

AMERICAN UNIVERSITY OF BEIRUT

LEVERAGING A BILSTM-BASED EMOTION
RECOGNITION TRANSFER LEARNING MODEL TO
IDENTIFY ABUSIVE LANGUAGE PATTERNS FOR
COMPLEX PHRASAL ANALYSIS IN CYBERBULLYING
DETECTION

by
MARITA GEORGES MATTA

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science in Business Analytics
to the Suliman S. Olayan School of Business
at the American University of Beirut

Beirut, Lebanon
January 2024

AMERICAN UNIVERSITY OF BEIRUT

LEVERAGING A BILSTM-BASED EMOTION RECOGNITION
TRANSFER LEARNING MODEL TO IDENTIFY ABUSIVE
LANGUAGE PATTERNS FOR COMPLEX PHRASAL
ANALYSIS IN CYBERBULLYING DETECTION

by
MARITA GEORGES MATTA

Approved by:

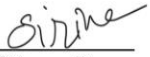
Dr. Bijan Azad, Professor - Director
Suliman S. Olayan School of Business

Signature 
Advisor


Dr. Fouad Zablith, Associate Professor
Suliman S. Olayan School of Business

Signature 
Co-Advisor

Dr. Sirine Taleb, Visiting Assistant Professor
Suliman S. Olayan School of Business

Signature 
Member of Committee

Dr. Walid Nasr, Associate Professor
Suliman S. Olayan School of Business

Signature 
Member of Committee

Date of Thesis Defense: 31/01/2024

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Matta Marita Georges
 Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

- As of the date of submission
- One year from the date of submission of my thesis.
- Two years from the date of submission of my thesis.
- Three years from the date of submission of my thesis.

 Marita Matta February 7, 2024

Signature

Date

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor and supervisor, Dr. Bijan Azad, for his invaluable guidance, unwavering support, and encouragement throughout the entire process of writing this thesis. His expertise, patience, and insightful feedback have been instrumental in shaping the direction of my research and enhancing the quality of this work.

I would also like to acknowledge the contributions of my co-advisor and member of my thesis committee, Dr. Fouad Zablith, Dr. Sirine Taleb, and Dr. Walid Nasr, who have provided valuable insights, encouragement, and support throughout this journey.

I am also immensely thankful to the Suliman S. Olayan School of Business at the American University of Beirut for providing the resources and environment conducive to academic growth and scholarly inquiry. The opportunities and facilities offered by the department have significantly contributed to the completion of this thesis.

Furthermore, I extend my heartfelt appreciation to my family for their unconditional love, understanding, and encouragement. Their support has been a constant source of motivation and inspiration, empowering me to pursue my academic goals with determination and perseverance.

This work would not have been possible without the collective support and encouragement of all those mentioned above. Thank you for believing in me and being part of this fulfilling academic endeavor.

ABSTRACT OF THE THESIS OF

Marita Georges Matta

for

Master of Science

Major: Business Analytics

Title: Leveraging a BiLSTM-based Emotion Recognition Transfer Learning Model to identify abusive language patterns for complex phrasal analysis in cyberbullying detection

In the last two decades, the penetration of Social Networking Sites and Social Media (SNS/SM) platforms has risen to include more than one-third of the global population. However, the use of SNS/SM has produced both positive and negative results so much so that there have been calls to researchers pay immediate and far greater attention to these contradictory effects of SNS/SM capabilities. A key negative impact is the fast and significant rise of cyberbullying.

Cyberbullying has emerged as a serious act on social networking sites/social media (SNS/SM) platforms in today's digital society. Statistics underscore this whereby 42% of individuals indicate they have experienced cyberbullying, more specifically 38% of females and 54% of males have also experienced some form of bullying. This form of dysfunctional social act is expressed via aggression, harassment, and toxic behavior poses severe consequences to increasing penetration of SNS/SM.

Simultaneously, there has been a great awareness of the need for moderating contents on (SNS/SM) to detect and reduce cyberbullying contents. It is also recognized that human content moderation on SNS/SM is impractical and too costly. Therefore, there is an increasing need for accurate methods of content moderation that are less reliant on human judgement and instead employ sophisticated machine learning methods. A key shortcoming of the current machine learning approaches to cyberbullying detection is that their accuracy needs to be improved significantly to be relied upon for practical deployment of non-human content moderation on SNS/SM.

Problem of using advanced data analytics and AI-based techniques in cyberbullying detection in the service of reducing the latter has been extensively studied. However, existing research method have faced challenges with the issue of false positives whereby many identified instances may not be cyberbullying. For instance, "I hate you" and "I hate thinking about the future" both contain the word "hate", yet the second sentence is a non-cyberbully phrase/sentence that contains the word hate as a metaphor for personal discomfort with the uncertainty associated with one's future.

Indeed, more accurate detection can come from a deeper contextually sensitive detection methods going beyond identifying simple hate words, whereby distinguishing

between harmful and innocuous expressions becomes the focal research problem. To address this, we will conduct a complex phrasal analysis to identify cyberbullying through the detection of the underlying emotion behind the seemingly toxic phrases. This approach aims to mitigate false positive predictions, which can occur when relying on hate keywords or an online hate corpus.

To achieve our goal, we employ an emotion recognition transfer learning model to comprehend the underlying emotion trigger of cyberbullying and enhance its detection. To accomplish this, we use a Bi-LSTM pre-trained emotion detection model with a 92% accuracy on the training set. Then, we adapt the knowledge gained from the first model to improve learning on a different but related task which is improving cyberbullying detection by integrating elements of context based on emotion identification within the expressed phrase. Based on extensive testing and training of the model on the data, we propose that LSTM and CNN multi-label-based classification model which is exhibiting an 80% accuracy on the training as superior approach to cyberbullying detection.

To evaluate and validate our results, we randomly split our dataset into training and validation sets, testing against an unseen dataset. A comparative analysis with (Gencoglu, 2021) model reveals a significant improvement in performance using McNemar's test and the Paired T-test. Notably, our model shows an 18% improvement on (Gencoglu, 2021) model¹.

In summary, this research has addressed the limitations inherent in keyword-based cyberbullying detection methodologies, thereby making a meaningful contribution to the field of cyberbullying detection in SNS/SM research.

Key Words: social networking sites, social media, cyberbullying detection, machine learning, complex phrasal analysis

¹ Aimed at mitigating bias in cyberbullying detection avoiding misclassifying text containing identity terms like black, jew, gay, white, Christian, or transgender in innocuous contexts.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	1
ABSTRACT.....	2
ILLUSTRATIONS.....	7
TABLES	8
INTRODUCTION.....	9
A. Goal and Scope.....	13
1. Research Questions.....	13
2. Methodology.....	14
3. Research Contribution.....	14
B. Structure of the Document	15
LITERATURE REVIEW	16
A. The Gap.....	18
1. Thesis Process.....	19
2. Research Model and Hypothesis	20
DATA COLLECTION METHODOLOGY	21
A. The Data Collection Process	21
1. The Emotional Dataset.....	21
2. The Cyberbullying Dataset.....	22
B. Data Processing and Analysis	23
1. Introduction	23

2. Data Cleaning	24
C. Feature Engineering	26
1. Polarity and Sentiment	26
2. countvectorizer	28
3. TF-IDF	29
D. Data Exploration.....	31
DATA ANALYSIS	35
A. Emotion Recognition Model	39
B. transfer learning	44
C. Cyberbullying Detection	45
EVALUATION AND RESULTS	49
A. Evaluation Metrics	49
B. Model Comparison	51
C. Statistical Tests Results.....	52
CONCLUSION	56
A. Research Summary	56
B. Research Contribution	57
C. Research Limitations	58
D. Future Research Directions	59
APPENDIX	61

REFERENCES.....72

ILLUSTRATIONS

Figure

1. Thesis Process Map	19
2. Proposed Research Model.....	20
3. Kaggle Emotion Dataset	22
4. HuggingFace Emotion Dataset.....	22
5. Kaggle Cyberbullying Dataset	23
6. Data Analysis Methodology Process	24
7. Data Processing Function.....	26
8. Feature Creation Functions	27
9. Count of Unique Emotions Labels	32
10. Tweets with less than 30 words.....	33
11. Distribution of tweet lengths for different labels	33
12. Word Cloud Visualizing Most Occurring Text	34
13. Hyperparameter Tuning Process for Multinomial Naïve Bayes	42
14. Construction of Our BiLSTM Neural Network Model	44
15. Construction of our LSTM & CNN Model.....	48
16. Models Accuracy on Unseen Data	54
17. Paired T-test Implementation and Results	54
18. McNemar’s Test Implementation and Results	55
19. “I hate when I think about you” (Gencoglu, 2021) Model Prediction.....	61
20. “What you are doing is bad” Enhanced Model Prediction	62
21. Optimal K Determination.....	63
22. Silhouette Score for K=2.....	64
23. Silhouette Score for K=3.....	65
24. Silhouette Score for K=4.....	66
25. Identified Clusters as Toxicity and Aggressivity	66

TABLES

Table	
1. Emotion Recognition Models	40
2. Enhanced Cyberbullying Detection Models	46

CHAPTER I

INTRODUCTION

Our lives have been completely transformed by the social networking sites and (SNS) and social media (SM) platforms, which has prioritized connectedness. The SNS/SM and their action possibilities have exploded, paving the way for penetration of SNS/SM in all aspects of life.

One of the basic ways the SNS/SM have changed our daily lives is the contradictory effects it has on the ongoing social communication and knowledge sharing (Majchrzak, et al. 2013). Indeed, there are undeniable advantages to use of SNS/SM, however, we cannot overlook the negative aspects it has brought forth. With the rise of sextortion, fraud, cyberbullying, and the construction of fake profiles, SNS/SM can encourage harmful behavior. The SNS/SM's anonymity and accessibility have afforded some people the capability to victimize and abuse others (Faraj & Azad 2012). Recognizing the possible harm that SNS/SM might cause is essential as it becomes more and more integrated into our daily lives. Certain individuals can now engage in bullying relying on the naivety of others, and can also cause mental distress via SNS/SM . In recent years, the press coverage and the research by companies operating SNS/SM ² have demonstrated that teenagers have been a specific cross section of population who are particularly vulnerable victims of cyberbullying.

Bullying is an intentional action that aims to inflict psychological and other injury on people whose victims lack the ability to protect themselves. It is based on systematic abuse and imbalance of power (Rigby, 2002). It involves intentional harm,

² Frances Hagan of facebook whistleblower

making it difficult for victims to defend themselves. Bullies now have a broad set of capabilities afforded to them, whereby they can harass their victims around-the-clock, causing perhaps unseen yet painful harm. The adjective "cyber" in the term "cyberbullying" refers to bullying that occurs through SNS/SM. It includes a range of negative actions, including sending hurtful or threatening messages, starting rumors, uploading embarrassing images or videos, and excluding someone from online networks or organizations (Slonje & al., 2013). Indeed, SNS/SM offer an environment that is favorable for cyberbullying to occur (Kowalski et al., 2014). The penetration and widespread use of SNS/SM appears to have no end in sight. As such there is an increasing need for means through which to detect cyberbullying and attempts to stem its rise. However, due to the limitations of the human moderation approach, SNS/SM companies are increasingly relying on developing digital means of moderation employing machine learning and other analytics tools.

At the same time, SNS/SM are also important data sources that give those interested in research, the ability to analyze data which can provide insights into how people engage in online bullying. Furthermore, these data are potential treasure trove for researchers who want to study cyberbullying to develop techniques for detecting cyberbullying more accurately by digital ML and other tools. In this study, we seek to examine the potential relationship between expressed emotions and cyberbullying, especially those expressed via written language and text on SNS/SM. Initially the study's goal is to examine the expressed emotion embedded within the cyberbullying text. Using emotion as a feature for detection we attempt to integrate context (using complex phrases) to order to prevent false-positive detection with the framework of non-human moderation by SNS/SM.

- **Motivation**

When SNS/SM were not yet commonplace, gathering user information was a difficult task. The development of smartphones caused a significant change in the environment since they made it possible to acquire consumer data virtually instantly. But this SNS/SM advance comes with a price, an abundance of data that may be frequently used for ill intent, including cyberbullying. In certain cases, the SNS/SM use meant to improve social connectedness has turned into a means to spread hatred.

Think about the powerful effect of a single tweet saying: “Our first child was born today!” While this tweet may bring joy to some, it may also serve as a painful reminder of unmet expectations for others who are dealing with reproductive issues that prevent them from becoming parents. A complex interplay of emotions is set in motion by the rapid nature of SNS/SM interactions and the immediate emotion response that such content triggers. But not everyone always navigates this complex web with compassion and understanding. An act of cyberbullying that is triggered like this. It points to the intricate relationship between SNS/SM and human emotion as well as subsequent cyberbullying communicative action which can lead to harm on the human target.

The range of actions cyberbullying agents can take within cyberbullying is almost unlimited. However, our complement of tools to study the effects of the cyberbullying on subjects is limited but growing. People who struggle with their own emotional reactions could unintentionally turn to making emotionally cold remarks or projecting their unresolved emotions onto other people. For instance, consider a scenario where a person in their twenties has been diligently searching for a job for six months. While, scrolling through LinkedIn, they come across posts from friends who

have successfully landed positions in a prestigious corporate setting. The emotions of fear, anger, or worry may arise in response to their own job search struggles. In such a situation, the individual might be tempted to express these negative emotions through cyberbullying in the comments section, projecting their frustration onto others.

Alternatively, they may choose to express best wishes for the candidate and productively doubling down on her own job search. This highlights the complex interplay between personal struggles and behavior on SNS/SM, emphasizing the need for individuals to navigate their emotions in a healthy and constructive manner within the online landscape.

SNS/SM comments are more than just words; they are a treasure trove of information for research to disentangle the above complex web of relationship between cyberbullying, triggered emotion and expressed sentiments/emotions. The expressed comments may be employed to reveal the emotion triggered that lead people to engage in cyberbullying. This thesis seeks to explore this treasure trove, revealing the complex relationships between digital exchanges, emotion reactions, and the problem of cyberbullying detection.

As a result, this thesis investigates the role of emotion in eliciting interactions that can exhibit cyberbullying expressions. Through an analysis of SNS/SM user comments, we aim to identify the emotion triggers that may underpin instances of cyberbullying.

A. Goal and Scope

The main objective of this thesis is to offer a technique is aimed to contribute and enhance the growing set of capabilities for non-human content moderation on SNS/SM by increasing the accuracy of cyberbullying detection. More specifically, this research intends focuses on the enhancement of practical techniques for the detection, avoidance, to reduce identification of false positive cases of cyberbullying which singularly focus on single words.

The research will concentrate on a multi-platform analysis of social media comments across well-known platforms including Kaggle, YouTube, and Twitter to meet this objective. Examining the linguistic and affective details included in user-generated text comments will be the main focus. Finding observable trends and indicators that point to possible cases of cyberbullying is the main goal. The thesis seeks to improve our knowledge of the mechanisms behind cyberbullying and offer practical insights that can guide the creation of specific enhanced accuracy detection solutions through this in-depth investigation.

1. Research Questions

Current efforts for cyberbullying detection include its detection and occurrence by proving its actual presence without actually knowing the root cause behind such actions.

As such, the thesis aims to answer the following research questions:

- "How do specific emotions manifested in social media content influence, trigger, and enable cyberbullying expressions?"

- "To what extent can analyzing complete sentences, compared to individual words, improve the accuracy of cyberbullying detection, and what specific semantic and syntactic patterns characterize cyberbullying language beyond the presence of explicit hate words?"

2. Methodology

Our methodology entails using transfer learning to anticipate the emotion context of cyberbullying incidents, designed for the enhancement of cyberbullying detection from single word to complex phrasal analysis. We will employ a comprehensive collection of user-generated emotion phrases to train an emotion detection model. This model will then be transferred to the cyberbullying dataset, containing textual comments from various SNS/SM. This approach aims to enhance and improve the detection of cyberbullying. The algorithm will be developed iteratively through training and validation, and its efficacy will be ensured by assessment measures including precision, recall, and F1-score.

3. Research Contribution

This study's main contribution is to clarify the motivations that underlie expressed cyberbullying phrases by identifying the emotion that influence people to express their emotions in this way. This thesis attempts to provide a detailed understanding of the emotional context surrounding cyberbullying text expressions by exploring and distinguishing those that signify toxic and aggressive textual comments.

This research is among few studies that contributes to this body of knowledge by going beyond superficial cyberbullying detection and those focused on single word

corpus. Our model extends the target cyberbullying expressions to include phrases which effectively provide a more meaningful context by integrating emotion detection to enhance the accuracy of detection and in particular lower the probability of false positives detection. This will open the door to a more complex and subtle method of cyberbullying identification and handling—enabling greater possibility for the use of non-human SNS/SM content moderation.

B. Structure of the Document

After establishing our main objective, the first part of our study is to create an emotion detection model. Then, we will then use the transfer learning model approach to apply it to our cyberbullying dataset which includes text comments extracted from many social networking sites. Afterwards, we will do a multi-label-classification model to predict both emotion and cyberbullying to create a more context-based model to enhance accuracy of detection and reducing the probability of detecting false positives. Finally, in order to evaluate the effectiveness of our model, we will compare it to a model found in a previous research paper which had the goal to mitigate biases in cyberbullying to avoid misclassifying text containing identity terms for example which might not necessarily indicate cyberbullying (Gencoglu, 2020). We will do this by predicting cyberbullying on an unseen dataset by both models and see the difference in prediction performance. Finally, we will use statistical analysis to evaluate performance and discuss our contribution as well as conclusion on improving cyberbullying expression detection process and product.

CHAPTER II

LITERATURE REVIEW

Several approaches to detect cyberbullying have been put out in scholarly literature. These approaches make use of variables such as user context, gender information, linguistic features, and graph properties. Textual content analysis has been instrumental in enabling the identification of instances of cyberbullying within online social networks. Compared to simple list-based matching of profane words, text analysis offers higher precision and fewer false positives. Consequently, most studies have primarily focused on applying text analysis techniques to online comments.

According to Hinduja & Patchin (2008) and Slonje et al. (2013), cyberbullying is defined as a purposeful, aggressive act committed by an individual or group using electronic communication technologies. This nefarious act usually targets a victim who finds it difficult to adequately protect themselves. Also, the recurring nature of cyberbullying, its reliance on technology for communication, and the inherent power imbalance between the aggressor and the victim are highlighted as the main characteristics of this type of bullying under this definition. Studies show that up to 25% of students report having experienced cyberbullying, a problem that is common among young adults. The detrimental effects on victims are severe, resulting in social isolation, emotional suffering, and difficulties in school. The authors contend that bullies can hide their identities and avoid reprisals because of the anonymity afforded by SNS/SM. Furthermore, victims have difficulties because of the non-ephemeral nature of social media posts, making it harder for them to escape cyberbullying.

The study conducted by Kontostathis et al. (2013) aimed to identify the most frequently used keywords by cyberbullies. They employed the bag-of-words model to evaluate the corpora of cyberbullying. The authors achieved an average precision of 91.25% by creating queries based on these phrases. In another study, Lempa et al. (2015) developed an Android app that employed two techniques for detecting cyberbullying. The first technique involved searching the text for sensitive words and phrases using a brute-force search algorithm. The second technique utilized keyword classification and relevancy matching by collecting seed words. These approaches yielded peak precisions of 89% and 91%, respectively. Huang et al. (2014) focused on identifying cyberbullying through textual analysis of cyber conversation datasets. They investigated whether analyzing features related to cyberbullying improved the accuracy of their detection model. The results indicated that social features played a significant role in detecting cyberbullying, suggesting that understanding the social context in which messages are exchanged is as important as the content itself. Future research can apply similar approaches to more detailed data on human behavior to detect cyber and physical social bullying in various settings, thereby contributing to a safer environment for individuals facing bullying. Addressing noise and errors in social media posts is crucial for effective cyberbullying detection, as emphasized by Zhang et al. (2016). Their proposed solution, a pronunciation-based convolutional neural network combined with keyword-based searching, demonstrates promising results based on their evaluation. Hosseinmardi et al. (2015) conducted a study on cyberbullying specifically on Instagram. They collected a dataset comprising Instagram images and their associated comments and utilized a crowd-sourcing platform to label the images as either cyberbullying or non-cyberbullying. The authors conducted several hypothesis

tests to examine the factors contributing to cyberbullying on Instagram. Their findings revealed that users engaging in cyberbullying were less likely to directly refer to themselves and more likely to refer to others in the third person. Furthermore, cyberbullying posts exhibited higher occurrences of negative emotions and lower occurrences of positive emotions, with a higher probability of cyberbullying found in posts related to religion, death, appearance, and sexual hints.

These studies evaluate the significance of keyword analysis in detecting instances of cyberbullying across different platforms and datasets. By employing textual data analysis, and incorporating context by detecting emotion in cyberbullying text, researchers can attain a more comprehensive understanding of the dynamics related to cyberbullying. Therefore, a deeper understanding of the problem can contribute to the better capabilities for non-human moderation on SNS/SM.

A. The Gap

Different approaches have been developed in the field of cyberbullying detection; these include textual analysis and automated keyword searching and classification. These approaches are particularly useful for locating and addressing cyberbullying incidents. The integration of emotion detection into cyberbullying detection, which serve as a way of understanding the user's engagement in cyberbullying behavior remains an aspect deserving further research attention.

This thesis seeks to investigate the connection between emotion and toxic online expressions. By integrating emotion as a feature for cyberbullying detection, my thesis will go beyond current research on cyberbullying by delving into the underlying emotions that trigger abusive hate expressions, providing a more comprehensive

approach for precise prediction, progressing from simple keyword detection to complex phrasal analysis.

Our ultimate objective is take steps in helping to develop non-human moderation of SNS/SM via ML and analytics methods to reduce incidence of false positives in cyberbullying expression prediction.

1. Thesis Process

The process map is shown in Figure 1, which reflects the tasks performed within the framework of this thesis. The first block deals with data collection, a topic that is further discussed in later parts. The second block then covers the steps involved in pre-processing and cleaning the data. The third and fourth blocks show how emotion will be integrated into transfer learning to improve cyberbullying detection, allowing context-based prediction. The findings from the complete procedure will be discussed in the final section.

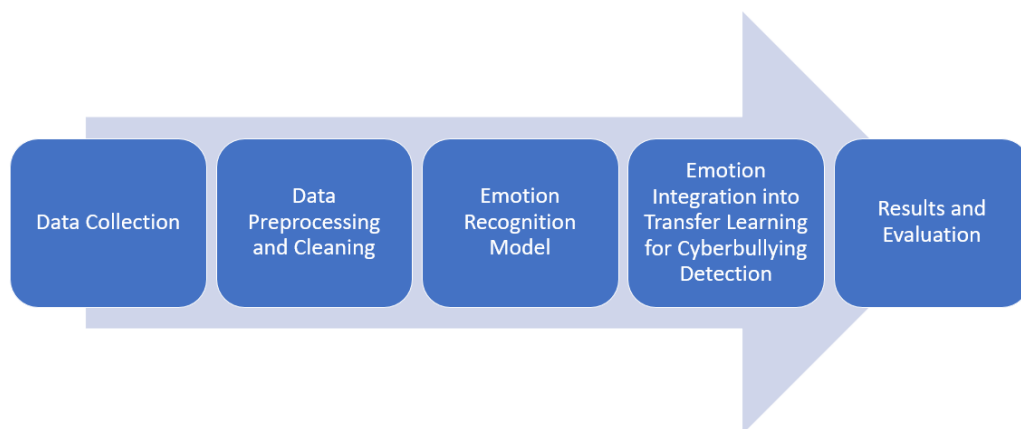


Figure 1 Thesis Process Map

2. Research Model and Hypothesis

I developed a research model illustrated in Figure 2, guiding my exploration of enhancing Social Networking Sites/Social Media (SNS/SM) bullying detection. The methodology emphasizes that analyzing the emotions underlying cyberbullying comments can provide enhanced context to comprehend the motivations behind this harmful behavior. This approach aims to increase detection accuracy of cyberbullying expressions and an enhanced contextual understanding of how better detect cyberbullying expressions.

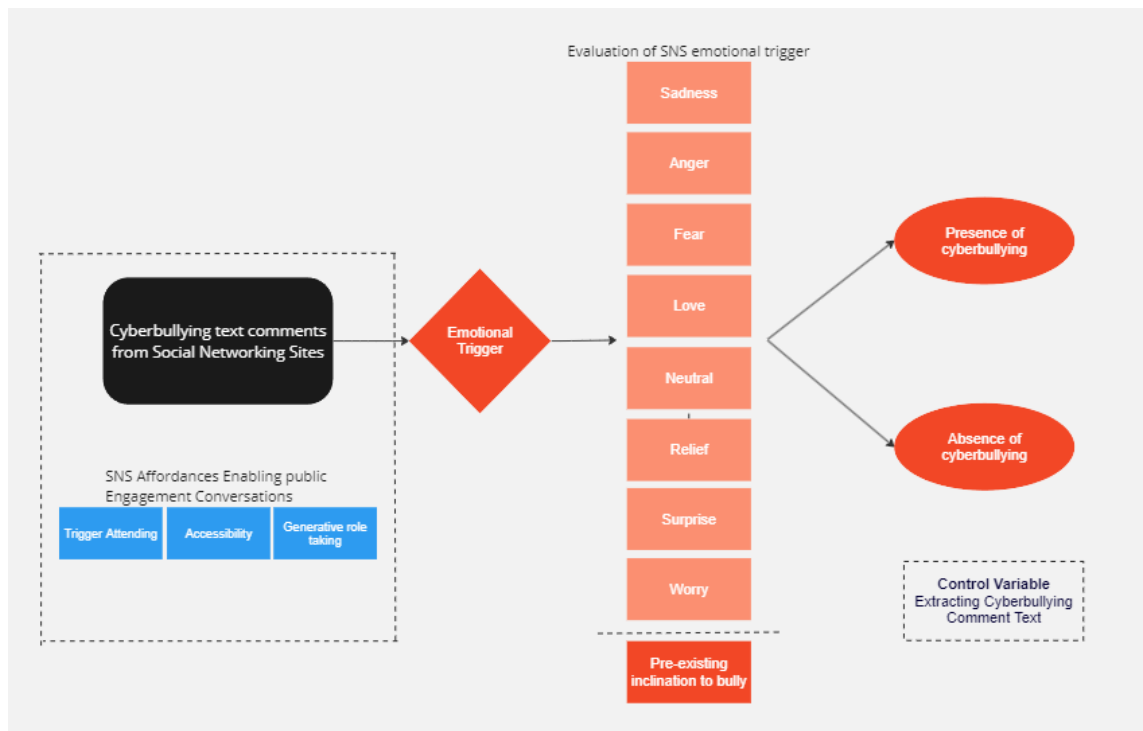


Figure 2 Proposed Research Model

CHAPTER III

DATA COLLECTION METHODOLOGY

A. The Data Collection Process

We started our data collection process by conducting an extensive search on the internet, which allowed us to extract publicly available datasets. These datasets, which we gathered from reliable web sources, serve as a vital basis for our study and guarantee the relevance and diversity of the data we took into account. The thorough compilation of these datasets improves our study's validity and robustness.

1. The Emotional Dataset

The emotion data was extracted from [Kaggle](#) and [Huggingface](#). The Kaggle dataset, which was first collected via data.world, is made up of tweets that have been marked with the appropriate emotions. Three columns make up the dataset: content, sentiment, and tweet_id. The raw tweet text is located in the "content" column, while the emotion connected with each tweet is listed in the "label" column. The Hugging Face dataset is an emotion dataset consisting of English Twitter messages categorized into six basic emotions: anger, fear, joy, love, sadness, and surprise.

	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...
...
39995	1753918954	neutral	@JohnLloydTaylor
39996	1753919001	love	Happy Mothers Day All my love
39997	1753919005	love	Happy Mother's Day to all the mommies out ther...
39998	1753919043	happiness	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...
39999	1753919049	love	@mopedronin bullet train from tokyo the gf ...

40000 rows × 3 columns

Figure 3 Kaggle Emotion Dataset

	text	label
0	i feel awful about it too because it s my job ...	sadness
1	im alone i feel awful	sadness
2	ive probably mentioned this before but i reall...	joy
3	i was feeling a little low few days back	sadness
4	i beleive that i am much more sensitive to oth...	love
...
416804	that was what i felt when i was finally accept...	joy
416805	i take every day as it comes i m just focussin...	fear
416806	i just suddenly feel that everything was fake	sadness
416807	im feeling more eager than ever to claw back w...	joy
416808	i give you plenty of attention even when i fee...	sadness

416809 rows × 2 columns

Figure 4 HuggingFace Emotion Dataset

2. *The Cyberbullying Dataset*

The cyberbullying dataset, extracted from [Kaggle](#) is a compilation of various datasets sourced from different platforms, focusing on the automated identification of cyberbullying incidents. The data is extracted from diverse social media platforms such

as Kaggle, Twitter, Wikipedia Talk pages, and YouTube. Each data entry consists of textual content accompanied by a label indicating whether it corresponds to bullying. The dataset encompasses various forms of cyberbullying, including hate speech, aggression, insults, and toxicity.

	Text	non-cyberbullying	cyberbullying	cyberbullying label	source
0	`- This is not ``creative``. Those are the di...	0.900000	0.100000	0.0	aggression
1	`` the term ``standard model`` is itself le...	1.000000	0.000000	0.0	aggression
2	True or false, the situation as of March 200...	1.000000	0.000000	0.0	aggression
3	Next, maybe you could work on being less cond...	0.555556	0.444444	0.0	aggression
4	This page will need disambiguation.	1.000000	0.000000	0.0	aggression
...
448875	She pretty I love this song I miss the old kel...	NaN	NaN	1.0	youtube
448876	Status-Online Im ZxkillergirizX! I'm Zxkillerg...	NaN	NaN	0.0	youtube
448877	JR so cute EXO M Better I agree like yeah yeah...	NaN	NaN	0.0	youtube
448878	!!	NaN	NaN	0.0	youtube
448879	great video and MERRY CHRISTMAS from greece :*...	NaN	NaN	0.0	youtube

448880 rows x 5 columns

Figure 5 Kaggle Cyberbullying Dataset

B. Data Processing and Analysis

1. Introduction

After acquiring our datasets, we will analyze them using a variety of techniques in order to get insights that support the objectives stated in the thesis. This section's main focus is on providing more detail and clarification on the analysis techniques used to handle data.

An extensive roadmap detailing the many tactics used is shown in Figure 6. Our procedure starts with cleaning and preparing datasets for analysis. Subsequently, we implement an emotion recognition model for transfer learning onto the cyberbullying dataset. The main goal is to improve cyberbullying detection accuracy by integrating emotion context into the learning process.

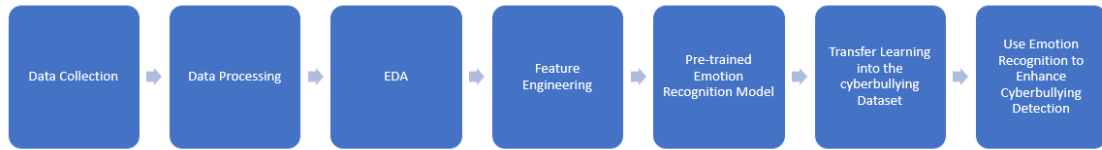


Figure 6 Data Analysis Methodology Process

In this section, we delve into the details of data processing, defining the steps taken to refine and optimize our datasets. We'll focus on the pre-processing measures implemented to clarify the data, enhance clarity, and create more effective groundwork for subsequent analyses.

2. Data Cleaning

Within the field of Natural Language Processing (NLP), preprocessing, often known as text cleaning, becomes a crucial task. It is an important first step that has the ability to change unstructured textual data into a format that is standardized and structured for analysis. This essential procedure lays the groundwork for further analytical efforts by guaranteeing that the textual environment is optimized, normalized, and enhanced for the more complex activities that come within the NLP framework. Within the scope of this thesis, we'll explore the importance of cleaning text for a thorough analysis of textual data.

In our text preprocessing pipeline, we employed a comprehensive approach to enhance the quality and standardization of our textual data. The core of this process lies in lemmatization, a technique facilitated by the WordNetLemmatizer. Lemmatization involves reducing words to their base or root form, ensuring uniformity, and aiding in the extraction of essential meaning from the text.

Additionally, we addressed the distinction of text by eliminating punctuation marks, converting all text to lowercase, and strategically removing common stopwords using an English stopwords list. The preprocessing function, `preprocess_text_column` is shown in Figure 7, refining the textual content for subsequent analysis.

In addressing the challenge of duplicated fields within our dataset, our methodology strategically confronted the issue through a purposeful approach. Notably, a 5% duplication was observed in the emotion dataset. While relatively small in comparison to our extensive dataset, we proactively chose to eliminate duplicates to mitigate potential data leakage. Data leakage occurs when information from the training set unintentionally influences the testing set, potentially leading to overfitting.

It is crucial to highlight that, considering various potential strategies such as aggregating based on the majority class label, prioritizing certain class labels, or even eliminating such duplicates since they are minor, we opted for the strategy of eliminating duplicates. This approach aligns with our commitment to maintaining data integrity and ensuring a robust foundation for subsequent analysis.

This granular text preprocessing lays the groundwork for a more accurate and meaningful exploration of the dataset, positioning our research to extract valuable insights from the refined textual data.

```

def preprocess_text_column(df, text_column_name):
    # Make a copy of the DataFrame to avoid SettingWithCopyWarning
    df_copy = df.copy()

    # Initialize the WordNet Lemmatizer
    lemmatizer = WordNetLemmatizer()

    # Define a regular expression pattern to remove punctuation
    punctuation_pattern = r'^[\w\s]!'

    # Get the English stopwords list
    english_stopwords = set(stopwords.words('english'))

    # Define a function for cleaning, tokenizing, and lemmatizing text
    def preprocess_text(text):
        # Remove punctuation and lowercase text
        text = re.sub(punctuation_pattern, ' ', text.lower())

        # Tokenize the text
        words = text.split()

        # Remove stop words and lemmatize each word
        words = [lemmatizer.lemmatize(word) for word in words if word not in english_stopwords]

        # Join the words back into a sentence
        return ' '.join(words)

    # Apply the preprocessing function to the text column of the copy
    df_copy[text_column_name] = df_copy[text_column_name].apply(preprocess_text)

    return df_copy

```

Figure 7 Data Processing Function

C. Feature Engineering

By extracting more variables from the emotion dataset, we performed feature engineering to improve the capabilities for more sophisticated analysis. These new fields are primarily intended to support the emotion recognition component and the overall objective of enhancing the emotion recognition aspect of cyberbullying detection. The dataset's finalized fields consist of:

1. *Polarity and Sentiment*

In the context of feature engineering for cyberbullying detection, the `get_sentiment` function, as depicted in Figure 8, is employed in processing text using the TextBlob library—a natural language processing tool in Python. The TextBlob library streamlines various text processing tasks, with a notable application in sentiment analysis. The sentiment property of the TextBlob object returns a named tuple with two essential components: polarity and subjectivity. In this specific instance, we focus on

extracting the polarity, a float value ranging from -1 (most negative) to 1 (most positive), thereby quantifying the emotion tone of the text. A polarity of 0 signifies a neutral sentiment.

Subsequently, the function assigns a sentiment label based on the calculated polarity:

- If polarity > 0, the sentiment is labeled as 'positive.'
- If polarity < 0, the sentiment is labeled as 'negative.'
- If polarity = 0, the sentiment is labeled as 'neutral.'

```
## Sentiment Tagging
def get_sentiment(text):
    # Create a TextBlob object for the input text
    blob = TextBlob(text)

    # Get the polarity (-1 to 1) where -1 is negative, 1 is positive, and 0 is neutral
    polarity = blob.sentiment.polarity

    # Determine the sentiment label based on polarity
    if polarity > 0:
        sentiment = 'positive'
    elif polarity < 0:
        sentiment = 'negative'
    else:
        sentiment = 'neutral'

    return sentiment, polarity
```

Figure 8 Feature Creation Functions

The incorporation of sentiment analysis becomes paramount in the broader framework of cyberbullying detection:

- Positive sentiments may indicate positive interactions or supportive content.
- Negative sentiments could be indicative of aggressive language or potentially harmful expressions.
- Neutral sentiments might suggest a lack of discernible emotion tone.

2. countvectorizer

Countvectorizer is a feature engineering technique employed in natural language processing, particularly in the context of textual data analysis. Using this technique, a set of text documents is converted into a matrix of token counts, where each row denotes a document and each column a single word across the corpus. The frequency with which each term appears in the corresponding documents is shown by the values in the matrix. It works through Tokenization which is the process of breaking down the text into individual words or tokens. Next, the Countvectorizer counts how many times each token appears across the whole corpus for every document. Ultimately, the output is a matrix in which every row represents a document, and every column is a distinct term found throughout the dataset. The frequency of each word in the corresponding documents is shown by the values in the matrix.

Countvectorizer is an essential tool for machine learning algorithms that extract features, particularly for tasks like sentiment analysis and text classification. It assists in transforming unprocessed text data into a representation that machine learning models can understand, enabling the algorithms to process textual data efficiently.

In the subsequent sections of this thesis, Countvectorizer is seamlessly integrated into the machine learning algorithm pipelines. Its role within these pipelines is explored in detail, demonstrating its significance in enhancing the overall efficiency and accuracy of the models developed for cyberbullying detection. The utilization of Countvectorizer contributes to capturing the essential textual features needed for robust machine learning based analysis.

By employing Countvectorizer, this thesis harnesses the power of feature engineering to transform raw text data into a structured format that can be effectively

utilized in machine learning algorithms, marking a pivotal step in achieving the goal of enhancing cyberbullying detection through advanced analysis techniques.

3. *TF-IDF*

In natural language processing (NLP), understanding the true meaning of individual words is crucial to deriving meaning from phrases. This is where a crucial feature engineering method called Term Frequency-Inverse Document Frequency (TF-IDF) excels. TF-IDF extends the Countvectorizer architecture by highlighting and exploring the subtle differences in the relative weights of words inside a document when compared to a larger corpus.

The two main components of TF-IDF are term frequency (TF) and inverse document frequency (IDF). TF indicates how frequently a term occurs in a certain document. It emphasizes the word's local importance and is calculated by dividing the word count by the total number of words in the text.

IDF, on the other hand, takes a broader approach, accounting for the word's occurrence across the entire corpus. It serves as a lens through which you may view the uniqueness and specificity of the word in connection to the totality. The IDF is calculated as the logarithm of the ratio of all documents in the corpus to all documents containing the word.

What makes it unique is the combination of TF and IDF (TF-IDF). TF-IDF identify words that are not only frequent within a document but also uncommon across the entire corpus, making it truly representative based on the document.

While Countvectorizer provides word frequencies, TF-IDF calculates TF and IDF independently.

The insights gained from TF-IDF empower various NLP applications, exploring hidden structures and meanings within text language. From pinpointing the most relevant keywords in document classification to uncovering trends in information retrieval, TF-IDF serves as a measure for unlocking the secrets of text. Its diverse applications include:

$$TF(t, d) = \frac{\textit{number of times } t \textit{ appears in } d}{\textit{total number of terms in } d}$$
$$IDF(t) = \log \frac{N}{1 + df}$$
$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$

While both Count Vectorizer and TF-IDF capture word frequencies, TF-IDF provides additional normalization by considering the importance of words across the entire dataset. This can be particularly advantageous when dealing with common words that appear frequently but may not carry significant meaning (e.g., "the," "and"). TF-IDF assigns higher scores to words that are not only frequent in a document but also rare in the overall corpus, thereby highlighting their distinctive importance.

In the subsequent sections of this thesis, TF-IDF is seamlessly integrated into the machine-learning algorithm pipelines. Despite its sophisticated approach, the choice between Countvectorizer and TF-IDF depends on the specific characteristics of the dataset and the objectives of the analysis. The thesis delves into the comparative analysis, showcasing why Countvectorizer with its consideration of term importance, was chosen over TF-IDF, demonstrating its effectiveness in contributing to the enhanced cyberbullying detection framework.

D. Data Exploration

With the dataset preparation completed, we proceed to initiate a data exploration process on the various columns we have synthesized. The analysis focuses on the cyberbullying and emotion text columns. Below is a comprehensive list of the multiple approaches taken, along with additional notes when necessary.

- Ran a count of the unique emotions label of the emotion dataset (Figure 9), the five higher occurring emotions are fear, anger, sadness, love, and joy.
- Ran a count of tweets with less than 30 words (Figure 10), we can definitely notice that the tweets below 11 characters are very low.
- Ran a visualization showing the distribution of tweet lengths for different labels using a KDE plot (Figure 11). The x-axis represents the length of tweets, and the y-axis represents the estimated density. Different colors represent different labels, allowing us to observe how tweet lengths vary across various emotions. We observe that the five most occurring emotions exhibit an average tweet length of around 200 characters, contrasting with other emotions that, on average, have a length of approximately 100 characters. It's important to note the presence of outliers, as evidenced by tweet lengths exceedingly around 450 characters, contributing to the right skewness observed in the plot. We can also here note that we are analyzing phrases and not sentences.
- Ran a word cloud showing words from the provided text data arranged within the shape of the woman (Figure 12) showing the most occurring texts in our corpus. The visualization prominently features the words

"feel" and "feeling," which aligns with expectations given the nature of our emotion dataset. This prevalence suggests that a significant portion of the text revolves around individuals expressing their emotions and feelings.

With the data exploration done, we get a basic set of information regarding our dataset that provides some minor but interesting insights. With this step done, we move on to the core foundation of the thesis, the data analysis.

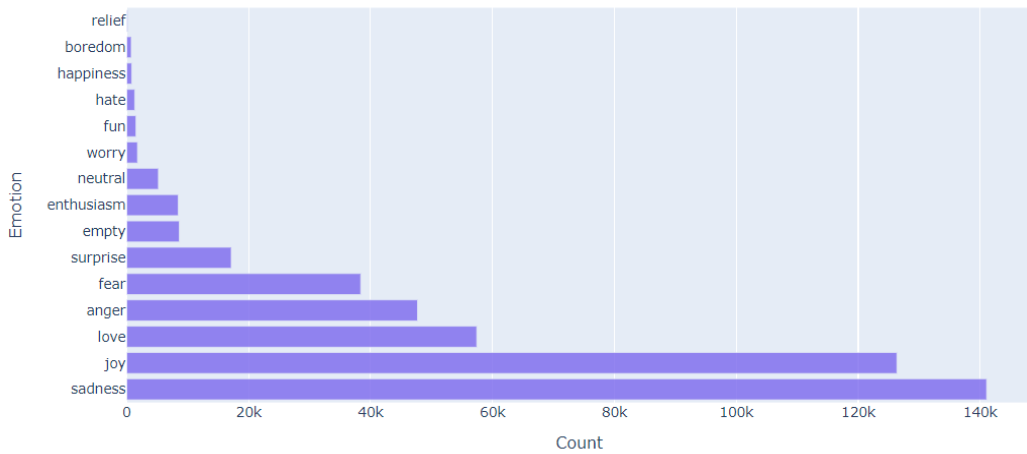


Figure 9 Count of Unique Emotions Labels

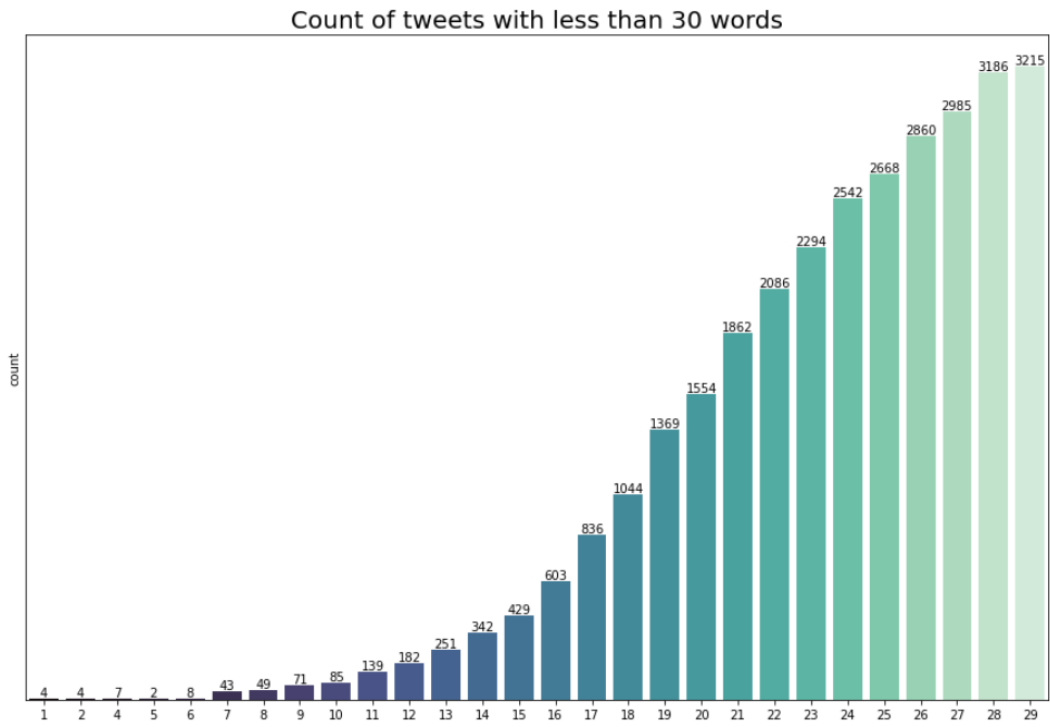


Figure 10 Tweets with less than 30 words

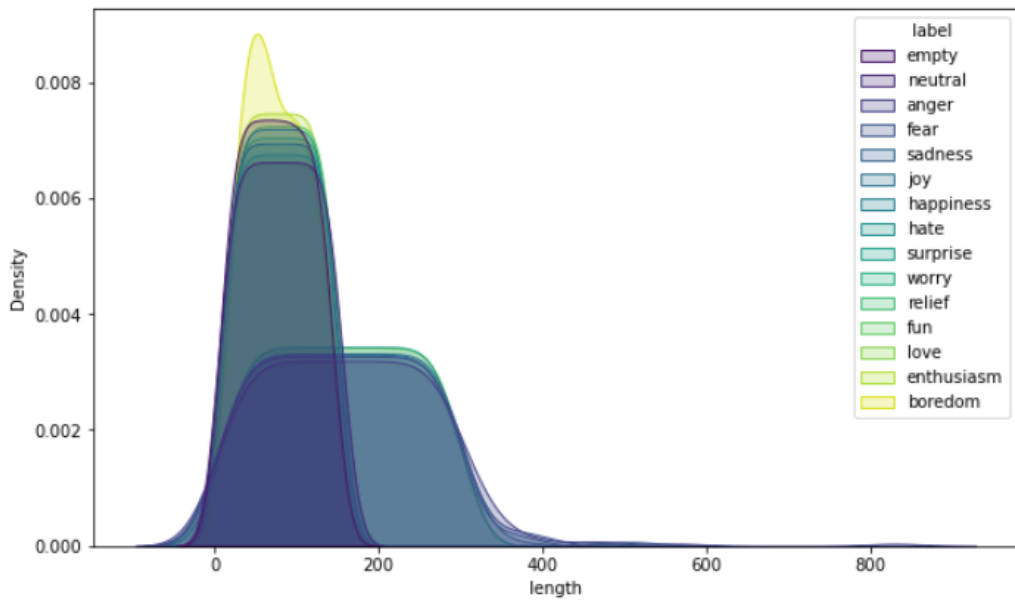


Figure 11 Distribution of tweet lengths for different labels

CHAPTER IV

DATA ANALYSIS

The previously described procedures played a pivotal role in the examination and handling of data related to cyberbullying and emotions, facilitating the attainment of outcomes that were consistent with our main objective. While many strategies were investigated and provided valuable insights, only one directly advances the main goal of the thesis, which is to improve cyberbullying detection via emotion recognition.

Before moving into a more thorough investigation of the mechanisms underlying the improvement of cyberbullying detection by emotion identification, we will give a brief explanation of the strategies used in this section.

- **Logistic Regression:** Predicting the likelihood that an instance will belong to one of two classes is the aim of the statistical technique known as logistic regression. It is not a regression algorithm, despite its name, but a classification one. The logistic (or sigmoid) function is used in logistic regression to convert a linear combination of input characteristics into values between 0 and 1, which indicates the likelihood of the positive class. Gradient descent is frequently used to optimize a logistic loss function in order to train the model. Because of its efficiency, interpretability, and simplicity, logistic regression is frequently used. This makes it especially useful in situations where it is necessary to characterize the relationship between the binary output and the input features. Additionally, it can be extended to handle multi-class classification through techniques like one-vs-rest like in our case.

- **Multinomial Naive Bayes:** Multinomial Naive Bayes is a probabilistic classification algorithm specifically designed for scenarios involving multiple classes. Unlike binary classification, where there are only two possible outcomes, Multinomial Naive Bayes accommodates more than two classes. It is particularly well-suited for text classification tasks, such as document categorization or sentiment analysis. The model is based on Bayes' theorem, making assumptions about the independence of features given the class label, which allows it to efficiently handle high-dimensional data like word frequencies in documents. In the context of multiclass classification, the algorithm estimates the probability of an instance belonging to each class and assigns it to the class with the highest probability. Multinomial Naive Bayes is known for its simplicity, scalability, and effectiveness in handling large and sparse feature spaces, making it a popular choice for text-based problems.
- **Decision Tree:** A Decision Tree is a versatile and intuitive machine learning algorithm used for both classification and regression tasks. It operates by recursively splitting the dataset into subsets based on the most significant features, resulting in a tree-like structure of decision nodes and leaves. In the context of classification, each leaf node represents a class label, and the path from the root to a leaf denotes the decision process. The algorithm selects the best features for splitting by evaluating their ability to maximize information gain or minimize impurity. Decision Trees are advantageous for their transparency, as the resulting model is easily interpretable and can be visualized. They can handle both numerical and categorical features, and their hierarchical structure naturally captures complex decision boundaries. However, they are

prone to overfitting, and strategies like pruning or ensemble methods (e.g., Random Forests) are often employed to enhance generalization performance.

- **Random Forest:** Random Forest is an ensemble learning algorithm that builds a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each decision tree in the forest is constructed based on a random subset of the training data and a random subset of features. This randomness injects diversity into the trees, mitigating overfitting and enhancing the model's robustness. The predictions from multiple trees are then combined to form a more stable and accurate overall prediction. Random Forests excel in handling high-dimensional data, capturing complex relationships, and providing variable importance measures. They are widely used due to their flexibility, scalability, and ability to deliver strong performance across various types of datasets. Additionally, they are less susceptible to overfitting compared to individual decision trees, making them a popular choice in machine learning applications.
- **Support Vector Classification (SVC):** Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. In the context of classification, the Support Vector Classification (SVC) variant is particularly notable. SVC aims to find the optimal hyperplane that best separates data points into different classes. The "support vectors" are the data points that lie closest to the decision boundary, and the margin is the distance between the hyperplane and these support vectors. SVM seeks to maximize this margin, providing a robust classification model. When the data is not linearly separable, SVC can use a kernel trick,

transforming the input space to a higher-dimensional one, making it possible to find a separating hyperplane. SVC is effective in handling complex decision boundaries and is less influenced by outliers. However, it might be sensitive to the choice of hyperparameters and can be computationally intensive with large datasets. Overall, SVC is valued for its ability to generalize well to diverse datasets and perform effectively in various scenarios.

- **Neural Network:** A Neural Network (NN) is a versatile and powerful machine learning model inspired by the structure and function of the human brain. Composed of layers of interconnected nodes, or neurons, neural networks are particularly adept at capturing intricate patterns and relationships within data. The network consists of an input layer, one or more hidden layers, and an output layer. Each connection between nodes is associated with a weight, and each node typically incorporates an activation function, introducing non-linearity to the model. During training, the network adjusts the weights based on the discrepancy between predicted and actual outputs using optimization algorithms such as gradient descent. Neural Networks are highly adaptable and excel in tasks like image recognition, natural language processing, and complex pattern recognition. However, they require substantial computational resources for training and might be susceptible to overfitting if not properly regularized. Despite these challenges, their capacity to learn intricate features and representations makes them a pivotal tool in contemporary machine-learning applications.

A. Emotion Recognition Model

A key component of the research technique in the thesis's data analysis phase is the application of transfer learning. In this case, a pre-trained emotion recognition model is applied to enhance cyberbullying detection. The practice of adapting a previously learned model to a specific objective is known as transfer learning.

The approach involves leveraging a model already proficient in recognizing emotion sentiments from text. By applying this pre-trained emotion recognition model to the cyberbullying dataset, the goal extends beyond mere cyberbullying prediction. The model is designed to discern the emotion triggers and sentiments underlying each phrase, aiming to increase the accuracy of cyberbullying detection by integrating context. This is crucial because certain words, such as "hate," can carry different emotion interpretations based on context. For example, "I hate you" and "I hate thinking about the future" convey distinct emotions—hate and worry, respectively.

The implemented machine learning pipeline plays a key role in achieving this objective. By incorporating techniques like TF-IDF vectorization and logistic regression models, the model is trained to predict both cyberbullying and emotion labels. TF-IDF captures the significance of words in the context of the dataset, while logistic regression models provide predictions for the presence of cyberbullying and associated emotions.

The aim of this approach is to move beyond a binary cyberbullying detection system and delve into the emotion intricacies embedded in the text. By recognizing emotion, the model contributes to a more comprehensive and context-aware cyberbullying detection system. This not only enhances the accuracy of identifying cyberbullying instances but also provides valuable insights into the emotion context, enabling a more comprehensive understanding of online interactions.

In Table 1, the table displays a range of evaluated and executed models. Following this, we will specify the selected model and elucidate the rationale behind our choice.

Table 1 Emotion Recognition Models

Machine Learning Model	Train Accuracy	Test Accuracy
Logistic Regression TF-IDF	93%	83%
Logistic Regression CountVectorizer	98%	73%
Multinomial Naive Bayes TF-IDF	92%	83%
Multinomial Naive Bayes CountVecorizer	89%	84%
Random Forest Word2Vec	96%	37%
Decision Tree CountVecorizer	54%	33%
LSTM	92%	86%
BiLSTM	92%	90%

The presented models underwent not only training but also hyperparameter tuning. In order to improve performance, hyperparameter tuning entails optimizing variables that are not part of the model itself, like learning rates or regularization strengths.

In Figure 13, an example illustrates the hyperparameter tuning process for Multinomial Naive Bayes, resulting in the identification of optimal hyperparameters. For instance, the best hyperparameters for the Multinomial Naive Bayes model were determined to be:

`cv__max_features': 5000`

`cv__ngram_range': (1, 2)`

`nb__alpha': 0.1`

This tuning process is crucial for refining the model's performance and achieving optimal results. The model's behavior before hyperparameter tuning was overfitting. Overfitting occurs when the training accuracy significantly surpasses the testing accuracy. In the initial state, the model exhibited a training accuracy of 89% compared to a testing accuracy of 79%. However, after hyperparameter tuning, the overfitting was mitigated, resulting in a more balanced performance with a training accuracy of 89% and an improved testing accuracy of 84%.

Hyperparameter Tuning Multinomial Naive Bayes

```
# Define the hyperparameters and their possible values
param_grid = {
    'cv__max_features': [1000, 5000], # Adjust these values as needed
    'cv__ngram_range': [(1, 2), (1, 3)], # Adjust these values as needed
    'nb__alpha': [0.1, 0.5], # Adjust these values as needed
}

# Create the grid search with cross-validation
grid_search_nb = GridSearchCV(pipe_nb, param_grid, cv=5, scoring='accuracy')

# Fit the grid search to the training data
grid_search_nb.fit(X_train, y_train)

# Get the best hyperparameters from the grid search
best_params = grid_search_nb.best_params_
print("Best Hyperparameters:", best_params)

# Evaluate the model with the best hyperparameters on the test set
best_model_nb = grid_search_nb.best_estimator_
y_train_pred_best_nb = best_model_nb.predict(X_train)
y_test_pred_best_nb = best_model_nb.predict(X_test)

# Calculate accuracy on the test set using the best model
test_accuracy_best_nb = accuracy_score(y_test, y_test_pred_best_nb)
train_accuracy_best_nb = accuracy_score(y_train, y_train_pred_best_nb)

print("Train Accuracy with Best Hyperparameters:", train_accuracy_best_nb)
print("Test Accuracy with Best Hyperparameters:", test_accuracy_best_nb)

Best Hyperparameters: {'cv__max_features': 5000, 'cv__ngram_range': (1, 2), 'nb__alpha': 0.1}
Train Accuracy with Best Hyperparameters: 0.7502416638187765
Test Accuracy with Best Hyperparameters: 0.8081186197033048
```

Figure 13 Hyperparameter Tuning Process for Multinomial Naïve Bayes

We implemented the BiLSTM (Bidirectional Long Short-Term Memory) model after verifying our findings on the validation dataset because of its exceptional performance. The BiLSTM model not only showcased superior accuracy but also demonstrated efficiency, requiring minimal training time. The decision to opt for the BiLSTM model was grounded in its capacity to capture complex patterns and dependencies in the data, aligning well with the specific requirements of our task. Its performance surpassed other models in terms of accuracy, instilling confidence in its effectiveness for our application. Additionally, the model's efficiency, highlighted by its swift convergence during training, further reinforced our decision. The combined strengths of high accuracy and training efficiency positioned the BiLSTM model as the optimal choice for subsequent stages of our analysis.

In Figure 14, the provided code shows the construction of the Bidirectional Long Short-Term Memory (BiLSTM) neural network model using TensorFlow and Keras for the purpose of classifying various emotions based on text data. The implemented BiLSTM model is trained on a labeled dataset, achieving a notable accuracy of approximately 90% on the test set. The architecture comprises an embedding layer, a bidirectional LSTM layer, and a dense layer tailored for multi-class classification. To address data imbalances, class weights are applied during training, and early stopping is employed to mitigate overfitting. The selection of the BiLSTM model is justified by its adeptness in capturing sequential dependencies within textual information, resulting in superior performance by underscoring the model's achievements in terms of accuracy, efficiency, and its suitability for the emotion classification task.

Data Analysis

BiLSTM Model

```
In [20]: # Define a mapping for your Labels
label_mapping = {'joy': 0, 'love': 1, 'fear': 2, 'anger': 3, 'neutral': 4, 'sadness': 5, 'surprise': 6, 'worry': 7,
                 'hate': 8, 'enthusiasm': 9, 'fun': 10, 'empty': 11, 'relief': 12, 'boredom': 13}

# Convert string Labels to numerical Labels using the mapping
y_train_numeric = np.array([label_mapping[label] for label in y_train])
y_test_numeric = np.array([label_mapping[label] for label in y_test])

# Tokenize the text data
max_words = 5000
tokenizer = Tokenizer(num_words=max_words, oov_token="<OOV>")
tokenizer.fit_on_texts(X_train)

X_train_sequences = tokenizer.texts_to_sequences(X_train)
X_test_sequences = tokenizer.texts_to_sequences(X_test)

# Pad sequences to have the same length
max_sequence_length = max(len(seq) for seq in X_train_sequences)
X_train_padded = pad_sequences(X_train_sequences, maxlen=max_sequence_length, padding='post')
X_test_padded = pad_sequences(X_test_sequences, maxlen=max_sequence_length, padding='post')

# Convert the target Labels to one-hot encoding
y_train_onehot = to_categorical(y_train_numeric)
y_test_onehot = to_categorical(y_test_numeric)

# Build the BiLSTM model
embedding_dim = 128
num_classes = y_train_onehot.shape[1]
model = Sequential()
model.add(Embedding(input_dim=max_words, output_dim=embedding_dim, input_length=max_sequence_length))
model.add(Bidirectional(LSTM(units=100)))
model.add(Dense(units=num_classes, activation='softmax'))

# Compile the BiLSTM model
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Calculate class weights to address class imbalance
class_weights = compute_class_weight('balanced', classes=np.unique(y_train_numeric), y=y_train_numeric)
class_weight_dict = dict(enumerate(class_weights))

# Define early stopping callback
early_stopping = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)

# Train the BiLSTM model
history = model.fit(X_train_padded, y_train_onehot, validation_data=(X_test_padded, y_test_onehot),
                  epochs=20, batch_size=64, class_weight=class_weight_dict, callbacks=[early_stopping])

# Evaluate the BiLSTM model
test_loss, test_accuracy = model.evaluate(X_test_padded, y_test_onehot)
print(f"Test Loss: {test_loss}, Test Accuracy: {test_accuracy}")

# Predict on the test set with the BiLSTM model
y_pred = model.predict(X_test_padded)
y_pred_classes = y_pred.argmax(axis=-1)

# Convert numerical Labels back to string Labels using the reverse mapping
reverse_label_mapping = {v: k for k, v in label_mapping.items()}
y_test_original = np.array([reverse_label_mapping[label] for label in y_test_numeric])
y_pred_original = np.array([reverse_label_mapping[label] for label in y_pred_classes])

# Evaluate the BiLSTM model
test_loss, test_accuracy = model.evaluate(X_test_padded, y_test_onehot)
print(f"Test Loss: {test_loss}, Test Accuracy: {test_accuracy}")

# Print the classification report
print(classification_report(y_test_original, y_pred_original))
```

Figure 14 Construction of Our BiLSTM Neural Network Model

B. transfer learning

In the context of enhancing cyberbullying detection, transfer learning model is employed by utilizing a pre-trained emotion recognition model to predict emotions within a cyberbullying dataset. This technique goes beyond traditional binary cyberbullying detection by recognizing emotion and acknowledging that words can carry distinct meaning based on context. By integrating this pre-trained model into a

machine learning pipeline, the approach simultaneously predicts both cyberbullying and associated emotions, providing a comprehensive understanding of online toxic behavior leading to a better prediction. Transfer learning thus enables the model to leverage prior knowledge of emotion, contributing to a more accurate cyberbullying detection framework.

C. Cyberbullying Detection

After performing transfer learning on the cyberbullying dataset to leverage it as a feature for improving detection, we will delve into the models employed for this enhancement. The processing and feature engineering techniques applied to the emotion dataset were replicated in this context.

We employed multiclass classification models to assess the results. The Table 2 showcases a variety of models that were evaluated and executed. Subsequently, we will identify the chosen model and provide an explanation for the reasoning behind our selection.

Table 2 Enhanced Cyberbullying Detection Models

Machine Learning Model	Train Accuracy	Test Accuracy
Logistic Regression TF-IDF	95% cyberbullying 65% emotion	95% cyberbullying 59% emotion
XGBoost CountVectorizer	94% cyberbullying 64% emotion	94% cyberbullying 64% emotion
Multinomial Naive Bayes TF-IDF	94% cyberbullying 59% emotion	94% cyberbullying 59% emotion
AdaBoost CountVectorizer	94% cyberbullying 47% emotion	94% cyberbullying 47% emotion
Random Forest Tfidf	99% cyberbullying 99% emotion	95% cyberbullying 63% emotion
MNB countvectorizer	94% cyberbullying 89% emotion	94% cyberbullying 87% emotion
LSTM & CNN	80% multi-label classification	81% multi-label classification

For the basic machine learning models, I employed multiclass classification models designed to categorize instances into predefined classes or categories. The model's objective was to accurately classify cyberbullying instances, incorporating emotion aspects. Each instance was assigned to a single class among multiple possible

classes, contributing to the model's effectiveness in discerning and categorizing various forms of cyberbullying based on both content and emotion context.

Then, I discovered an improved approach to achieve my objective by implementing a multi-label classification model designed specifically for tasks where instances may belong to multiple classes simultaneously. In contrast to multiclass classification, where each instance is assigned exclusively to one class among several options, the model I employed is trained to predict multiple labels for each instance, addressing the distinctions of a multi-label classification problem. The implemented neural network model incorporates LSTM and CNN layers to concurrently predict multiple labels. To facilitate this, I transformed the labels into binary format using the `MultiLabelBinarizer`. The model is compiled with binary cross entropy loss and sigmoid activation in the output layer. Notably, the final layer features nodes equal to the total number of classes, with each node indicating the presence or absence of the corresponding class through binary classification. This comprehensive approach enhances the model's ability to handle predicting multiple labels for each instance. Incorporating emotion recognition into the multi-label classification model enhances cyberbullying detection by providing a detailed understanding of textual content. The integration of LSTM and CNN layers enables the model to concurrently analyze the emotion and contextual aspects of communication, improving its capability to predict multiple labels that capture both cyberbullying categories and associated emotion variations.

Figure 15 shows the provided code of the neural network model for multi-label text classification using a combination of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) layers. The model is trained over 50 epochs on a

labeled dataset, with training and validation accuracy reported after each epoch. The training process exhibits an accuracy of 80%. The test accuracy is reported as 60%, indicating that the model's performance on unseen data is relatively good. The training time for each epoch is considerable, likely due to the complexity of the model architecture and the large dataset.

LSTM & CNN

```

# Assuming you have already imported and processed your data (X_train, X_test, y_train, y_test)

# Convert labels to strings
y_train['cyberbullying'] = y_train['cyberbullying'].astype(str)
y_train['Emotion_Label'] = y_train['Emotion_Label'].astype(str)

y_test['cyberbullying'] = y_test['cyberbullying'].astype(str)
y_test['Emotion_Label'] = y_test['Emotion_Label'].astype(str)

# Tokenize and pad sequences
max_words = 10000 # You can adjust this based on your data
max_len = 100 # You can adjust this based on your data

tokenizer = Tokenizer(num_words=max_words)
tokenizer.fit_on_texts(X_train)

X_train_seq = tokenizer.texts_to_sequences(X_train)
X_test_seq = tokenizer.texts_to_sequences(X_test)

X_train_padded = pad_sequences(X_train_seq, maxlen=max_len, padding='post')
X_test_padded = pad_sequences(X_test_seq, maxlen=max_len, padding='post')

# Convert labels to binary format
mlb = MultiLabelBinarizer()
y_train_binary = mlb.fit_transform(y_train.values)
y_test_binary = mlb.transform(y_test.values)

# Build the neural network model (LSTM + CNN)
num_classes = len(mlb.classes_)
model = Sequential([
    Embedding(input_dim=max_words, output_dim=128, input_length=max_len),
    Bidirectional(LSTM(64, return_sequences=True, dropout=0.2, recurrent_dropout=0.2)),
    Conv1D(128, 5, activation='relu'),
    GlobalMaxPooling1D(),
    Dense(64, activation='relu'),
    Dropout(0.5),
    Dense(num_classes, activation='sigmoid')
])

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Define early stopping
early_stopping = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)

# Train the model with early stopping
model.fit(X_train_padded, y_train_binary, epochs=50, batch_size=32, validation_split=0.1, callbacks=[early_stopping])

# Evaluate the model on the test set
y_pred = model.predict(X_test_padded)
y_pred_binary = (y_pred > 0.5).astype(int)

# Convert predictions back to original labels
y_pred_labels = mlb.inverse_transform(y_pred_binary)

# Evaluate the model on the test set
y_test_binary = mlb.transform(y_test.values)
accuracy = accuracy_score(y_test_binary, y_pred_binary)
print(f"Test Accuracy: {accuracy}")

```

Figure 15 Construction of our LSTM & CNN Model

CHAPTER V

EVALUATION AND RESULTS

Previously, we delved into the inner workings of the techniques we employed. Now, we shift our focus to discussing the outcomes and insights derived from the experiment. We will begin by exploring the evaluation metrics used in our research. Subsequently, we'll address the findings by comparing our model to another research model from (Gencoglu, 2021). Finally, we will delve into the analysis results and our contribution.

A. Evaluation Metrics

We examine the crucial procedure of evaluating the efficacy and performance of the created models in the thesis section on assessment metrics. Selecting the right assessment criteria is critical to determining how effectively the models achieve their stated goals. This section provides a thorough examination of the approaches used to assess the models' performance, predictive power, and overall dependability. We seek to offer a comprehensive view of the models' strengths and limits by closely examining a wide range of metrics being recall, F1 score, accuracy and precision. Insights into the behavior of the models in diverse settings can only be obtained through this thorough evaluation procedure, which also guarantees the reliability and validity of the suggested fixes.

- **F1-Score:** The harmonic mean of precision and recall. The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the

worst possible F1 score. It is a useful metric when we want to consider both false positives and false negatives and seek a balance between precision and recall.

$$F1 = 2 \times [(Precision \times Recall)/(Precision + Recall)]$$

- **Precision:** The number of true positive predictions divided by the total number of positive predictions. It measures the accuracy of the positive predictions and is also known as the positive predictive value.

$$Precision = True\ Positives / (False\ Positives + True\ Positives)$$

- **Recall:** The number of true positive predictions divided by the total number of actual positives. It measures the ability of the classifier to capture all the positive instances.

$$Recall = True\ Positives / (False\ Negatives + True\ Positives)$$

- **Receiver Operating Characteristic Curve (AUC-ROC):** The Receiver Operating Characteristic (ROC) curve is a graphical representation of a binary classification model's performance across various classification thresholds. It plots the trade-off between the true positive rate (sensitivity) and the false positive rate, allowing for an assessment of the model's ability to discriminate between positive and negative instances. The curve is generated by adjusting the classification threshold, and each point on the curve represents a different balance between correctly identifying positive instances and incorrectly classifying negative instances. The Area Under the ROC Curve (AUC-ROC) is a summary metric that quantifies the overall performance of the model, with higher values indicating better discrimination. An AUC-ROC of 0.5 corresponds

to random chance, while a value of 1 signifies perfect predictions. This evaluation method is especially valuable in scenarios with imbalanced class distributions or varying costs associated with false positives and false negatives.

In the realm of my thesis, when delving into emotion recognition, I opted against utilizing the Receiver Operating Characteristic (ROC) curve for evaluation. Given the intricacies associated with extending ROC to multiclass classification scenarios, I found it to be less intuitive and potentially confusing for interpreting results. Instead, I leaned toward more interpretable metrics like F1-score, recall, and precision. These metrics offered a clearer understanding of my model's performance across various emotion states. However, when transitioning to the domain of cyberbullying detection in my thesis, the ROC curve took center stage. In this context, where the primary focus is often on binary outcomes—identifying instances of bullying or non-bullying—the ROC curve serves as a valuable tool. Its ability to illustrate the trade-offs between true positive and false positive rates proved especially useful, considering the difference between false positives and false negatives. Thus, the choice of evaluation metrics emerged as a strategic decision, aligning with the specific demands of each classification task within my research.

B. Model Comparison

In the comparative analysis phase of my thesis, I conducted an assessment of the accuracy of the chosen models to measure the impact of my contributions. In addition, I sought to compare my model's performance with that of a notable research paper (Gencoglu, 2021), which aimed to enhance cyberbullying detection through the integration of fairness constraints. The research paper posed a crucial question: 'Can we

mitigate the unintended bias of cyberbullying detection models by guiding the model training with fairness constraints?' The authors proposed a model training scheme designed to employ fairness constraints and validated their approach across various datasets. Their findings demonstrated that different types of unintended biases could be successfully mitigated without compromising the overall model quality. This research significantly contributes to the ongoing pursuit of unbiased, transparent, and ethically grounded machine learning solutions for cyber-social health.

We will now delve into the statistical tests employed and the results obtained as part of my contribution. This analysis aims to provide a rigorous evaluation of the effectiveness and significance of the enhancements introduced to the cyberbullying detection models. Through meticulous examination, we seek to elucidate the impact and validity of the implemented changes in comparison to established benchmarks and evaluate the overall statistical robustness of the proposed contributions.

C. Statistical Tests Results

After extensive research, I chose to employ the Paired T-test and McNemar's test for the comprehensive evaluation and comparison of my proposed enhancements against the established results. The evaluation process involved running both models on an unseen dataset, comparing actual versus predicted results to calculate accuracy. Subsequently, I conducted paired T-tests and McNemar's tests to assess the statistical significance and determine whether there is a substantial difference in performance. These tests aim to rigorously validate the impact of my contributions and provide meaningful insights into the effectiveness of the proposed enhancements in comparison to existing results.

The paired T-test assesses the statistical significance of mean differences in model performance metrics, offering insights into whether the observed distinctions are meaningful. McNemar's test, tailored for binary outcomes in paired data, providing a focused evaluation of misclassification rates, aiding in the determination of significant differences in the models' abilities to correctly classify instances. These tests collectively contribute to a comprehensive comparative analysis, enabling robust statistical validation of the effectiveness of proposed enhancements and aiding in informed decision-making regarding model selection and improvement strategies.

The presented results from the paired T-test and McNemar's test offer critical insights into the comparative analysis between (Gencoglu, 2021) model and our proposed model for cyberbullying detection. The accuracy figures indicate a substantial 18% improvement in my model having a 66% accuracy compared to (Gencoglu, 2021) having a 48% accuracy, signifying the potential effectiveness of my contributions. The paired T-test result with a t-statistic of 17.27 and an exceedingly small p-value of $5.01e-66$ suggesting that there is a significant difference in the mean performance between the two models, providing statistical support for the superiority of my model. Furthermore, McNemar's test, reflected in the chi-squared statistic of 1193.9 and a p-value of $1.42e-36$, indicates that there is also a significant difference in predictions, emphasizing the robustness of my model's classification capabilities. These results collectively strengthen the argument for the efficacy of our enhancements, reinforcing the importance and impact of our contributions to the field of cyberbullying detection.


```
# Print the accuracy scores
print("IEEE Model:", ieee_model_accuracy)
print("My model:", mymodel_accuracy)
```

```
IEEE Model: 0.48414414414414414
My model: 0.6337837837837837
```

Figure 16 Models Accuracy on Unseen Data

Paired T-test

```
from scipy import stats

# data for Model 1 and Model 2
model1_predictions = fairnesscyberbullying_pred['fairnesscyberbullying_pred']
model2_predictions = new_data['CB_pred']
true_labels = fairnesscyberbullying_pred['CB_Label']

# Calculate the differences in accuracy between Model 1 and Model 2
differences = ieee_model_accuracy - mymodel_accuracy

# Perform the Paired T-Test
t_statistic, p_value = stats.ttest_rel(model1_predictions, model2_predictions)

# significance level (alpha)
alpha = 0.05

# Print the Paired T-Test results
print("Paired T-Test Results:")
print(f"t-statistic: {t_statistic}")
print(f"p-value: {p_value}")

# Check if the p-value is less than the significance level
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant difference in performance.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference in performance.")
```

```
Paired T-Test Results:
t-statistic: 17.27804757222069
p-value: 5.012677939285059e-66
Reject the null hypothesis: There is a significant difference in performance.
```

Figure 17 Paired T-test Implementation and Results

McNemar's Test

```
from sklearn.metrics import confusion_matrix
from scipy.stats import chi2_contingency

model1_predictions = model1_predictions
model2_predictions = model2_predictions
true_labels = new_data['CB_Label']

# Convert predictions to binary
threshold = 0.5
model1_binary = (model1_predictions > threshold).astype(int)
model2_binary = (model2_predictions > threshold).astype(int)

# Create a confusion matrix for the new data
conf_matrix = confusion_matrix(true_labels, model2_binary) # Fix: Use new_model2_binary here

# Extract values from the new confusion matrix
a = conf_matrix[0, 0]
b = conf_matrix[0, 1]
c = conf_matrix[1, 0]
d = conf_matrix[1, 1]

# Perform McNemar's test for the new data
statistic = ((b - c) ** 2) / (b + c)
p_value = chi2_contingency([[b, c], [d, a]])[1]

# Set your chosen significance level (alpha)
alpha = 0.05

# Print McNemar's test results for the new data
print("McNemar's Test Results:")
print(f"Chi-squared statistic: {statistic}")
print(f"p-value: {p_value}")

# Check if the p-value is less than the significance level
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant difference in predictions.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference in predictions.")

McNemar's Test Results:
Chi-squared statistic: 1193.90135301353
p-value: 1.4233898601412619e-36
Reject the null hypothesis: There is a significant difference in predictions.
```

Figure 18 McNemar's Test Implementation and Results

CHAPTER VI

CONCLUSION

A. Research Summary

In the last two decades, the penetration of Social Networking Sites and Social Media (SNS/SM) platforms has risen to include more than one-third of the global population. However, the use of SNS/SM has produced both positive and negative results so much so that there have been calls to researchers pay immediate and far greater attention to these contradictory effects of SNS/SM capabilities (Majcrack et al 2013). A key negative impact is the fast and significant rise of cyberbullying and the commensurate need for moderating contents on (SNS/SM) (Colantes et al., 2020). It is also recognized that human content moderation on SNS/SM is impractical and too costly. Viewing technology not as a fixed object but as an interactive element within a system, shaping and being shaped by human actions and social contexts (Faraj & Azad, 2012), is crucial for understanding complex phenomena like cyberbullying in social networking sites. This interactive view is essential because, for example, anonymity and ease of content sharing on SNS/SM can create fertile ground for cyberbullying behavior. Therefore, there is an increasing need for accurate methods of content moderation that are less reliant on human judgement and instead employ sophisticated machine learning methods. A key shortcoming of the current machine learning approaches to cyberbullying detection is that their accuracy needs to be improved significantly to be relied upon for practical deployment on SNS/SM.

In this thesis, we have proposed a robust framework to enhance cyberbullying detection by integrating a Bidirectional Long Short-Term Memory (BiLSTM) transfer

learning emotion recognition model into the process. Going from conventional models that rely on the identification of specific abusive keywords, our approach offers a more comprehensive analysis of complex phrases, enabling a more accurate understanding of potential cyberbullying instances in SNS/SM content. Past research, e.g., the work by Kontostathis et al. (2013) and Huang et al. (2014), has achieved improved accuracy through keyword searching. However, our model addresses some of the key limitations of such approaches, specifically that cyberbullying methods that rely solely on single words may not always identify occurrence of cyberbullying due to high incidence of false positives.

Our study began by acquiring an emotion dataset, followed by thorough pre-processing, cleaning, and feature engineering to construct a relatively more robust BiLSTM model. This model, after successful training, was transferred to an unseen cyberbullying dataset, serving as the foundation for future deployment of the LSTM and CNN aspects aimed at improved detection of cyberbullying and emotion. This multi-label classification model was specifically designed and helped to relatively more accurately classify cyberbullying instances while incorporating emotion aspects, by demonstrating an accuracy of 80% on the training set.

B. Research Contribution

The primary contribution of this research is in taking initial steps to improve the prediction capability of cyberbullying expression within SNS/SM text by employing a multi-label LSTM and CNN classification model, achieving an 80% accuracy on the training set. More specifically the model improved detection accuracy from 48% to 66% (or 18%) based on an unseen dataset in comparison to the model propose by

(Gencoglu, 2021)³. In order to quantitatively assess the significance of the difference between this thesis' model and comparable base I performed, statistical analysis, including the paired t-test and McNemar's test, which demonstrated there is significant difference in performance of two models, with BiLSTM model performing better. In summary, this research has introduced a practical and effective approach to improved cyberbullying prediction, by emphasizing the importance of analyzing complex phrases for more accurate detection. In addition, this research has addressed the limitations inherent in keyword-based methodologies, thereby making a meaningful contribution to the field of cyberbullying research.

C. Research Limitations

It is important to highlight some limitations of this research. First, the potential biases within the emotion dataset present a concern, potentially impacting the model's effectiveness. Second, the computational resource requirements for trying to implement a BERT model pose practical challenges, potentially limiting the improved accuracy of the proposed framework. The need for a lot of computer power might make it hard for many people to use, especially those with less access to strong computers or in places where resources are limited. Third, another limitation pertains to the model's generalization to diverse textual data. The extent to which the model can effectively extrapolate beyond the specific content types used for training remains an issue to be addressed. The training data predominantly consists of specific linguistic styles, cultural contexts, or cyberbullying instances, such as attack, sexism, racism, and toxicity.

³ the latter was focused on mitigating unintended bias in cyberbullying detection through fairness constraints.

Therefore, the model may have difficulty to adapt to a broader range of cases with varying language intricacies and cultural distinctions. Finally, the risk of overfitting to a narrow subset of data, especially one dominated by such sensitive and specific content, poses challenges to the model's performance when faced with the rich diversity and evolving language trends inherent in real-world cyberbullying situations.

D. Future Research Directions

Future research in this domain should focus on mitigating the impact of emotion bias within the training dataset. Investigating strategies to enhance the diversity and inclusivity of emotion expressions is crucial for developing a more robust cyberbullying detection model. Researchers could explore collecting emotion data from a wide range of sources, demographics, and cultural contexts to create a more representative dataset. Additionally, incorporating mechanisms to detect and address biases within the emotion dataset during model training could be a promising avenue. By promoting a more balanced emotion dataset, future models can strive to better capture the diverse emotion expressions inherent in cyberbullying instances.

Furthermore, addressing the computational resource challenges associated with BERT model implementation is essential for increasing the accuracy of the proposed framework. Future research could explore optimizing existing models or developing new architectures that maintain high performance while reducing computational demands. Finally, future work should delve into working on diverse data types, including exploring images and videos related to cyberbullying. The goal is to develop models that excel in real-world scenarios, demonstrating adaptability to the ever-evolving landscape of cyberbullying expressions across different contexts and

demographics, with a specific emphasis on integrating visual and auditory dimensions present in images and videos. This targeted exploration of multimedia content would contribute to a more precise cyberbullying detection framework capable of handling the diverse forms in which cyberbullying occurs.

APPENDIX

APPENDIX I

MODEL COMPARISON

```
# Example sentence
sentence = ["I hate when I think about you"]

# Preprocess the sentence for prediction
processed_sentence = get_bert_embeddings(sentence)

# Make predictions
prediction_class = constrained_model.predict_classes(processed_sentence, batch_size=cf.hyperparams['batch_size'])[0]
prediction_probs = constrained_model.predict(processed_sentence, batch_size=cf.hyperparams['batch_size'])[0]

# Interpret the result
predicted_class_label = "cyberbully" if prediction_class == 1 else "non-cyberbully"

# Display the results
print("Predicted Class Label:", predicted_class_label)
print("Prediction Probabilities:", prediction_probs)

✓ 3.3s

Predicted Class Label: cyberbully
Prediction Probabilities: [0.7330631]
```

Figure 19 “I hate when I think about you” (Gencoglu, 2021) Model Prediction

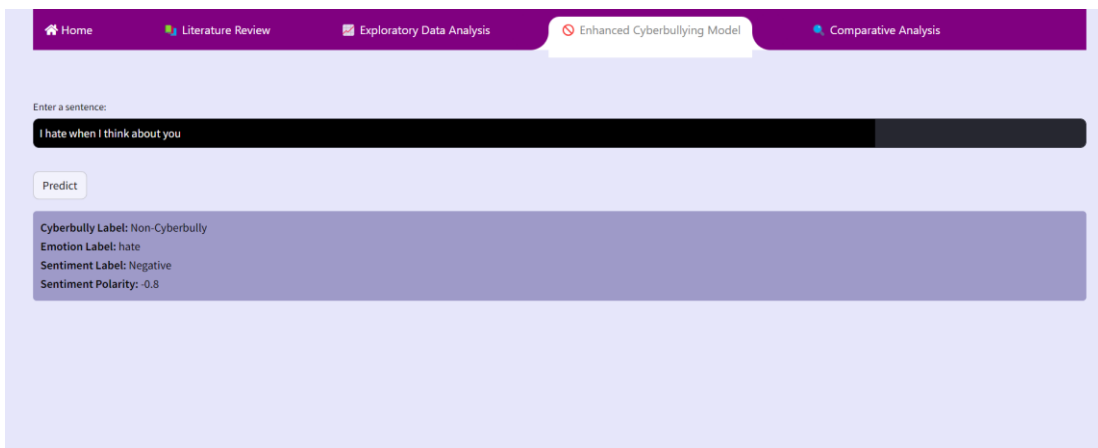


Figure 20 “I hate when I think about you” Enhanced Model Prediction


```
# Example sentence
sentence = ["what you are doing is bad"]

# Preprocess the sentence for prediction
processed_sentence = get_bert_embeddings(sentence)

# Make predictions
prediction_class = constrained_model.predict_classes(processed_sentence, batch_size=cf.hyperparams['batch_size'])[0]
prediction_probs = constrained_model.predict(processed_sentence, batch_size=cf.hyperparams['batch_size'])[0]

# Interpret the result
predicted_class_label = "cyberbully" if prediction_class == 1 else "non-cyberbully"

# Display the results
print("Predicted Class Label:", predicted_class_label)
print("Prediction Probabilities:", prediction_probs)

✓ 2.7s

Predicted Class Label: cyberbully
Prediction Probabilities: [0.8071867]
```

Figure 21 “What you are doing is bad” (Gencoglu, 2021) Model Prediction

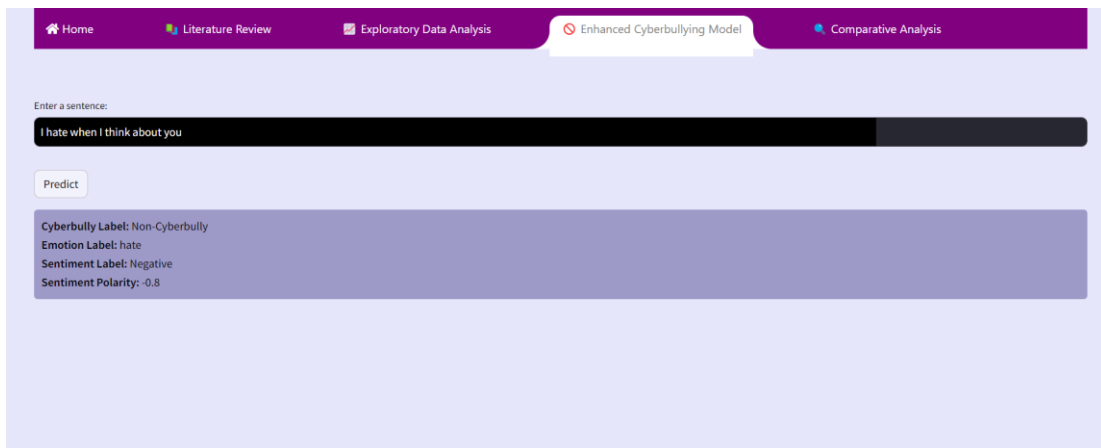


Figure 20 “What you are doing is bad” Enhanced Model Prediction

APPENDIX II

CLUSTERING

Before delving into the emotion analysis, I conducted clustering analysis on my cyberbullying dataset to extract valuable insights. The following visuals provide a clear illustration of the results obtained during this process.

- Figure 23 indicates that the optimal K value is 2.
- Figures 24, 25, and 26 further support the choice of $K = 2$, as evidenced by the silhouette scores revealing distinct clusters.
- Figure 27 clarifies that these two identified clusters correspond to toxicity and aggressivity in the cyberbullying dataset.

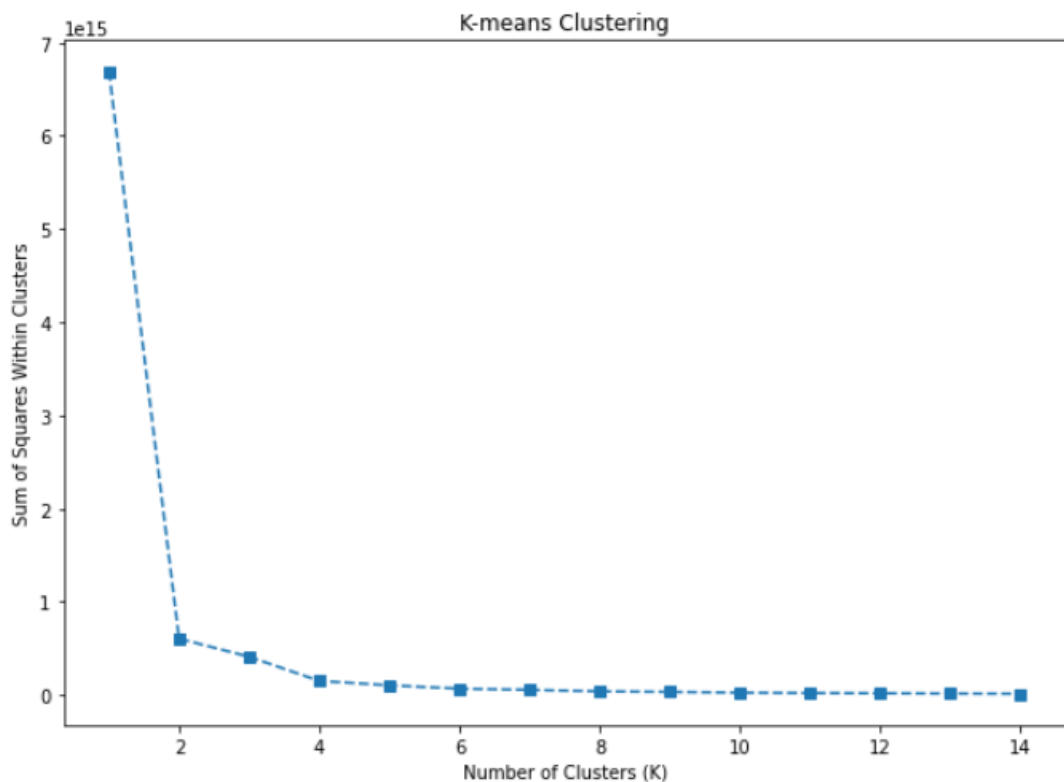


Figure 21 Optimal K Determination

Silhouette analysis for KMeans clustering with n_clusters = 2

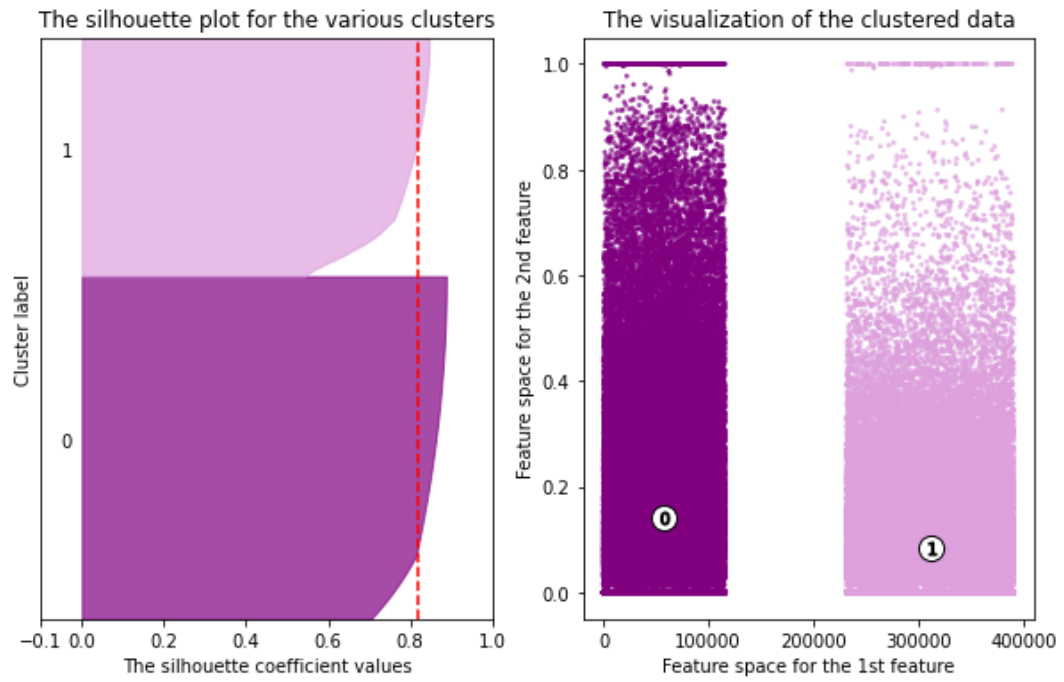


Figure 22 Silhouette Score for K=2

Silhouette analysis for KMeans clustering with n_clusters = 3

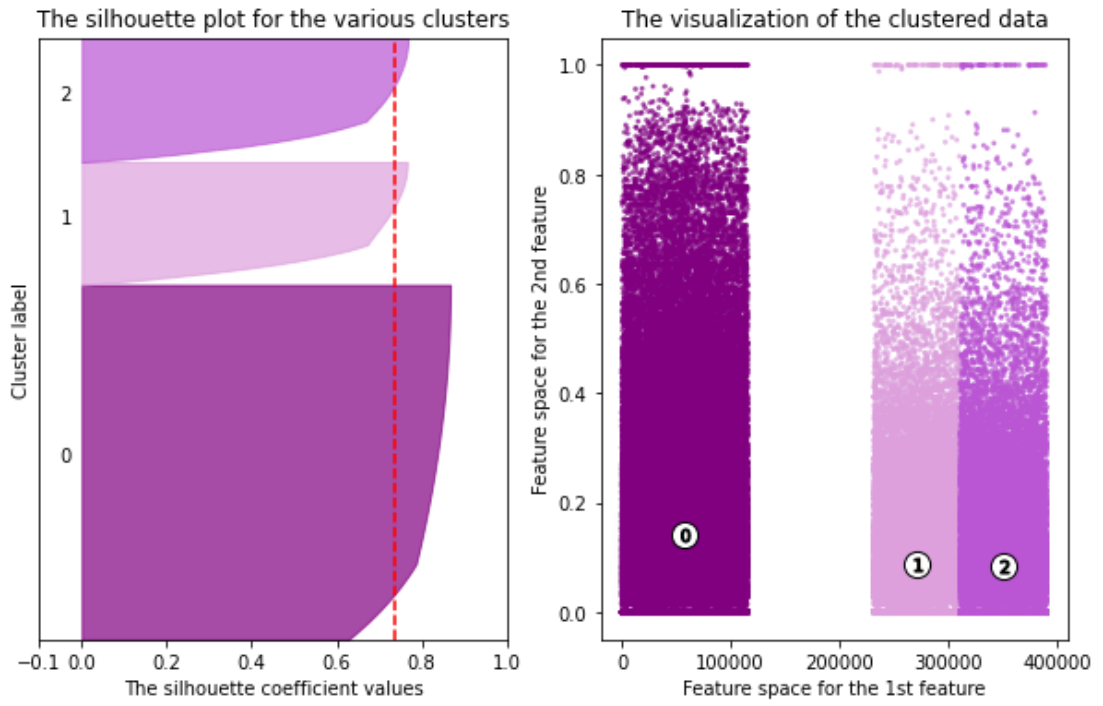


Figure 23 Silhouette Score for K=3

Silhouette analysis for KMeans clustering with n_clusters = 4

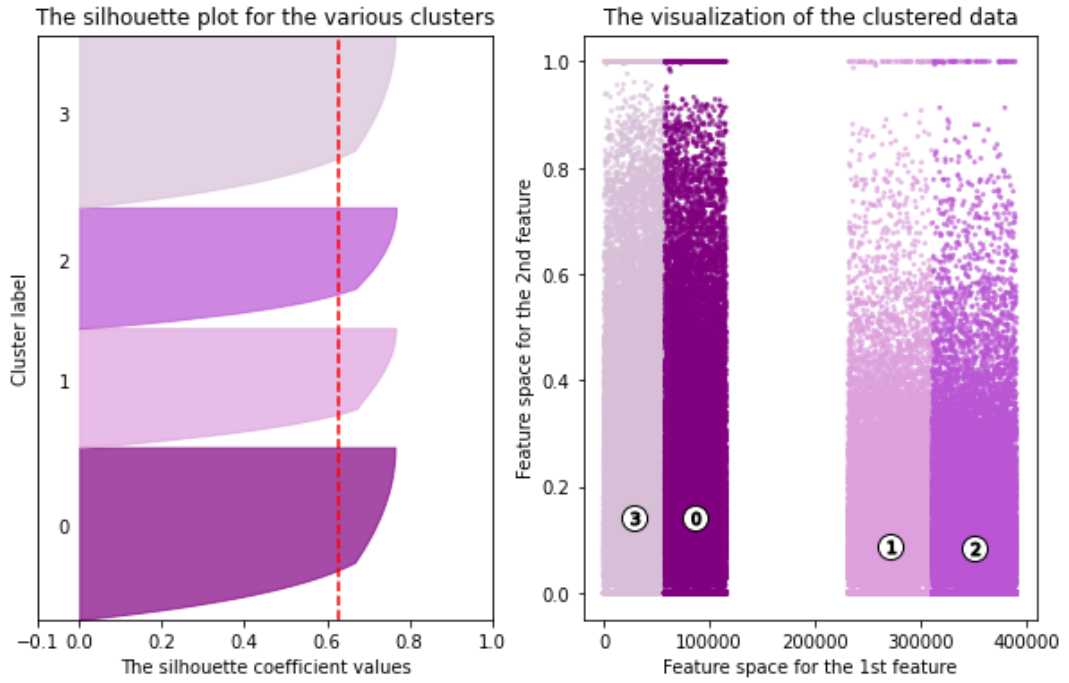


Figure 24 Silhouette Score for K=4

Cyberbullying Behavior

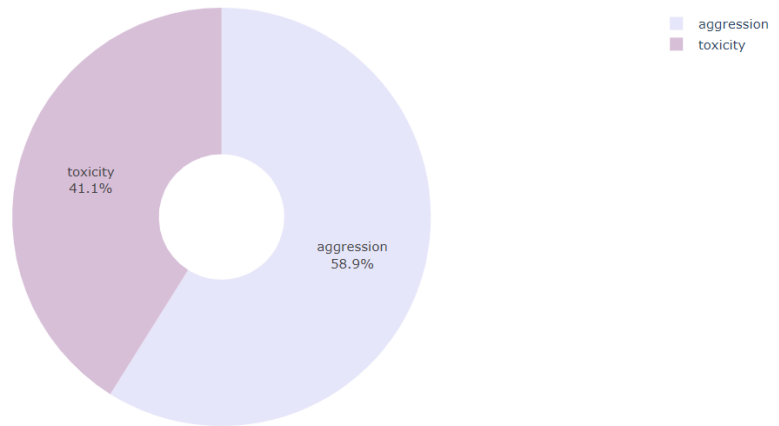


Figure 25 Identified Clusters as Toxicity and Aggressivity

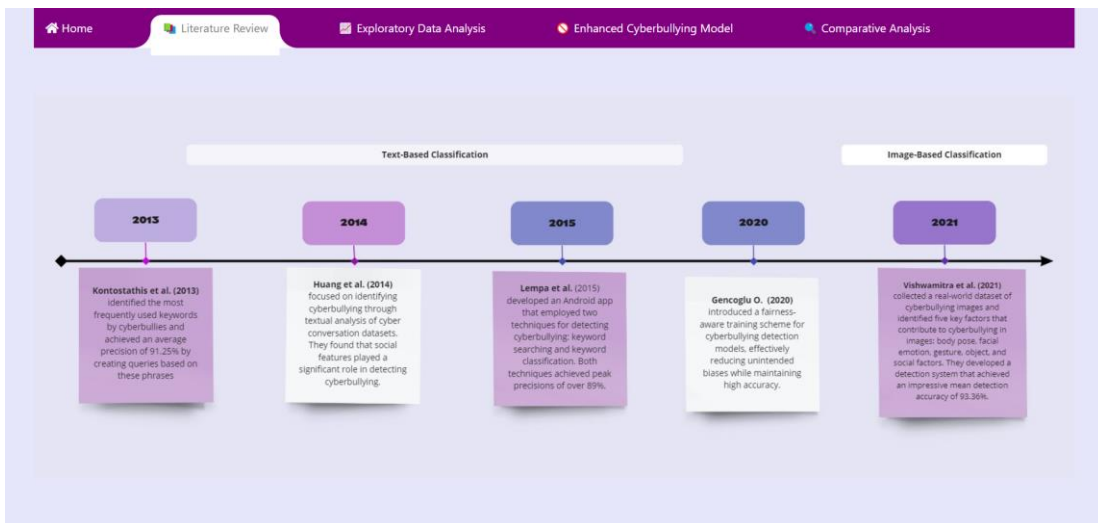
APPENDIX III

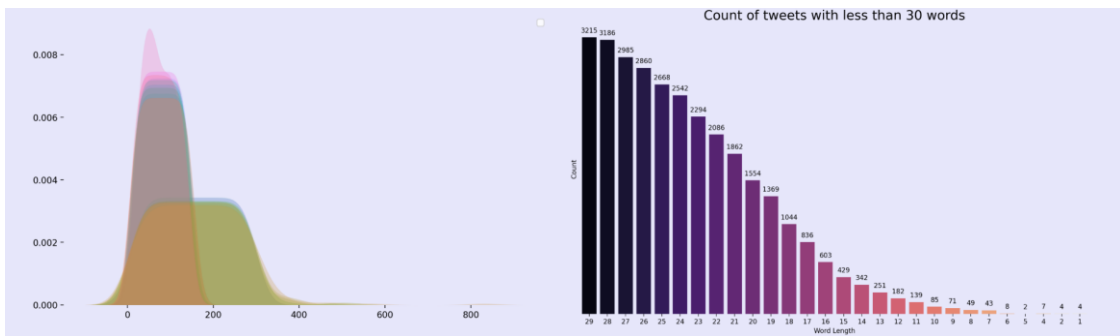
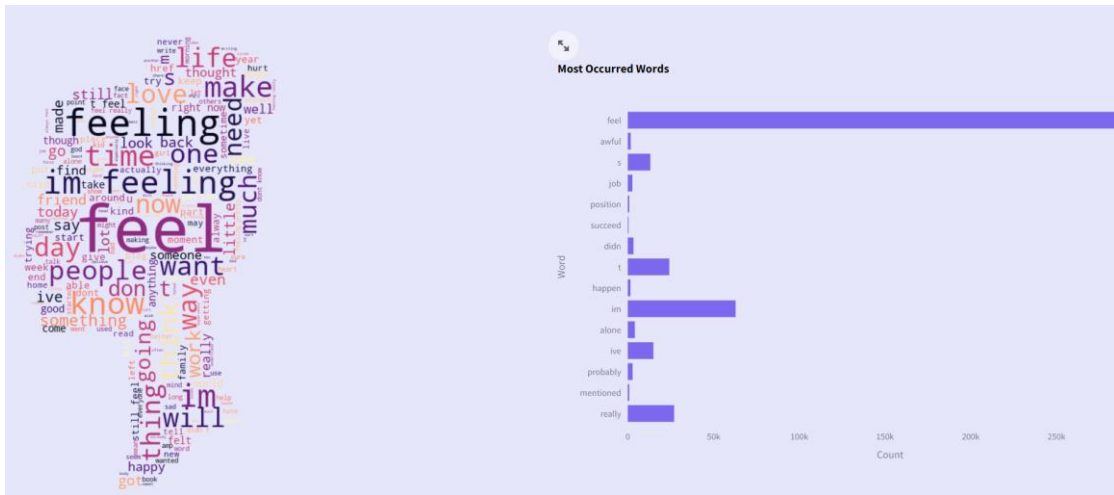
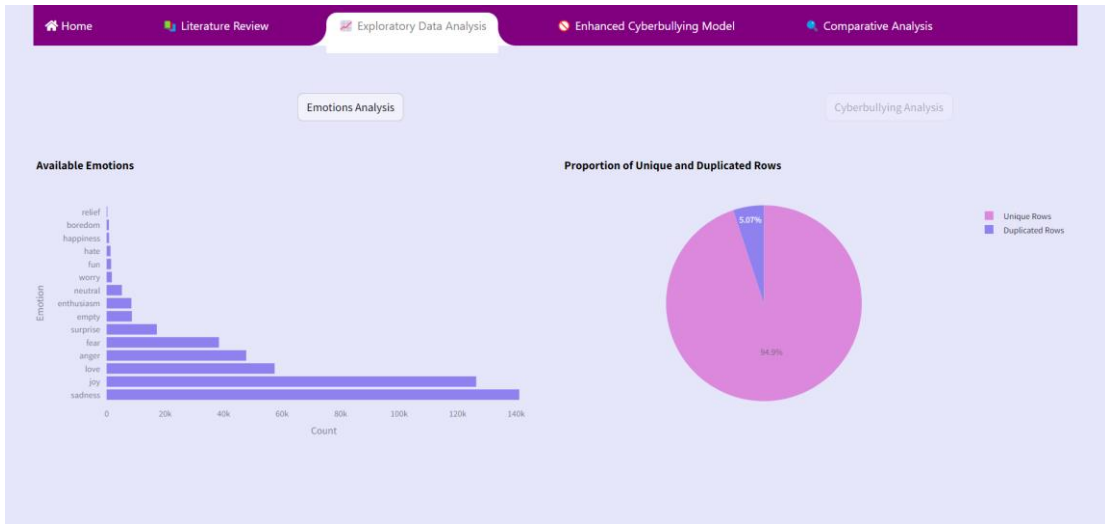
STREAMLIT APP

The screenshot shows a Streamlit web application interface. At the top, there is a navigation bar with five tabs: Home, Literature Review, Exploratory Data Analysis, Enhanced Cyberbullying Model, and Comparative Analysis. The main content area features a title and a detailed abstract. The abstract discusses the role of emotions in cyberbullying on social media, the goal of developing a BiLSTM-based model for detection, and the aspiration to create a safer digital environment. The background of the page includes an illustration of a person sitting at a laptop with speech bubbles containing social media-related icons and text like '#Sm'.

Leveraging a BiLSTM-based Emotion Recognition Transfer Learning Model for complex phrasal analysis in cyberbullying detection

This research aims to investigate the role of emotions, particularly those triggered by images or text, in the context of cyberbullying on social media platforms. The study seeks to understand how users harness emotional cues as triggers for cyberbullying incidents and to develop targeted strategies for the early detection and effective mitigation of cyberbullying. The research will shed light on the specific emotional elements and nuances within social media platforms that serve as catalysts for cyberbullying behavior. By gaining a deep understanding of these emotional triggers, the research will empower both scholars and practitioners to explore viable constraints and interventions aimed at discouraging the exploitation of these triggers in the perpetration of cyberbullying. The study will contribute to the development of robust countermeasures and preventive measures that can be seamlessly integrated into social media platforms. Ultimately, the overarching goal is to cultivate a safer and more nurturing online environment where the presence of emotional triggers leading to cyberbullying can be promptly identified and effectively mitigated. The research aspires to create a digital sphere where users can engage in positive and empathetic interactions, free from the perils of cyberbullying.

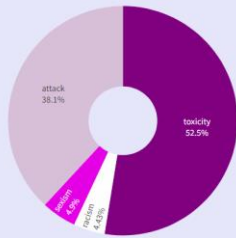




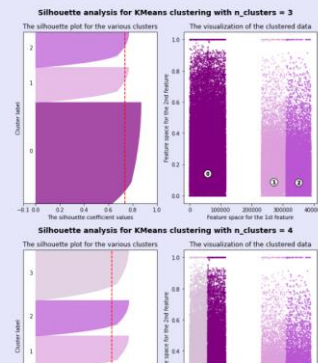
Emotions Analysis

Cyberbullying Analysis

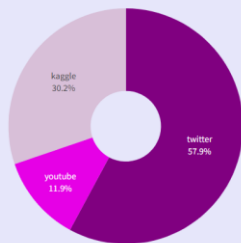
Toxic Behaviors



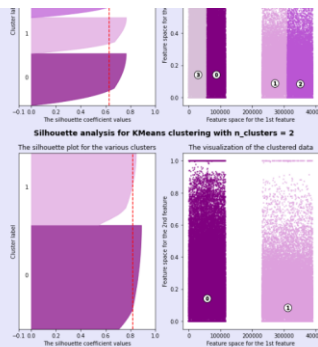
- toxicity
- attack
- sexism
- racism



Sources



- twitter
- kaggle
- youtube



Correlation Matrix Heatmap



	Non-Cyberbully	Cyberbully	cyberbullying
Non-Cyberbully	1	-1	-0.8695
Cyberbully	-1	1	0.8695
cyberbullying	-0.8695	0.8695	1

Enter a sentence:

wf, you can't show this, that is bad for your reputation

Predict

Cyberbully Label: Non-Cyberbully
Emotion Label: worry
Sentiment Label: Negative
Sentiment Polarity: -0.5999999999999999

APPENDIX IV

PROJECT REFERENCE

Here is the link to my Streamlit application and my GitHub repository, containing all the utilized data and implemented models:

<https://github.com/maritamatta/cyberbullying-emotion>

<https://cyberbullying-emotion-app.streamlit.app/>

REFERENCES

- Slonje, R., Smith, P. K., & Frisé, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior*, 29(1), 26-32.
- Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, 140(4), 1073–1137.
- Rigby, K. (2002). *New perspectives on bullying*. Jessica Kingsley Publishers.
- Faraj, S., & Azad, B. (2012). The materiality of technology: An affordance perspective. *Materiality and organizing: Social interaction in a technological world*, 237, 258.
- Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, 29(2), 129-156.
- Vishwamitra, N., Hu, H., Luo, F., & Cheng, L. (2021, January). Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*.
- Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7* (pp. 49-66). Springer International Publishing.
- Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., ... & Dillon, E. (2016, December). Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 740-745). IEEE.
- Huang, Q., Singh, V. K., & Atrey, P. K. (2014, November). Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-aware Multimedia* (pp. 3-6).
- Kontostathis, K., Reynolds, A., Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques", *Proceedings of the 5th annual acm web science conference*, pp. 195-204, 2013.
- P. Lempa, M. Ptaszynski and F. Masui, "Cyberbullying Blocker Application for Android", *presented at the 7th Language & Technology Conference (LTC'15)*, 2015.
- Jones, K. S. (2003). What is an affordance?. *Ecological psychology*, 15(2), 107-114.
- Olweus, D. (2012). Cyberbullying: An overrated phenomenon?. *European journal of developmental psychology*, 9(5), 520-538.

Falahatpisheh, Z., & Khajeheian, D. (2020). Affordances and IT design: A typology for social media and platform affordances. Paper presented at the 1-7.

J. Vitak, "Social media affordances: Enhancing or disrupting relationships", *Presentation at the 65th Annual International Communication Association (ICA) Conference*, 2015.

J. W. Treem and P. M. Leonardi, "Social media use in organizations: Exploring the affordances of visibility editability persistence and association", *Commun. Yearb.*, vol. 36, no. 1, 2012.

S. K. Evans, K. E. Pearce, J. Vitak and J. W. Treem, "Explicating affordances: A conceptual framework for understanding affordances in communication research", *J. Comput. Mediat. Commun.*, vol. 22, no. 1, pp. 35-52, 2017.

Patchin, J. W., & Hinduja, S. (2010). Cyberbullying and self-esteem. *Journal of school health*, 80(12), 614-621.

Dong, X., & Wang, T. (2018). Social tie formation in Chinese online social commerce: The role of IT affordances. *International journal of information management*, 42, 49-64.

Abhishek, G. S., Ingole, H., Laturia, P., Dorna, V., Maheshwari, A., Iyer, R., & Ramakrishnan, G. (2021). SPEAR : Semi-supervised data programming in python.

Lu, X. (2022). Deep learning based emotion recognition and visualization of figural representation. *Frontiers in psychology*, 12, 818833.

Yamada, T., Hashimoto, H., & Tosa, N. (1995, November). Pattern recognition of emotion with neural network. In *Proceedings of IECON'95-21st Annual Conference on IEEE Industrial Electronics (Vol. 1, pp. 183-187)*. IEEE.

Al-Hashedi, M., Soon, L. K., Goh, H. N., Lim, A. H. L., & Siew, E. G. (2023). Cyberbullying Detection Based on Emotion. *IEEE Access*.

O. Gencoglu, "Cyberbullying Detection With Fairness Constraints," in *IEEE Internet Computing*, vol. 25, no. 1, pp. 20-29, 1 Jan.-Feb. 2021.