Confusion matrix on the Test set

B. Results of the study on the test set

The figure 11 show the classification report done on the test set. The model has a precision of 86% for "Not Cyberbullying" and 93% for "Cyberbullying," meaning it's slightly better at correctly identifying cyberbullying instances. The recall is 92% for "Not Cyberbullying" and 87% for "Cyberbullying," indicating that the model is more likely to miss some cyberbullying cases. The overall accuracy of the model is 89%, which shows it performs fairly well. The macro and weighted averages for precision, recall, and F1-score are all around 0.89 to 0.90, suggesting that the model is balanced and consistent in its predictions across both classes.

The confusion matrix on the test set is shown in the figure 12.

The confusion matrix shows that the model is fairly accurate on the test set, correctly identifying most cases, but it still misses some cyberbullying instances, with 743 false negatives where cyberbullying was not detected.

TABLE III
TEST RESULTS

Metrics	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Values	89.48%	92.75%	86,89%	89,72%

The test results shown in the table 3, indicate that the model has an accuracy of 89.48%, with a precision of 92.75%, recall of 86.89%, and an F1-Score of 89.72%. These metrics suggest that while the model is generally accurate and precise in detecting cyberbullying, there is a slight trade-off in recall,

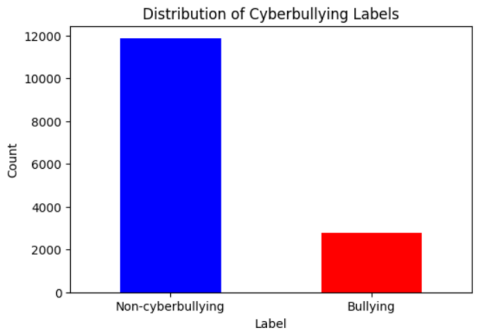

Fig. 13. Class distribution of the evaluation dataset

meaning it occasionally misses some cyberbullying cases. The F1-Score, which balances precision and recall, confirms that the model performs solidly overall.

VI. Evaluation on Facebook and Twitter dataset

A. Facebook and Twitter dataset

The data is collected from Twitter and Facebook groups, this dataset [3] is based on suspicious activities like racism, discrimination, abusive language, threatening, which mostly comes in cyberbullying. The tagging of the data is based on suspicious words which are being used in the tweets and comments. Suspicious data is tagged with 1 and Non-suspicious data is tagged with 0 manually after scraping the data. The dataset contains up to 20 thousand rows of labels. Around 12 thousands of the data is tagged with a negative sentiment like (racism, discrimination, abuse) while 8 thousands of the data is tagged positive or neutral sentiment which shows the data is not suspicious.

We did some cleaning to the dataset, just to remove the non values and empty and unnecessary columns, then we extracted the sentiment features by using the trransformers, the same method that we used to our training dataset. Finally we passed the data to our tokenizer and then to our model.

The class distribution in this dataset is the inverse of ours; here, the cyberbullying class is actually the minority class as we see in the figure 13.

B. Results

The overall results on the unseen dataset are shown in Table 4. As we can see, the accuracy is 76.71%, which might initially seem reasonable. However, when we examine the other metrics, we find that the model's performance is quite poor. The precision is only 32.41%, indicating that a significant number of instances classified as positive (cyberbullying) are actually false positives. The recall is 36.59%, meaning the model is missing a large portion of actual cyberbullying cases. The F1-Score is quite low at 29.08%, reflecting that the model is struggling to correctly identify and classify instances of cyberbullying. These results suggest that the model is not well-suited to this unseen dataset.

[3]https://www.kaggle.com/datasets/syedabbasraza/suspicious-communication-on-social-platforms

TABLE IV
Test results

Metrics	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Values	76,71%	32.41%	36,59%	29,08%

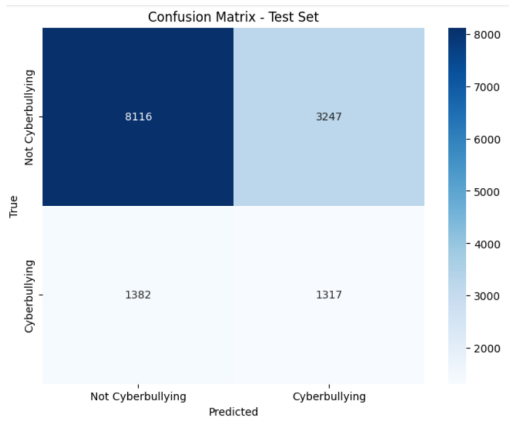

Fig. 14. Confusion matrix on the evaluation dataset

The confusion matrix shown in the figure 14, indicates that the model correctly classified 8,116 instances of "Not Cyberbullying" but misclassified 3,247 of them as "Cyberbullying" (False postives). Conversely, it correctly identified 1,317 "Cyberbullying" cases but failed to recognize 1,382 of them, classifying these as "Not Cyberbullying" (False negatives).

The results suggest that while the model can generalize to some extent, it still struggles with accurately detecting cyberbullying in unseen data, which might indicate that the model needs more diverse training data to improve its robustness and generalization capabilities.

VII. Conclusion

The main goal of this research was to evaluate the performance and generalization capabilities of our proposed model across different social media datasets. Our model, which leveraged sentiment features, achieved promising results with 94.56% accuracy during training, 89.63% on the validation set, and 89.48% on the test set, significantly outperforming traditional machine learning models.

However, when tested on an unseen dataset with a different class distribution, the model's accuracy dropped to 76.71%, and other metrics like precision, recall, and F1-score showed poor performance. This indicates that our model struggled to generalize to new data, particularly from Facebook and Twitter, suggesting that it may require further adjustments and exposure to more diverse data to enhance its generalization and overall effectiveness.

Notably, the BERT model has the potential to deliver even more accurate results if trained on a larger and more varied dataset. Incorporating all the features proposed in this research, beyond just sentiment, could also improve the model's ability to detect cyberbullying more effectively. Furthermore, a multilingual approach could be a valuable direction for future

work, enabling the model to identify cyberbullying across different languages and apply evaluations to other social media platforms like TikTok and Instagram.

Additionally, exploring the reasoning behind the model's classification of certain texts or posts as cyberbullying could provide valuable insights into its decision-making process, further refining our approach and ensuring that our model is not only accurate but also interpretable and reliable in real-world applications.

REFERENCES

[1] MAYA DOLLARHIDE. Social media: Definition, importance, top websites and apps. https://www.investopedia.com/terms/s/social-media.asp.

[2] MARCOS ZAMPIERI, SHERVIN MALMASI, PRESLAV NAKOV, SARA ROSENTHAL, NOURA FARRA, AND RITESH KUMAR. *SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval)*. (2019).

[3] What is bullying. https://cyberbullying.org/what-is-bullying.

[4] What is cyberbullying. https://cyberbullying.org/what-is-cyberbullying.

[5] How cyberbullying occurs. https://www.stopbullying.gov/cyberbullying/what-is-it.

[6] Types of cyberbullying. https://socialmediavictims.org/cyberbullying/types/.

[7] AHMET AKER, ALFRED SLIWA, FAHIM DALVI, AND KALINA BONTCHEVA. *Rumour verification through recurring information and an inner-attention mechanism*. Online Social Networks and Media **13**, 100045 (2019).

[8] SAED REZAYI, VIMALA BALAKRISHNAN, SAMIRA ARABNIA, AND HAMID R ARABNIA. Fake news and cyberbullying in the modern era. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 7–12 (2018).

[9] YUYI LIU, PAVOL ZAVARSKY, AND YASIR MALIK. Non-linguistic features for cyberbullying detection on a social media platform using machine learning. 12 (2019).

[10] VIMALA BALAKRISHNAN, SHAHZAIB KHAN, AND HAMID R. ARABNIA. *Improving cyberbullying detection using twitter users' psychological features and machine learning*. Computers Security **90**, 101710 (2020).

[11] MOHAMAD AHMADINEJAD, NASHID SHAHRIAR, AND LISA FAN. *Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset*. Thèse de Doctorat, PhD thesis, Faculty of Graduate Studies and Research, University of Regina (2023).

[12] NAIMUL ISLAM, REZAUL HAQUE, PIYUSH KUMAR PAREEK, MD BABUL ISLAM, IMAM HOSSAIN SAJEEB, AND MAHEDI HASSAN RATUL. Deep learning for multi-labeled cyberbully detection: Enhancing online safety. In *2023 International Conference on Data Science and Network Security (ICDSNS)*, pages 1–6. IEEE (2023).

[13] T NITYA HARSHITHA, M PRABU, E SUGANYA, S SOUNTHARRAJAN, DURGA PRASAD BAVIRISETTI, NAVYA GADDE, AND LAKSHMI SAHITHI UPPU. *Protect: a hybrid deep learning model for proactive detection of cyberbullying on social media*. Frontiers in artificial intelligence **7**, 1269366 (2024).

[14] SULIMAN MOHAMED FATI, AMGAD MUNEER, AYED ALWADAIN, AND ABDULLATEEF O BALOGUN. *Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction*. Mathematics **11**(16), 3567 (2023).

[15] YONG FANG, SHAOSHUAI YANG, BIN ZHAO, AND CHENG HUANG. *Cyberbullying detection in social networks using bi-gru with self-attention mechanism*. Information **12**(4), 171 (2021).

[16] JATIN KARTHIK TRIPATHY, S SIBI CHAKKARAVARTHY, SURESH CHANDRA SATAPATHY, MADHULIKA SAHOO, AND V VAIDEHI. *Albert-based fine-tuning model for cyberbullying analysis*. Multimedia Systems **28**(6), 1941–1949 (2022).

[17] DAN OTTOSSON. Cyberbullying detection on social platforms using largelanguage models, (2023).

[18] KE-LI CHIU, ANNIE COLLINS, AND ROHAN ALEXANDER. *Detecting hate speech with gpt-3*. arXiv preprint arXiv:2103.12407 (2021).

[19] SAYANTA PAUL AND SRIPARNA SAHA. *Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification*. Multimedia Systems **28**(6), 1897–1904 (2022).

[20] HONG FAN, WU DU, ABDELGHANI DAHOU, AHMED A EWEES, DALIA YOUSRI, MOHAMED ABD ELAZIZ, AMMAR H ELSHEIKH, LAITH ABUALIGAH, AND MOHAMMED AA AL-QANESS. *Social media toxicity classification using deep learning: real-world application uk brexit*. Electronics **10**(11), 1332 (2021).

[21] ADITYA DESAI, SHASHANK KALASKAR, OMKAR KUMBHAR, AND RASHMI DHUMAL. Cyber bullying detection on social media using machine learning. , **40**, page 03038. EDP Sciences (2021).

[22] SAYYIDA TABINDA KOKAB, SOHAIL ASGHAR, AND SHEHNEELA NAZ. *Transformer-based deep learning models for the sentiment analysis of social media data*. Array **14**, 100157 (2022).

[23] MARIA PONTIKI, DIMITRIOS GALANIS, HARIS PAPAGEORGIOU, ION ANDROUTSOPOULOS, SURESH MANANDHAR, MOHAMMAD AL-SMADI, MAHMOUD AL-AYYOUB, YANYAN ZHAO, BING QIN, ORPHÉE DE CLERCQ, ET AL. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30 (2016).

[24] ALOK SHANKAR. *Understanding google's "attention is all you need" paper and its groundbreaking impact*. (2023).

[25] ASHISH VASWANI. *Attention is all you need*. arXiv preprint arXiv:1706.03762 (2017).

[26] JACOB DEVLIN. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805 (2018).

[27] VICTOR SANH, LYSANDRE DEBUT, JULIEN CHAUMOND, AND THOMAS WOLF. *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108 (2019).

[28] JDMCK LEE AND K TOUTANOVA. *Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805 **3**(8) (2018).

[29] JASON BROWNLEE. Gentle introduction to the adam optimization algorithm for deep learning, (July 3, 2017).

[30] Reducelronplateau.

[31] What is early stopping?

[32] SHREEJAL TRIVEDI. Understanding focal loss—a quick read, (May 2, 2022).