AMERICAN UNIVERSITY OF BEIRUT

UNMASKING IMPLICIT ABUSE:
A DATA-CENTRIC APPROACH TO DETECT ONLINE ABUSIVE LANGUAGE

by

NOUHA ABDELLAH ABARDAZZOU

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
December 2023

AMERICAN UNIVERSITY OF BEIRUT

UNMASKING IMPLICIT ABUSE:
A DATA-CENTRIC APPROACH TO DETECT ONLINE ABUSIVE LANGUAGE

by
NOUHA ABDELLAH ABARDAZZOU

Approved by:

Signature

Dr. Wael Khreich, Assistant Professor — Advisor
Suliman S. Olayan School of Business

Signature

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

Signature Sirine Taleb

Dr. Sirine Taleb, Lecturer — Member of Committee
Suliman S. Olayan School of Business

Date of thesis defense: December 4, 2023

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Abardazzou Nouha Abdellah
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.

December 15, 2023

Signature Date

ACKNOWLEDGEMENTS

This master's thesis has been a journey filled with challenges and triumphs, and I owe immense gratitude to those who have made it all possible.

To my parents, sister, and brother, your unwavering encouragement has been a driving force.

To my invaluable support system, your presence has been indispensable throughout these past two years.

To Dr. Wael Khreich, my academic mentor, your guidance has sparked a new passion within me.

And to the Middle East Partnership Initiative (MEPI), I thank you for the opportunity.

This thesis is a testament to your collective contributions, and I hold you in the highest regard.

ABSTRACT
OF THE THESIS OF

Nouha Abdellah Abardazzou for Master of Science in Business Analytics
Major: Business Analytics

Title: Unmasking Implicit Abuse: A data Centric Approach to Detect Online Abusive Language

The rise of social media platforms has witnessed a disturbing increase in online abusive behavior, causing psychological harm, especially among children. This study focuses on detecting implicit abusive language, often overlooked in favor of explicit abuse. Implicit abuse conceals derogatory language within seemingly positive expressions, making it harder to identify. Using Twitter data, we collected and annotated a dataset distinguishing implicit, explicit, and non-abusive language. Our research leveraged traditional, deep learning, and transfer learning models to detect online abusive language. The Ensemble BERT model achieved a remarkable F1 score of 0.72 and AUC of 0.81 in detecting Implicit Abuse versus not abusive content. This research provides a deeper understanding of the nuances of online abuse and offers a significant step toward creating a safer online environment that promotes healthy digital interactions and the well- being of users.

TABLE OF CONTENTS

ILLUSTRATIONS

Figure

TABLES

Table

CHAPTER 1

INTRODUCTION

With the rise of social media platforms, there has been a disturbing surge in abusive behavior online, resulting in detrimental psychological effects (Duggan, 2017; Munro et al., 2011). For instance, the research by Munro et al. (2011) highlights the negative consequences of online abuse on children, linking such experiences to the development of mental health issues such as depression and anxiety. The Pew Research Center's report further emphasizes the prevalence of abusive behavior, with 40% of US adults reporting encountering online abuse and 18% experiencing severe forms of harassment, including sexual harassment. Moreover, according to the same report by Duggan (2017), 13% of American internet users have discontinued their use of online services due to witnessing abusive and disruptive behavior from fellow users. These statistics underscore the urgent need for automated abuse detection and moderation systems to foster a safer digital environment and mitigate the detrimental effects of online abuse.

Abusive language is typically identified by its exceptionally impolite and derogatory nature. Previous definitions of abusive language, encompass the following aspects: "any strongly impolite, rude, or hurtful language using profanity, that can demonstrate the degradation of someone or something, or express intense emotion" (Fortuna et al., 2018; Nobata et al., 2016; Park & Fung., 2017). Additionally, Caselli et al. (2020) define abusive language as "hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions”.

While explicit abusive language, such as threats, sexual harassment, and aggression, has received considerable research attention, the detection and understanding of implicit abusive language have been relatively overlooked (Chatzakou et al., 2017a; Chen et al., 2012a; Waseem et al., 2017). Unlike explicit abusive language which is easily recognizable due to its explicit obscenities and slurs, implicit abusive language can be equally harmful but disguised through sarcasm, humor, stereotypes, ambiguous words, or the absence of explicit profanity (Frenda, 2018.; MacAvaney et al., 2019; Nobata et al., 2016; Sanguinetti et al., 2020).

Implicit abusive language, frequently hidden within ostensibly positive expressions or conveyed through subtler forms of aggression, presents a significant challenge for detection (Waseem et al., 2017). This challenge is amplified by the deliberate use of metaphors, homophones, abbreviations, and other linguistic forms, further complicating the process of detection (Caselli et al., 2020). This complexity renders traditional approaches, such as keyword-based methods commonly used to identify explicit abusive language, obsolete when it comes to effectively detecting implicit abuse, as the latter lacks explicit markers (Gao et al., 2017). Furthermore, the dynamic nature of the internet, where language evolves rapidly, compounds the limitations of traditional keyword-based methods. New abusive terms and expressions continually emerge, making these conventional techniques less effective (Raisi & Huang, 2016).

To address implicit abusive language, a deeper understanding of context and language nuances becomes essential (Ptaszynski et al., 2016). For instance, the interpretation of sarcastic comments often relies on contextual knowledge and the user’s background understanding (Nobata et al., 2016). The prevalence of metaphors in abusive samples compared to non-abusive ones underscores their significance in abuse detection (Mishra, 2018). Moreover, despite the increasing recognition of implicit abusive language as a significant issue, there remains a notable absence of representative benchmark datasets for evaluation. Consequently, a clear practical approach to detect such language is yet to be established (Elsherief et al., 2021).

Our research aims to contribute to the field of abusive language detection by focusing on the often-overlooked aspect of implicit abusive language (Chen et al., 2012b). Using a data-driven approach, we aim to identify important signals that indicate implicit abuse, improving moderation and prevention efforts. Our primary goal is to create a carefully annotated dataset covering implicit and explicit instances of abusive language, capturing the complexities of online abusive behavior. We will also develop specialized machine learning (ML) and natural language processing (NLP) models designed for detecting online abusive language. These models will be trained on our dataset to analyze subtle language cues that signify abusive behavior, automating the detection process and empowering content moderation.

Our research explores pivotal questions concerning abusive language detection. Firstly, we assess the extent to which ML and NLP approaches can effectively detect implicit abusive language in online communication. This investigation entails employing diverse traditional and modern ML models on our meticulously curated dataset. Additionally, we scrutinize which distinct patterns and characteristics differentiate implicit abusive language from explicit and non-abusive forms, aiming to optimize detection methods by leveraging these differentiating factors. Lastly, we rigorously evaluate what is the best accuracy level achieved by our detection models in identifying implicit abusive language. By addressing these questions, our research aims not only to provide substantial insights and advancements in the field of implicit abusive language detection but also seeks to limit the negative impact on individuals who experience online abuse. Our work is to offer a more comprehensive understanding of the intricacies involved in online abuse detection, ultimately contributing to a healthier online environment, and fostering digital spaces that are safer and more supportive for all users.

In summary, our research endeavors to propel the field of abusive language detection forward by focusing on implicit forms of abusive language and employing a data-centric approach. Our primary objectives are threefold: first, to create an annotated dataset that distinguishes between implicit, explicit, and non-abusive language, facilitating a more comprehensive understanding of online abuse; second, to develop specialized detection models capable of identifying less-explored types of online abuse, particularly implicit abuse, with the aim of ensuring a safer online experience; and third, to address pivotal research questions surrounding the complexities of implicit abusive language detection, ultimately contributing to a healthier online environment.

CHAPTER 2

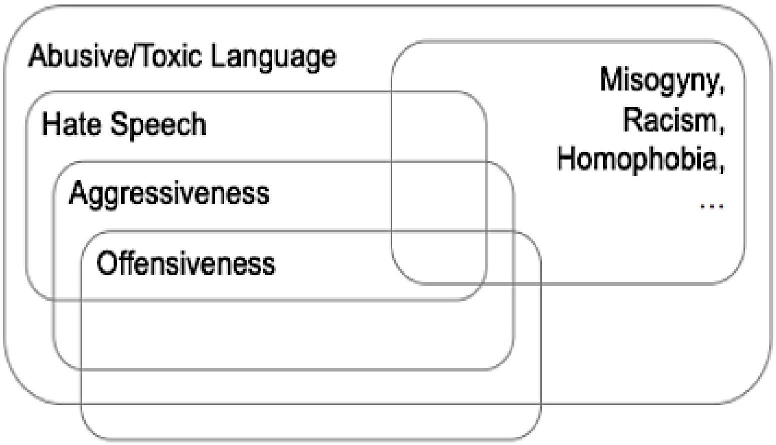
LITERATURE REVIEW

The literature review in this thesis provides a comprehensive overview of research on detecting online abuse, with a specific focus on implicit abusive language. It explores terminology, approaches, and challenges associated with detecting implicit abuse. This chapter lays the foundation for developing more effective methods in this area by synthesizing existing literature.

2.1. Abusive Language Terminology

Abusive language serves as an overarching concept that encompasses various sub-categories, such as Hate Speech, Harassment, Cyberbullying, Aggression Identification, and Offensive Language Identification (Davison et al., 2009; Kumar et al., 2018; Tokunaga, 2010; Waseem & Hovy, 2016; Zampieri et al., 2019). The definition of abusive language provided by Fortuna and Nunes (2018) aligns with this understanding, as it includes hate speech, derogatory language, and profanity, collectively referred to as "hurtful language". This definition highlights the intention of abusive language to demean and insult individuals or groups (Nobata et al., 2016; Vidgen et al., 2021; Waseem et al., 2017). Poletto et al. (2021) proposed a comprehensive framework illustrated in Figure 1 that organizes the various terms and concepts associated with abusive language to establish a structured framework for comprehending abusive language. This framework includes hate speech, aggressiveness, and offensiveness, providing a systematic approach to studying the different dimensions of abusive language.

Figure 1 Taxonomy of abusive language adopted from Poletto et al. (2021)



Kshirsagar et al. (2018) defines hate Speech as language that expresses hatred toward a targeted group with the intention to derogate, humiliate, or insult its members. This category focuses on instances where individuals or groups are targeted based on their characteristics, and the language used aims to demean and degrade them. Harassment is often characterized as the repeated act of sending nasty, mean, and insulting messages with the explicit intention to annoy others. It involves a persistent pattern of intentionally bothersome communication that seeks to disturb and distress the recipient (Davison et al., 2009). On the other hand, Cyberbullying encompasses the deliberate and repeated infliction of harm on an individual through electronic text mediums. Unlike Hate Speech and Harassment, Cyberbullying specifically focuses on the repeated aspect and imbalance of power, where individuals are targeted and subjected to ongoing harmful behavior (Dadvar et al., 2012). Lastly, the concept of cyber aggression refers to aggressive online behavior with the intent to harm. It often manifests as communication that attacks individuals based on certain characteristics, frequently employing rude language. Cyber aggression encompasses various forms of hostile interactions online, where the primary objective is to cause harm to the targeted individuals through verbal assaults and personal attacks (Chatzakou et al., 2017b; Poletto et al., 2006).

2.2. Abusive Language Detection

According to Schmidt and Wiegand (2017), the first classification model for abusive language can be traced back to Spertus (1997). In their work, Spertus employed a decision tree algorithm combined with manually created rules to classify messages as abusive or non-abusive. Over time, abusive language classifiers improved, especially with the emergence of neural networks, making manual feature engineering techniques outdated. This section provides a literature review on feature selection and machine learning models used for detecting abusive language online.

2.2.1. Features Selection

2.2.1.1. Dictionaries & exicons

In their work Sood et al. (2012) explored the efficacy of a lexicon-based strategy for detecting offensive language. They utilized a shared profanity list obtained from the website phorum.org as a basis for their analysis. The researchers developed a system that would label a comment as offensive if it contained any of the words listed on phorum.org. However, their findings revealed several challenges with this approach. They identified three main reasons for the technique’s poor performance. First, misspellings and variations in offensive language posed difficulties in accurately capturing and matching profane terms. Second, the inability of the lexicon-based approach to adapt to evolving offensive language limited its effectiveness in keeping up with emerging expressions. Finally, the context-specific nature of profanity added complexity, as the offensiveness of certain words could depend on the context in which they were used. This study highlights the need for more sophisticated strategies to overcome the limitations of lexicon-based approaches and achieve higher performance in offensive language detection.

2.2.1.2. Textual Features

In their respective studies, Chen et al. (2012a) and Nobata et al. (2016) have shown the superiority of N-Grams over Bag of Words (BOW) characteristics. For detailed definitions of these concepts, please refer to Section 4.3. of this thesis. Additionally, previous research has explored different content-based aspects that can be utilized as features. These aspects include comment length (Dadvar et al., 2013; Davidson et al., 2017), the ratio of capital letters (Dadvar et al., 2014; Huang et al., 2014), the presence of special characters (Chatzakou et al., 2017b), and the number of emoticons (Dadvar et al., 2013). These text features enhance machine learning models' performance in tasks such as abusive language detection, providing valuable insights for a more accurate and nuanced analysis of offensive language across different contexts.

2.2.1.3. Semantic Features

The significance of semantic features in conveying word meanings and establishing lexical relationships within a language has been highlighted by researchers (Al-Garadi et al., 2016; Cheng et al., 2015). Studies detecting abusive content have utilized word embeddings, such as FastText, Word2Vec, and GloVe, which capture semantic associations, syntactic correlations, and grammar-based information. While word embeddings have proven valuable, they have limitations in capturing the multiple meanings of a word across different contexts (Djuric et al., 2015; Zhao et al., 2016).

Semantic features are crucial for understanding word meanings and relationships. Word embeddings like FastText, Word2Vec, and GloVe capture semantic associations. However, character-level models have limited use in detecting abusive content, requiring further research for evaluation in this domain.

2.2.1.4. Syntactic Features

Syntactic characteristics, such as part-of-speech (POS) tagging and dependency relations, have been recognized as essential elements in text analysis (Al-Garadi et al., 2016; Chen et al., 2012a; Dadvar et al., 2013; Nobata et al., 2016; Xu & Zhu, 2010). For instance, Xu and Zhu (2010) highlighted that the extensive use of adjectives in a comment can reveal a specific viewpoint. Furthermore, several studies have emphasized the significance of first- and second-person pronouns in online postings, which provide insights into the intended audience and the potential to provoke or irritate others when combined with offensive terms (Al-Garadi et al., 2016; Chen et al., 2012a; Dadvar et al., 2013; Nobata et al., 2016). By incorporating syntactic features such as POS tagging and dependency relations, researchers enhance text analysis and gain valuable insights into the structure and intention of abusive content.

2.2.1.5. Sentiment Features

The research community has shown significant interest in leveraging sentiment features for the detection of abusive language, as it can be linked to social psychological phenomena like aggressive and antisocial behavior change (Ali et al., 2022; Chatzakou et al., 2017a; Yin et al., 2009). Chatzakou et al. (2017a) employed the SentiStrength tool to evaluate the sentiment of text, detecting positive and negative sentiment. Similarly, Yin et al. (2009) incorporated the presence of pronouns and foul language as sentiment features in their classification feature set. Recent studies have further explored the application of sentiment analysis features to address specific issues. For example, Ali et al. investigated the correlation between news coverage by mainstream news channels and the occurrence of hate speech and Islamophobic sentiment (Ali et al., 2022). These studies highlight the importance of sentiment features in comprehending abusive language and its impact within diverse social contexts.

2.2.2. Machine Learning Models

2.2.2.1. Traditional Machine Learning Models

Logistic Regression (LR), Naïve Bayes (NB), Decision Trees (DT), Random Forests (RF), and Linear Support Vector Machine (SVM) classifiers are commonly employed in various applications, including abusive language detection (Davidson et al., 2017; Nobata et al., 2016; Razavi et al., 2010; Salminen et al., 2019; Xu et al., 2012). Among these approaches, SVM has consistently demonstrated superior performance, making it the preferred choice for many researchers.

In their paper, Razavi et al. (2010) made a significant contribution to the field of hostile language detection by introducing an innovative approach. They pioneered the use of lexicon-based features and semantic rules integrated into a NB classifier, achieving an accuracy of 0.68. Notably, their work marked an important milestone as the first lexicon-based method for abuse detection, where they constructed a comprehensive dictionary consisting of insulting and abusive words and phrases, each assigned a weight indicating its abusive impact. This pioneering approach laid the foundation for subsequent advancements in the field of hostile language detection.

Davidson et al. (2017) aimed to detect hate speech by employing a range of features and models. They utilized text preprocessing techniques, including lowercasing, and stemming, followed by the creation of bigram, unigram, and trigram features weighted by Term Frequency-Inverse Document Frequency (TF-IDF), and syntactic information was captured through the POS tag. Additionally, they incorporated quality indicators, sentiment scores, and binary/count indicators for various tweet characteristics. For model selection, LR with L1 regularization was initially used for dimensionality reduction. LR with L2 regularization emerged as the best-performing model, with linear SVMs also exhibiting favorable results. The final model achieved an overall precision of 0.91 and an F1 score of 0.90. This study showcases the application of diverse features and models for hate speech detection while acknowledging the challenges in accurately identifying hate speech.

Nobata et al. (2016) conducted an extensive study on detecting abusive language by utilizing various textual features. They identified 13 feature types, including the presence of polite expressions and modal words, and combined them to detect offensive language. Their approach involved N-Grams features, linguistic features, syntactic features, and distributional semantic features obtained through Word2vec. Experimental results on multiple datasets demonstrated that combining multiple features significantly improved the performance of offensive language detection compared to using a single feature. The study provided valuable insights into the importance of feature fusion and the effectiveness of different feature types in detecting abusive language.

Salminen et al. (2019) explored the automatic detection of online hate speech using various features, such as N-Grams (TF and TF-IDF), semantic and syntactic features, and Word2vec/doc2vec embeddings. Multiple machine learning algorithms were evaluated, including LR, DT, RF, Adaboost, and SVM. The study highlighted the effectiveness of N-Grams features (TF and TF-IDF) in detecting hate speech, with consistent performance from LR and DT. SVM achieved the highest F1 score among the classifiers, and combining SVM with TF-IDF weighted N-Grams achieved an impressive 0.96 F1 score for binary classification. This study showcased the effectiveness of various features and machine learning algorithms in detecting online hate speech.

Table 1 summarizes the features used and the reported performance of the traditional machine learning models presented above. It is important to note that the performances of all models studied in this chapter were evaluated on distinct datasets and are not directly comparable.

Table 1 Evaluation Metrics & Features of Traditional ML Models

Paper	Accuracy	Precision	Recall	F1-Score	Features	Best Model	Algorithms
Razavi et al. (2010)	0.68	-	-	-	Dictionary	NB	NB
Davidson et al. (2017)	-	0.91	0.9	0.9	N-Grams, TF-IDF, POS, Binary/count indicators for tweet characteristic	LR (L2)	LR, SVM
Nobata et al. (2016)	-	-	-	-	Linguistic features: number of polite words, modal words	-	-
Salminen et al. (2018)	-	-	-	0.96	N-Grams, TF-IDF, Semantic features, Word2vec/ doc2vec embeddings.	SVM with TF-IDF weighted N-Grams	LR, DT, RF, Adaboost, SVM

2.2.2.2. Deep Learning Models

In their study, Badjatiya et al. (2017) focused on hate speech detection and experimented with various classifiers, including LR, RF, SVM, Gradient Boosted Decision Trees (GBDTs), and Deep Neural Networks (DNNs). They utilized task-specific embeddings generated from FastText, Convolutional Neural Networks (CNN), and Long Short-Term Memory Networks (LSTMs) to define the feature spaces for these classifiers, comparing them with baselines such as char N-Grams, TF-IDF vectors, and Bag of Words vectors (BoWV). The research findings demonstrated the significant outperformance of deep learning models in hate speech detection, with a remarkable F1 score of 0.839, surpassing the baseline method that employed SVM with the highest F1 score of 0.816. Notably, the combination of LSTM and GBDT yielded the highest F1 score of 0.93, highlighting the effectiveness of leveraging multiple architectures in hate speech detection. Overall, Badjatiya et al.'s work reveals the superiority of deep learning models over traditional approaches for detecting hate speech.

Zampieri et al. (2019) conducted a study focusing on classifying messages as offensive or not offensive, utilizing SVM, bidirectional Long Short-Term Memory (BiLSTM), and CNN models. In their experiments, the BiLSTM model demonstrated superior precision in detecting offensive messages, achieving a precision score of 0.81. For the detection of non-offensive language, the precision score was 0.83. Comparatively, the SVM and CNN models achieved precision scores of 0.66 and 0.78, respectively, for offensive message detection, and precision scores of 0.80 and 0.87 for the detection of non-offensive messages. These findings highlight the effectiveness of the BiLSTM model in accurately identifying offensive content, while the SVM and CNN models also exhibited reasonable performance in the classification task.

An overview of the features and performance of the Deep Learning models presented is provided in Table 2.

Table 2 Evaluation Metrics & Features of Deep Learning Models

Paper	Accuracy	Precision	Recall	F1-Score	Features	Best Model	Algorithms
Badjatiya et al. (2017)	-	0.93	0.93	0.93	N-Grams, TF-IDF, BoWV, Random embeddings, GloVe	LSTM & GBDT	LR, SVM, CNN, LSTM, GBDT
Zampieri et al. (2019)	-	0.81	0.48	0.60	Unigram, FastText	BiLSTM	SVM, BiLSTM, CNN

2.2.2.3. Transfer Learning Models

The introduction of Bidirectional Encoder Representations from Transformers (BERT) has significantly impacted abusive language detection in the field of natural language processing (Caselli et al., 2021; Devlin et al., 2019; Safaya et al., 2020). BERT and other transformer-based classification models have emerged as state-of-the-art approaches in abusive language classification, outperforming alternative methods in recently shared tasks (Mandl et al., 2020; Risch et al., 2021).

Safaya et al. (2020) aimed to enhance detection results by combining BERT with CNN in their study. By processing Twitter tweets, converting hashtags into plain text format, and utilizing a twelve-layer BERT model with CNN, they achieved improved classification performance. Evaluation on three language datasets (Arabic, Greek, Turkish) showed that the proposed BERT-CNN approach achieved the highest F1 score of 0.851, outperforming other models such as SVM with TF-IDF (0.760), Multilingual BERT (0.796), Bi-LSTM (0.801), CNN-Text (0.805), and BERT (0.841). The improvement in F1 score of 1% over BERT alone demonstrated the effectiveness of the BERT-CNN fusion in enhancing the detection performance.

Caselli et al. (2021) proposed the HateBERT model, a retrained BERT model for detecting abusive language phenomena in English. The study employed the same pre-processing steps and hyperparameters for HateBERT and the generic BERT model. They conducted a comparative analysis using the generic BERT model on three English general-purpose datasets: OffensEval (Zampieri et al., 2019), AbusEval (Caselli et al., 2020), and HatEval (Basile et al., 2019). HateBERT consistently outperformed the generic BERT model in terms of macro F1 scores. In the OffensEval dataset, HateBERT achieved a macro F1 score of 0.809, surpassing BERT's score of 0.803. Similarly, in the AbusEval dataset, HateBERT achieved a macro F1 score of 0.765, outperforming BERT's score of 0.727. For the HatEval dataset, HateBERT obtained a macro F1 score of 0.516, while BERT scored 0.480. These findings highlight the effectiveness of HateBERT in improving the detection of abusive language phenomena.

An overview of the features and performance of the Transfer Learning models presented is provided in Table 3.

Table 3 Evaluation Metrics & Features of Transfer Learning Models

Paper	Accuracy	Precision	Recall	F1-Score	Features	Best Model	Algorithms
Safaya et al. (2020)	-	-	-	0.85	TF-IDF	BERT-CNN	SVM, Multilingual BERT, Bi-LSTM, CNN, BERT
Caselli et al. (2021)	-	-	-	0.80	-	HateBERT	HateBERT, BERT

2.3. Implicit Abusive Language

2.3.1. Implicit Abusive Language Definition

Implicit abusive language is often defined by the lack of immediate implication or denotation of abuse (Caselli et al., 2020; Waseem et al., 2017; Wiegand et al., 2021). Waseem et al. (2017) brought attention to the utilization of positive language that carries negative connotations, creating an impression of benignity while harboring a mean undertone. Previous studies, including Caselli et al. (2020), Waseem et al. (2017) and Wiegand et al. (2021) have defined explicitness and implicitness in the detection of abusive language based on the presence or absence of specific keywords, slurs, or profanity. However, Vidgen et al. (2020) have highlighted that definitions primarily centered around slurs and profanity do not fully align with the broader linguistic or social interpretations of implicit abusive language due to the inherently subjective nature of implicitness. Further examination of implicit abusive language reveals its reliance on ambiguous terms and figurative language, such as metaphor or sarcasm (Mishra et al., 2020; Mohammad et al., 2016). Mishra et al. (2020) elaborate that implicit abuse often employs various forms of figurative language, including sarcasm, irony, metaphor, rhetorical questions, analogies, and comparisons. Metaphors and sarcasm emerge as prominent features, enabling the communication of intensified emotions and sentiments compared to literal expressions (Mohammad et al., 2016). Additionally, Mishra (2018) observed a higher prevalence of metaphors in abusive samples, underscoring their significance in capturing implicit abusive language.

In this study, we adhere to the following definition of implicit abusive language, drawing from various existing definitions found in the literature. Implicit abusive language is a form of verbal expression that exhibits a subtler manifestation of abusive tendencies. It is characterized by the presence of ambiguous terms and figures of speech, such as metaphor or sarcasm. Although explicit abusive words may not be explicitly present, the overall understanding of the sentence, text, or discourse implies a hurtful and abusive intention. This nuanced form of abusive language operates through indirect means, requiring a deeper analysis of contextual cues and underlying implications.

Table 4, presented below, provides examples of implicit tweets collected from our dataset.

Table 4 Example of Collected Implicit Tweets

Implicit Tweets
@USER If you have read the actual complaint, you will find it is a complete joke. A bit like you.
@USER I can't even begin to understand how he cannot seem to function without her holding his leash (I mean hand)
@USER keep watching your soaps...it's obviously where you get your logic?...from?
@USER She liked the curtains that much she is wearing them ,what a taste.
@USER Time for you to change your glasses obviously you cannot read or look up facts
@USER The hair looks like the aftermath of a Cali. wildfire.

2.3.2. Implicit Abusive Language Challenges

The identification of explicit abusive language is generally considered a straightforward task, as it involves recognizing explicit obscene expressions and slurs using keyword-based approaches (Wiegand et al., 2018). However, the detection of implicit abuse presents significant challenges due to its nuanced nature, relying on subtle linguistic cues and contextual factors that elude the direct offensiveness typically associated with abusive language (Frenda, 2018; MacAvaney et al., 2019; Nobata et al., 2016). A primary challenge in detecting implicit abusive language lies in the need for a deeper understanding of context and language nuances (Sap et al., 2020). Unlike explicit abuse, which can be identified through specific keywords, implicit abuse requires deciphering hidden meanings that may not be immediately apparent (Breitfeller et al., 2019; Mendelsohn et al., 2020; Ptaszynski et al., 2016). For instance, the interpretation of sarcastic comments often relies on contextual knowledge and the user's background understanding (Nobata et al., 2016). According to Mishra et al. (2018) the contextual aspect highlights the inherent relationship between abuse and the wider conversation within which it occurs, suggesting that individual comments may be challenging to classify without considering their respective contexts. Another significant challenge pertains to the situational and topical nature of abusive language. Contextual features that can indicate implicit abuse vary depending on the specific context or topic being discussed (Chandrasekharan et al., 2018). Therefore, modeling online conversations and their evolution toward abusive language could provide valuable insights into detecting implicit abuse more accurately (Mishra et al., 2020). Furthermore, the continuous evolution of internet language further complicates the detection of implicit abusive language. The emergence of new abusive terms, expressions, and linguistic variations poses challenges for traditional keyword-based methods, as they may struggle to keep pace with these evolving linguistic dynamics (Raisi & Huang, 2016). According to Nobata et al. (2016), classifiers trained on more recent data tend to outperform those trained on older data due to the evolving nature of internet jargon. Overall, effectively detecting implicit abusive language necessitates the utilization of sophisticated techniques that go beyond surface-level analysis. It requires a comprehensive understanding of language nuances, the evolving landscape of internet language, and the contextual factors influencing the interpretation of abusive expressions. Addressing these challenges is crucial for the development of accurate and effective detection methods to combat the pervasive issue of implicit abusive language online.

2.3.3. Available Implicit Abusive Language Datasets

To address the challenge of detecting and understanding implicit abuse, it is crucial to develop annotated datasets that specifically target implicit abusive language. Existing datasets in this field have primarily focused on explicit abusive text, often employing keyword-based approaches, which overlook implicit forms of abuse (Caselli et al., 2020; Wiegand et al, 2018). Consequently, these datasets lack a representative number of instances containing implicit abuse and suffer from a high proportion of false negatives (Basile et al., 2019; Mozafari et al., 2019; Wiegand et al., 2019). Moreover, identifying and annotating implicit abuse poses additional challenges, as it requires understanding the context and the specific cultural background knowledge and experience to decipher the hidden meaning behind such statements (Breitfeller et al., 2019; Mendelsohn et al., 2020; Ptaszynski et al., 2016).

2.3.3.1. The AbuseEval Dataset

Caselli et al. (2020) introduced the AbuseEval dataset by re-annotating the publicly available Offensive Language Identification Dataset (OLID). This new dataset addresses critical issues in the annotation of offensive/abusive language, such as message explicitness, target presence, and contextual requirements.

- **Dataset**

To construct the AbuseEval dataset, Caselli et al. (2020) leveraged the Offensive Language Identification Dataset (OLID) dataset originally introduced by Zampieri et al. (2019). OLID was developed for the SemEval 2019 shared task on offensive language detection and consists of a collection of English tweets. The OLID dataset employs a hierarchical annotation scheme that includes three levels: subtask A determines whether a message is offensive or not, subtask B identifies whether the offensive message has a target or not, and subtask C categorizes the target of the offensive message as an individual, a group, or other entities. This hierarchical annotation scheme enables a more nuanced analysis of offensive language in online communication.

- **Data Annotation**

Caselli et al. (2020) proposed annotation guidelines for the AbuseEval dataset, utilizing a decision tree approach to distinguish between explicit and implicit abuse. The guidelines introduced a differentiation between utterances, characterized by coherent language, and non-utterances, encompassing unconnected words or hashtags. Non-utterances were intentionally excluded from the annotation process, although their potential for conveying abusive language was acknowledged. Annotators were instructed to label instances as explicit abuse when they contained negative context profanities targeting individuals or groups, including explicit markers like profanity and negative connotations, as well as abusive hashtags. Implicit abuse was identified based on linguistic constructions involving sarcasm, irony, and rhetorical questions, aligning with the definition of abuse that implies or infers without explicit markers. Non-abusive messages were classified as either context-dependent or genuinely non-abusive. The guidelines aimed to minimize subjective interpretation in the annotation of different forms of abusive language within the dataset.

- **Implicit vs. Explicit Classification**

Caselli et al. (2020) utilized a BERT model to classify abusive messages in the AbuseEval dataset. The evaluation of the model's performance revealed relatively lower precision and recall scores for the implicit abuse class compared to the other categories. Specifically, the precision for implicit abuse was approximately 0.234, and the recall was around 0.098. In contrast, the non-abusive class achieved a precision of 0.864 and a recall of 0.936, while the explicit abuse class had a precision of 0.640 and a recall of 0.509. The study revealed challenges in predicting the implicit abuse class due to the limited amount of available training data. Despite this limitation, the BERT model showed relatively stable performance in predicting explicit instances, even though they were less represented compared to non-abusive categories. This highlights the unique nature and complexity of implicit abuse detection, underscoring the need for further research and development in this area.

- **Limitation of the AbuseEval Dataset**

The AbuseEval dataset introduces an innovative annotation scheme for classifying abusive language, but it is not without limitations. One key limitation is its reliance on a predefined lexicon of abusive words inherited from the OLID/OffensEval dataset. This approach, where any message containing profanity is automatically labeled as offensive, may exclude other forms of abusive language not covered by the lexicon. It oversimplifies the distinction between offensive and explicitly abusive language, potentially leading to misrepresentations. Additionally, while the annotation process of the AbuseEval dataset recognizes the importance of contextual information, the conservative approach of labeling messages as non-abusive when context cannot be retrieved may result in mislabeling implicit forms of abuse. To address these limitations, the authors propose directions for future research. Firstly, they recommend grounding data sets for abusive language detection in contextual information, emphasizing the need for annotations that consider the context of occurrence rather than treating messages in isolation. Secondly, they suggest collecting data directly from "hateful" users who are more likely to employ abusive/offensive language, instead of relying solely on keyword-based retrieval methods. This approach aims to reduce bias and enhance the identification of implicit and complex expressions of abuse/offense.

2.3.3.2. The GAB Hate Corpus (GHC)

The GAB Hate Corpus dataset, developed by Kennedy et al. (2018) is a large-scale annotated corpus from gab.com, providing theoretically justified labels for hate-based rhetoric, target group, and rhetorical framing to facilitate hate speech classification.

- **Dataset**

Kennedy et al. (2018) collected a dataset consisting of 27,665 posts from the social network service gab.com. To ensure sufficient textual content, posts were included in the sampling process only if they contained at least three non-hyperlink tokens.

- **Data Annotation**

The annotation process involved training annotators to categorize posts using a comprehensive coding typology developed by the authors, which drew from multiple research fields such as law, computational science, psychology, and sociology. The typology encompassed various subtasks. Firstly, posts were coded as "assaults on human dignity" (HD) or "calls for violence" (CV) based on explicit targeting of social groups, following the definition of "hate-based rhetoric." Another subtask involved identifying and evaluating hate-based slurs, specifically focusing on their harmful usage rather than casual usage. Instances targeting groups or their characteristics were labeled as "Vulgarity and/or Offensive language" (VO). Additionally, texts were analyzed for framing effects, distinguishing between implicit (IM) and explicit (EX) rhetoric. The distinction between implicit and explicit rhetoric was guided by the framework introduced by Waseem et al. (2017), considering the presence of implicit or explicit framing effects. Implicit rhetoric relied on cultural knowledge to evoke derogatory beliefs, sentiments, or threats, while explicit rhetoric directly conveyed discriminatory attitudes. This annotation process emphasizes the importance of differentiating between implicit and explicit rhetoric in the examination of hate-based discourse, aligning with the concept of denotation and connotation within language as highlighted by Waseem et al. (2017).

- **Implicit vs. Explicit Classification**

Kennedy et al. (2018) conducted experiments on the Gab Hate Corpus (GHC) to develop algorithms for hate-based rhetoric detection. They used LIWC and TF-IDF features, SVM with linear kernels, and fine-tuning BERT models. BERT achieved the highest performance, with an F1 score of 0.44, accuracy of 0.92, precision of 0.46, and recall of 0.42 for classifying hate content. LIWC features provided less information but improved performance when combined with TF-IDF. The study underlined the challenge of hate speech prediction and the need for iterative modeling and background knowledge integration.

- **Limitation**

The GHC dataset faces a limitation regarding annotator agreement for implicit and explicit rhetoric labels, casting doubts on their reliability. To address this challenge and enhance the understanding of implicit hate speech, Kennedy et al. (2018) recommended the incorporation of additional extra-linguistic data is recommended. Considering diverse contexts and user accounts can significantly influence the interpretation of the text in relation to expressed prejudices. Collaborations between NLP researchers and social scientists are essential in developing comprehensive datasets and modeling strategies that effectively capture speaker intentions.

2.3.3.3. The Aggression Identification Dataset

The Aggression Identification Dataset developed by Kumar et al. (2018) is a diverse collection of comments from Facebook Pages and Twitter, encompassing various contentious topics in India and featuring data in English, Hindi, and other Indian languages, with labeled categories of Overtly Aggressive, Covertly Aggressive, and Non-Aggressive.

- **Dataset**

Kumar et al. (2018) collected a dataset consisting of approximately 18,000 tweets and 21,000 Facebook comments. The data was obtained by crawling public Facebook Pages and Twitter, with a focus on pages and issues expected to generate significant discussion among Indians, particularly in Hindi. More than 40 Facebook pages were recognized and crawled to collect the data, while for Twitter, popular hashtags related to contentious themes such as the beef ban, India vs. Pakistan cricket match, election results, and opinions on movies were used for data collection. Importantly, the data collection process did not involve sampling based on language, resulting in the inclusion of data from English, Hindi, and other Indian languages.

- **Data Annotation**

Kumar et al. (2018) conducted a document-level annotation process, encompassing posts, comments, and discourse units. Annotators received comprehensive guidelines that defined the typology of verbal aggression, distinguishing between overt and covert aggression. Overt aggression involved explicit expressions of aggression using specific lexical items, aggressive lexical features, or syntactic structures. Covert aggression, on the other hand, comprised indirect attacks disguised as insincere polite expressions, employing conventionalized polite structures. Verbal aggression was further classified into target-based categories, including physical threat, sexual threat/aggression, identity threat/aggression (with various subtypes), and non-threatening aggression. Additionally, comments containing abusive language were labeled as "abuse," while tweets or comments in languages other than English or Hindi were categorized as non-aggressive.

- **Implicit vs. Explicit Classification**

The TRAC-2018 shared task employed the Aggression Identification Dataset to develop a three-way classification system for Overtly Aggressive, Covertly Aggressive, and Non-aggressive text data. Aroyehun et al. (2018) emerged as the winners by achieving a remarkable F1 score of 0.642%. Their winning approach involved a combination of recurrent neural networks (RNN) and CNN, alongside data augmentation and pseudo-labeling techniques. Notably, their method showcased strong generalization capabilities, outperforming other approaches even when trained on the Facebook dataset and evaluated on the Twitter dataset. This achievement highlights the significance of their work in the context of text classification.

- **Limitations**

During the task, while the Aggression Identification dataset was generally regarded as high quality by most participants, two significant limitations were identified. Firstly, the dataset contained comments in English that exhibited code-mixed Hindi-English data, as well as instances of other languages like German. Although these occurrences constituted a small portion of the data, it was necessary to filter them out to ensure consistency. The second and more substantial limitation pertained to the annotation process itself. Some participants raised concerns about potential inaccuracies in the annotations. Aggression is a highly subjective phenomenon; different annotators may have varying judgments regarding the same comment. Consequently, certain annotations appeared implausible and require further scrutiny and validation to enhance reliability. These limitations highlight the importance of addressing language variations and ensuring the accuracy and consistency of annotations in aggression identification datasets.

2.3.3.4. Latent Hatred Dataset

The Latent Hatred Dataset, developed by ElSherief et al. (2021), provides fine-grained labels for implicit hate speech, sourced from online hate group accounts on Twitter, and specifically targets active hate groups in the United States.

- **Dataset**

ElSherief et al. (2021) conducted a comprehensive data collection to build the Latent Hatred Dataset, consisting of 12,143 tweets. The data was sourced from online hate groups and their followers on Twitter, utilizing the authors' own taxonomy for categorizing implicit hate speech. This taxonomy, informed by social science and relevant NLP literature, includes categories such as White Grievance, Incitement to Violence, Inferiority Language, Irony, Stereotypes and Misinformation, and Threatening and Intimidation. The data collection period spanned from January 1, 2015, to December 31, 2017, capturing a substantial range of hate group activities prior to the removal or suspension of numerous accounts.

- **Data Annotation**

ElSherief et al. (2021) used a two-stage annotation process to label implicit hate speech. In the first stage, Amazon Mechanical Turk (MTurk) annotators were employed to assign high-level labels to the tweets. Annotators were given hate speech definition and examples and reached a majority agreement for most tweets. The resulting consensus labels included 933 explicit hate, 4,909 implicit hate, and 13,291 not hateful tweets. In the second stage, the authors applied more fine-grained category definitions to label the 4,909 implicit hate tweets, enhancing the understanding of implicit hate speech.

- **Implicit vs. Explicit Classification**

In the task of distinguishing implicit hate speech from non-hate, ElSherief et al. (2021) conducted experiments using SVM and BERT baselines for text classification. The SVM models achieved competitive performance with F1 scores up to 0.644 in binary implicit hate speech classification. However, fine-tuned BERT models showed improvement, gaining up to 6 additional points in the F1 score. The BERT-base model exhibited significantly better macro precision (0.721 vs. a maximum of 0.614) compared to linear SVMs, indicating a deeper understanding of the text beyond simple keyword-matching.

- **Limitations**

The dataset and work by ElSherief et al. (2021) introduce a theoretical taxonomy of implicit hate speech and provide a large-scale benchmark corpus with fine-grained labels. While this initial effort enables a better understanding and modeling of implicit hate speech, there are several limitations. Specifically, the dataset is tailored to active hate groups in the United States, focusing on intentionally veiled acts of intimidation, threats, and abuse. Additionally, the authors identify eight challenges in detecting implicit hate speech, including coded hate symbols, discourse relations, entity framing, commonsense, metaphorical language, colloquial speech, irony, and identity term bias. To address these challenges, ElSherief et al. (2021) suggested that future research could explore the development of models for deciphering coded language, lifelong learning of hateful language, contextualized sarcasm detection, and bias mitigation techniques for named entities in hate speech detection systems.

Table 5 provides a summary of the four implicit abusive language datasets discussed in this section. The table includes key information such as dataset sizes, labels, identified limitations, and recommendations.

Table 5 Compilation of Datasets on Abusive Language with Implicit Labels

Datasets	***Size***	***Labels***	***Limitations***	***Recommendations***
AbuseEval Dataset	*13,240*	*- Implicit abuse* *- Explicit abuse* *- Not abuse*	- Reliance on a predefined lexicon of abusive words. - Conservative annotation.	- Include context information. - Collect data directly from "hateful" users.
GAB Hate Corpus	*12,000*	*- Overtly Aggressive* *- Covertly Aggressive* *- NonAggressive*	- Limited annotator agreement for implicit and explicit rhetoric labels.	-Include extra-linguistic data for better understanding of implicit hate speech. -Capture diverse contexts and user accounts to enhance the interpretation of expressed prejudices.
Aggression Identificati on Dataset	*27,665*	*- Implicit* *- Explicit*	- Presence of code-mixed and multilingual data. - Potential inaccuracies in the annotation process due to subjective judgments.	-Filter out code-mixed and non-English instances to ensure dataset consistency. -Validate and scrutinize annotations to enhance reliability and accuracy.
Latent Hatred Dataset	*22,584*	*- Explicit hate* *- Implicit hate* *- Not hate*	- US focused dataset, limiting generalizability.	-

In conclusion, the literature review has highlighted several limitations in existing datasets used for abusive language detection research. These limitations include the predominant emphasis on explicit forms of abusive text, neglecting implicit manifestations of abuse, and relying heavily on keyword-based approaches (Caselli et al., 2020; Wiegand et al., 2018). Consequently, the datasets lack sufficient instances representing implicit abuse, leading to a significant proportion of false negatives in classification (Basile et al., 2019; Mozafari et al., 2019; Wiegand et al., 2019). Moreover, the heterogeneity in dataset size, labeling schema, class balance, and overall quality poses challenges in comparing and generalizing results across studies (Vidgen et al., 2019; Vidgen & Derczynski, 2021).

The identification and annotation of implicit abuse present additional complexities, requiring contextual comprehension and cultural background knowledge to decipher the concealed meanings of statements (Breitfeller et al., 2019; Mendelsohn et al., 2020; Sap et al., 2020). Consequently, the quality of annotation varies, impacting the inter-rater reliability of the datasets (Vidgen et al., 2019; Van Aken et al., 2018). Additionally, dataset degradation occurs when researchers rely on references to comments or posts instead of accessing the actual content, leading to reduced dataset sizes over time due to the removal of references based on platform policies (Vidgen et al., 2019). Unintended biases inherent in the datasets, such as topic bias, can also result in undesirable model behavior (ElSherief et al., 2021; Wiegand et al., 2019). Furthermore, regular updates of the datasets are necessary to incorporate evolving language patterns, particularly in internet jargon, to ensure the ongoing effectiveness of classifiers (Nobata et al., 2016).

Acknowledging the limitations in current abusive language detection datasets, we have opted to collect our own dataset. Existing datasets predominantly emphasize explicit abuse, neglect implicit manifestations, and rely on keyword-based methods, leading to significant drawbacks. To address these issues, our dataset collection prioritizes a balanced representation of explicit and implicit abuse, integrates context-aware comprehension, and proposes a comprehensive approach to constructing a diverse and representative dataset that addresses the identified challenges. In the next chapter, we will present the details of our data collection methodology, providing insights into the approach employed to ensure the reliability of our dataset.

CHAPTER 3
DATA COLLECTION AND ANNOTATION

In this chapter, we present a detailed justification for the choices made in the data collection and annotation processes. Drawing upon insights from relevant literature, our goal is to address the limitations identified in Chapter 2, particularly the absence of a labeled dataset tailored to the specific context of implicit online abuse. Our aim is to create a representative dataset that can effectively capture explicit, implicit, and none-abusive language. This section encompasses two main components: Data Collection and Data Annotation.

3.1. Dataset Collection

The data collection process is systematically designed to ensure the representativeness and reliability of the dataset. It comprises several key steps, namely the selection of data sources, collection of online conversations and data filtering. Each step plays a crucial role in enhancing the quality and integrity of the collected data. The following section provides a detailed overview of these steps.

3.1.1. Dataset Collection Process

3.1.1.1. Selection of Data Sources

To ensure the diversity and representativeness of our dataset, we adopt a targeted approach to gather data from users who exhibit "hateful" behavior, as recommended by Caselli et al. (2020). This approach overcomes limitations associated with conventional keyword-based retrieval methods, which may fail to capture the entirety of abusive language instances.

To identify these hateful users, we initiate our data collection process by investigating instances of online backlash that have recently led to cyberbullying or hate-fueled incidents. A noteworthy case under examination involves Meghan Markle, who has been subjected to persistent hate-fueled harassment online. By extensively researching and analyzing press articles[1], we identify relevant hashtags associated with the backlash (e.g., #Meghan_Markle_Revealed). Subsequently, we performed a targeted search on Twitter to compile a list of users who dedicate entire accounts to propagating online hate against Meghan Markle.

In selecting data sources, it is imperative to choose those that offer a comprehensive and diverse representation of online conversations and comments. Twitter, being a widely used social media platform with extensive user interactions, serves as our primary data source. The real-time nature of Twitter enables us to capture the dynamic aspect of abusive language in online contexts. It is important to acknowledge that Twitter users constitute a biased sample of the population, with approximately 80% of tweets originating from just 10% of users (Wojcik et al., 2019). Despite this inherent bias, Twitter data has been widely employed in research and has demonstrated its potential for various applications in social good, as supported by our Literature Review.

3.1.1.2. Collection of Conversations

Our data collection methodology is informed by the insights provided by Mishra et al. (2020) and Caselli et al. (2020), highlighting the significance of modeling online conversations and anchoring dataset collection in contextual information. To enhance our comprehension of abusive language and its implicit manifestations, we give priority to

[1] https://sports.yahoo.com/twitter-data-shows-meghan-markle-131412609.html

collecting and analyzing conversations rather than isolated messages. This approach enables us to consider the broader context in which abusive language arises, facilitating the identification of underlying meanings behind abusive statements.

To implement this methodology, we use Twitter's API V2[2], a powerful tool provided by Twitter. We rely on the "conversation_id"[3] query parameter which links replies to the original Tweet that initiated the conversation, as shown in table 6. It groups all responses, even sub-threads, under the same conversation_id. This approach enables us to extract and analyze entire conversation threads, providing a comprehensive understanding of the context and nuances surrounding abusive language in these conversations.

Table 6 Example of Scrapped Conversations from Twitter

Author ID	**In_reply_to_user_id**	**Text**
77911548475	1599509527	@USER @USER Diana cheat first and many more times Charles was aching for his love Diana was a slut
1599509527	77911548475	@USER @USER Were u even alive when diana was? If yes, the delusional statement stands, if no you're youthful ignorance is charming
77911548475	1599509527	@USER @USER Diana was the cheating fool FIRST MULTIPLE TIMES
1599509527	77911548475	@USER @USER And don't call her a fool. Or any dead person a fool. U don't know her to be that familiar. No one on Twitter does.

3.1.1.3. Data Filtering

To enhance the relevance of our collected comments, we implement specific criteria for data filtering. Initially, we consider comments that consist of fewer than 50 words, emphasizing concise and focused content suitable for analysis. This approach

[2] https://developer.twitter.com/en/docs/twitter-api/early-access

[3] https://developer.twitter.com/en/docs/twitter-api/conversation-id

prioritizes comments that are more likely to convey meaningful information pertaining to abusive language, while also easing the burden on annotators. Furthermore, we exclude comments that solely comprise of emojis, URLs, or memes from our dataset. Although these elements hold significance in online communication, they often lack substantial textual content directly related to abusive language. By excluding such comments, we ensure that our dataset primarily focuses on textual expressions that bear greater relevance to our study.

Through this meticulous filtering process, we curate a dataset that incorporates comments with adequate textual content for analysis while excluding those that may not significantly contribute to our understanding of abusive language. By applying these criteria, we strive to enhance the quality and relevance of our collected data, thereby facilitating more accurate and insightful analysis in our research.

3.1.1.4. Active Collection of Dataset

Our research follows a data-centric approach. Initially, we manually labeled 100 tweets across various abusive language categories, laying the foundation for training our SVM + TF-IDF model. The effectiveness of this specific model is supported by prior research, including Salminen et al. (2018), which we discussed in our literature review.

Subsequently, the model was applied to the remaining unlabeled dataset, generating scores for each sentence, as shown in Table 7. These scores helped prioritize instances with a high likelihood of containing implicit abusive language. A single experienced human annotator verified these scores, marking an essential step in our dataset collection process. It's worth noting that this verification process is just one phase of our dataset collection, with plans for further annotation rounds to enhance dataset quality.

In summary, our approach combines the SVM + TF-IDF model with human annotators for iterative dataset improvement, ensuring a balanced, high-quality dataset for our research thesis.

Table 7 Annotation based on our suggested approach.

Index	Text	Proba_label	Proba_Score
2653	@USER Yes, they push it too far, people should take those papers to court. They are so low in their mind.	Implicit	0.9431106448173523
4501	@USER More like a filthy douchebag	Explicit	0.3630037307739258
1715	@USER Sounds like you’re complaining and can’t say anything nice	None	0.4919069707393646

3.1.2. Dataset Size

Through meticulous application of rigorous filtering criteria and thoughtful considerations, our initial dataset, originally comprising roughly 7,000 tweets, has undergone refinement to a final selection of 1,938 tweets. This reduction is a crucial step, allowing us to create a focused and manageable dataset that aligns seamlessly with our research objectives.

Additionally, the downsizing of our dataset greatly improves the efficiency of our annotation process, making it more manageable for annotators and enhancing the overall coherence and effectiveness of our study. This systematic curation process enables precise data scrutiny, streamlining our research procedure while maintaining data integrity. Thus, finding the right balance between dataset size and quality ensures a reliable analysis of our research findings.

3.2. Data Annotation Process

Annotation Guideline: Our annotation process hinges on a robust guideline, depicted in Figure 2, and detailed in appendix. This guideline, inspired by Kumar et al. (2018) and enriched with insights from established guidelines of El Sherief et al. (2017) and Caselli et al. (2020), lays the foundation for identifying and classifying abusive language within our dataset. It furnishes annotators with clear instructions, definitions of explicit, implicit, and non-abusive content, and illustrative examples for precise labeling.

Annotation Task Distribution: The distribution of annotation tasks comprises Task 1 and Task 2. Task 1 entails annotators determining the presence of abusive content, while Task 2 is concerned with distinguishing between explicit and implicit forms within the previously identified abusive content.

Annotators: Our annotation team consists of three female graduate students, averaging 24 years in age, with four years of academic experience. They are well-prepared to make informed decisions during annotation, guided by detailed guidelines and examples. While the annotators' gender composition may introduce some gender-related bias, our current priority is ensuring high-quality and consistent annotations. In future work, we will explore strategies to mitigate potential gender-related biases for a more inclusive dataset.

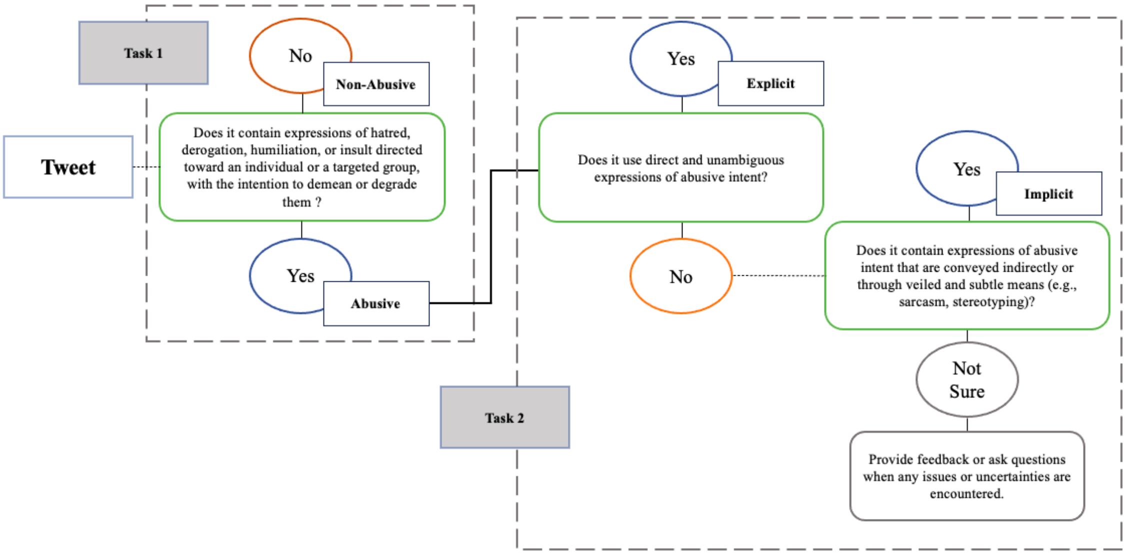
Validation and Review: Continuous validation and review are integral to maintaining annotation accuracy. Annotators actively provide feedback and seek clarifications, fostering a collaborative environment. This approach ensures adherence to guidelines and addresses uncertainties promptly.

Inter Annotator Agreement (IAA): To assess the consistency among multiple annotators, we computed Fleiss' Kappa, a statistical measure introduced by Fleiss (1971).

Fleiss' Kappa measures the extent of agreement observed among multiple annotators, accounting for what can be attributed to chance. Our computed Fleiss' Kappa value of 0.49 indicates a moderate level of agreement among annotators, surpassing what would be expected by random chance. Detecting implicit abuse poses challenges, contributing to the moderate agreement observed. Additionally, a limitation arises from all annotators being non-native English speakers, potentially leading to variations in interpreting sarcasm or nuanced language expressions. This linguistic diversity may impact overall agreement levels, highlighting the complexities of annotating abusive content, especially when conveyed implicitly.

Despite the challenges we faced, our agreement rate emphasizes the reliability and consistency of our annotations. This aligns with the acknowledged difficulty in capturing subtle linguistic nuances, implying a commendable level of agreement given the inherent complexities of the task.

Figure 2 Annotation Guideline Process


Task 1
No
Non-Abusive
Tweet
Does it contain expressions of hatred, derogation, humiliation, or insult directed toward an individual or a targeted group, with the intention to demean or degrade them ?
Yes
Abusive
Yes
Explicit
Does it use direct and unambiguous expressions of abusive intent?
No
Yes
Implicit
Does it contain expressions of abusive intent that are conveyed indirectly or through veiled and subtle means (e.g., sarcasm, stereotyping)?
Not Sure
Task 2
Provide feedback or ask questions when any issues or uncertainties are encountered.

In this chapter, we have detailed our data collection and annotation processes, emphasizing the creation of a reliable dataset for abusive language detection. The next chapter will focus on the development and evaluation of our proposed abusive language detection models.

CHAPTER 4

EXPERIMENTAL METHODOLOGY

In this chapter, we detail our experimental methodology. We cover data preparation, pre-processing, feature generation, and the development of machine learning models, including traditional, deep learning, and transfer learning models. We also discuss model training, parameter tuning, and multi-class evaluation, emphasizing strategies and metrics for assessing performance. This chapter forms the foundation of our research, ensuring rigor in our investigations.

4.1. Data Preparation

The dataset used in this study consists of 1,938 tweets, categorized into three distinct classes: Implicit Abusive Language (ImpAbu), Explicit Abusive Language (ExpAbu), and Non-Abusive Language (NonAbu). The classes exhibit a balanced distribution with 663 Implicit Abuse, 665 Explicit Abuse, and 610 Non-Abusive tweets. This balance ensures equitable representation, minimizing biases and providing a robust foundation for model training. The dataset was stratified into 60% for training, 20% for validation, and 20% for testing, carefully preserving class proportions in each subset. The training set was used to train our models, while the validation set was utilized for hyperparameter optimization. The test set remained untouched during training, enabling unbiased evaluation of the models' performance on unseen instances.

Table 8 Dataset Distribution

Classes	Label	Size
Implicit Abusive Language	ImpAbu	663
Explicit Abusive Language	ExpAbu	665
Non-Abusive Language	NonAbu	610

4.2. Data Pre-processing

In our study, we rigorously explore the impact of distinct text preprocessing steps on refining the effectiveness of our online abusive language detection system. A set of comprehensive experiments was meticulously devised to investigate the individual and collective contributions of various preprocessing techniques. The preprocessing encompasses a total of 10 steps encompass:

1. **Removal of HTML tags**: Regular expressions are utilized to eliminate HTML tags commonly found in online content, preserving the abusive language present in the text.
2. **Removal of accented characters**: Accented characters are replaced with their respective ASCII codes to standardize the text and handle spelling variations.
3. **Removal of user mentions**: User mentions, indicated by '@' symbols followed by usernames, are removed to address privacy concerns.
4. **Removal of repeated characters**: Consecutive repeated characters occurring more than twice are reduced to a single occurrence, handling exaggerated expressions commonly seen in online abusive language.
5. **Removal of special characters (except for !, *,.):** Irrelevant special characters are removed to simplify the analysis, while certain characters like '!' and '*' are retained due to their significance in bypassing banned words (e.g., F**k).
6. **Replacement of multiple whitespaces**: Multiple consecutive whitespaces are replaced with a single whitespace to enhance text readability and consistency.

7. **Conversion to lowercase**: The entire text is converted to lowercase to avoid discrepancies due to capitalization.
8. **Removal of emojis**: Emojis are eliminated to prevent interference with the textual analysis of abusive language, this aligns with our primary data collection objectives that prioritize the textual content of tweets.
9. **Removal of stop words**: Common stop words that do not carry significant meaning in the context of abusive language are removed.
10. **Lemmatization**: Words are lemmatized, reducing them to their base or root form to capture the fundamental meaning and handle variations.

The preprocessing experiments are structured as detailed in Table 9. Each experiment is conducted in conjunction with a baseline assessment, allowing us to measure the incremental impact of each preprocessing step on the model's performance.

Table 9 Preprocessing Experiments and Steps

Experiment	**Preprocessing Steps**
Preprocessing 1	Steps 1 to 7 (Removal of HTML tags, accented characters, user mentions, repeated characters, special characters, whitespaces, conversion to lowercase)
Preprocessing 2	Steps 1 to 7 + Step 8 (Emoji Removal) + Step 10 (Stop Word Removal)
Preprocessing 3	Steps 1 to 7 + Step 8 (Emoji Removal) + Step 10 (Stop Word Removal) + Step 11 (Lemmatization)

By systematically dissecting the preprocessing stages and evaluating their effects, we aim to identify optimal strategies that enhance the efficacy of our online abusive language detection system. The results of these experiments will be presented and discussed in the subsequent chapter, providing insight into the intricate dynamics of text preprocessing and its implications for online abusive language detection.

4.3. Features Selection

Several experiments are conducted at the features level, testing N-Grams, TF-IDF features, and word embeddings GloVe and BERT.

- **N-Grams**: N-Grams refer to ordered sequences of N adjacent words, capturing the context of each word to some extent (Bengfort et al., 2018). We experimented with different ranges of N-Grams and obtained the best results using a combination of words Unigrams and Bigrams.
- **TF-IDF**: Term Frequency-Inverse Document Frequency (TF-IDF) is utilized to weigh the importance of words in documents (Webb & Sammut, 2010). We tested TF-IDF unigrams, bigrams, and a combination of unigrams and bigrams techniques.
- **GloVe**: Global Vectors for Word Representation (GloVe) is an unsupervised method that creates word-to-word co-occurrence vector representations based on a global corpus. The resulting vector representations exhibit intriguing linear substructures within the word vector space (Pennington et al., 2014). We conducted experiments using GloVe embeddings for both traditional machine learning models and deep learning models.
- **BERT**: Bidirectional Encoder Representation from Transformers (BERT) is introduced by Devlin et al. in 2018 and captures words’ multiple meanings based on context. In this study, BERT is employed to extract machine-readable data representations from text, and it is used in combination with traditional machine-learning models.

4.4. Machine Learning Models

This section presents a diverse range of machine learning models utilized to detect implicit, explicit, and non-abusive language online. These models fall into three main categories: traditional machine learning, deep learning, and transfer learning.

4.4.1. Traditional Machine Learning Models

We explored various traditional machine-learning models for online abusive language detection. The following models were investigated: LR, RF, Extreme Gradient Boosting (XGBoost), and SVM. The performance of these models was extensively compared and evaluated.

4.4.1.1. Logistic Regression (LR)

LR is a linear supervised machine learning technique used for classification tasks. It employs the logistic function to predict a dependent variable based on the interplay of a set of independent variables (Hosmer & Lemeshow, 2013; Stoltzfus, 2011).

4.4.1.2. Random Forest (RF)

RF involves the creation of a set of decision trees in a randomized manner. These individual trees exhibit subtle variations in terms of their tendency to overfit and their predictive capabilities. The combination of outcomes from these multiple trees, through averaging, serves to mitigate overfitting while preserving the predictive potency of the model (Breiman, 2001).

4.4.1.3. Extreme Gradient Boosting (XGBoost)

XGBoost stands out as a scalable and adept interpretation of the Gradient Boosting Machines Framework (Friedman, 2001). This framework, akin to a collection of decision trees combined harmoniously, yields a more powerful model (Chen et al., 2015; Chen & Guestrin, 2016). Notably deviating from RFs, XGBoost employs a sequential arrangement of decision trees, facilitating each subsequent model to amend the errors of its forerunner (Guido & Müller, 2016).

4.4.1.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification tasks. Its primary objective is to identify the optimal hyper-plane that effectively separates classes, maximizing the margin between data points and the hyper-plane (Cristianini & Ricci, 2008). For text classification in this study, we utilized LinearSVC, a variant of SVM, particularly well-suited for classifying implicit, explicit, and non-abusive online language. Detailed descriptions of our SVM experiment combined with various feature representations are provided in the subsequent sections.

4.4.2. Deep Learning Models

Given the sequential nature of text data, neural network architectures that consider this characteristic have shown significant success in various natural language processing (NLP) tasks. In this section, we introduce and analyze four deep learning models—LSTM, Bi-LSTM, CNN, and the hybrid CNN+Bi-LSTM—for the specific application of online abusive language detection.

4.4.2.1. Long Short-Term Memory (LSTM)

LSTM is a deep learning model specifically designed to capture sequential information (Hochreiter & Schmidhuber, 1997). Leveraging its ability to capture long-term dependencies within input sequences, LSTM models effectively comprehend the context and relationships among words. As each word is processed sequentially through LSTM units, the model accumulates a comprehensive understanding of the entire text, retaining crucial information over extended sequences.

4.4.2.2. Bidirectional Long Short-Term Memory (BiLSTM)

To address the limitation of LSTM's unidirectional nature, BiLSTM was introduced, which has demonstrated superior performance in various natural language processing (NLP) applications (Wang et al., 2015; Huang et al., 2015). BiLSTM consists of two LSTM layers—one processes the data from left to right, and the other processes data from right to left. The outputs from both LSTM layers are concatenated and flattened, enhancing the model's context understanding by considering information from both ends of the sequence.

4.4.2.3. Convolutional Neural Network (CNN)

CNN is a feed-forward artificial neural network primarily employed for hierarchical document classification (Yamashita et al., 2018). Originally popular in computer vision tasks, CNNs have showcased promising performance in natural language processing (NLP) domains as well. In NLP, CNNs act as feature extractors, adept at identifying informative N-Grams within text sequences, considering their local ordering but not global position. This capability allows CNNs to capture crucial textual aspects relevant to the specific prediction task at hand.

4.4.2.4. CNN+BiLSTM Hybrid Model

The CNN+BiLSTM Hybrid Model represents an ingenious fusion of two potent neural network architectures, namely CNN and BiLSTM. This hybrid approach capitalizes on the respective strengths of both networks, thereby enhancing its effectiveness in natural language processing tasks. By combining the feature extraction capabilities of CNN and the contextual understanding of BiLSTM, the model gains a comprehensive understanding of textual content, making it well-suited for the detection and classification of implicit, explicit, and non-abusive language in online contexts.

4.4.3. Transfer Learning Models

The emergence of pre-trained models, fine-tuned to specific tasks, provides researchers with a unified architecture applicable to different contexts. The transfer learning models explored in this study include BERT, RoBERTa, HateBERT, and an ensemble of these models.

4.4.3.1. Bidirectional Encoder Representation from Transformers (BERT)

BERT is a state-of-the-art language model introduced by Devlin et al. (2018). It utilizes a transformer-based architecture, specifically designed to handle long-range dependencies in sequential data, such as natural language text. Pre-trained on a large corpus of text data from Wikipedia, BERT employs an unsupervised learning approach to predict masked words in a sentence or to predict the next sentence in a pair of sentences during pre-training. This process enables BERT to learn comprehensive representations of word meaning and relationships within sentences.

4.4.3.2. Robustly optimized BERT approach (RoBERTa)

RoBERTa is an enhanced version of BERT released by Liu et al. (2019). In contrast to BERT, RoBERTa is trained on a larger dataset of 160 GB of uncompressed text and employs a higher batch size during training. Additionally, RoBERTa adopts a dynamic token masking strategy, randomly masking some tokens in each pass to better utilize contextual information for predicting masked tokens. These optimizations result in significantly improved performance across various language tasks, making RoBERTa a powerful and robust language representation model.

4.4.3.3. HateBERT

HateBERT is a specialized version of the BERT model designed for detecting specific forms of online toxicity, including offensive language, hate speech, and abusive language (Caselli et al., 2020). Unlike modifying the BERT architecture, HateBERT achieves its improved performance by retraining BERT on data that is more relevant to social media platforms, where most online discussions take place. Notably, HateBERT outperforms the standard BERT model in all three classification tasks and achieves state-of-the-art results on the AbusEval dataset.

4.4.3.4. Ensemble BERT

Ensemble BERT model employed in our final experiment is a powerful combination of BERT, RoBERTa, and HateBERT. It is designed to harness the strengths of each individual model, thereby enhancing the overall performance and predictive capabilities. BERT, RoBERTa, and HateBERT are state-of-the-art language models, each excelling in different aspects of natural language processing tasks. By integrating these models into an ensemble, we aim to capitalize on their complementary features and achieve superior results in detecting implicit, explicit, and non-abusive language online. The ensemble approach allows us to effectively leverage the pre-training knowledge and fine-tuned expertise of each individual model, resulting in a comprehensive and robust solution for abusive language detection in online contexts.

4.5. Models Training and Parameter Tuning

To ensure robust model performance, we employed a stratified data-splitting approach for all models. We utilized Grid Search Cross-Validation with 5-fold validation to identify the best hyperparameters for each classifier. The primary evaluation metric was the AUC scores computed using the 'One vs Rest' approach. Recording results for each hyperparameter combination facilitated a comprehensive analysis of model performance across various settings.

4.5.1. Traditional Machine Learning Models

We conducted a comprehensive assessment of the traditional machine learning models used in this study. Each model underwent an extensive hyperparameter search to optimize its performance on our relatively small dataset for multi-class classification. Hyperparameter tuning focused on key parameters specific to each model. For LR, {'C'} regularizes the strength that influences the trade-off between fitting the training data and preventing overfitting. In the case of RF and XGBoost, hyperparameters {'max_depth'} determine the maximum depth of decision trees and {'n_estimators'} specify the number of trees in the ensemble, which were fine-tuned to control model complexity. Additionally, for Support Vector Machine, we adjusted hyperparameters {'C'}, which represents the regulation strength, and 'gamma,' the kernel coefficient, to enhance model performance. The results, documented in the accompanying table 10, showcase the macro-averaged AUC scores achieved during the validation phase. This rigorous approach enhances the credibility of the conclusions drawn in this study, considering the data's size and complexity.

Table 10 Hyperparameter Tuning for Traditional Machine Learning Models

Models	**Param_Grid**	**Best Parameters**	**AUC Score Validation**
LR+Ngrams	: [0.01, 0.1, 1, 10, 100,1000]}	{'C': 1}	0.759
LR+TFIDF	{'C': [0.01, 0.1, 1, 10, 100,1000]}	{'C': 10}	0.774
LR+Glove	{'C': [0.01, 0.1, 1, 10, 100,1000]}	{'C': 0.1}	0.778
LR+BERT	{'C': [0.01, 0.1, 1, 10, 100,1000]}	{'C': 0.1}	0.795
RF+Ngrams	{'n_estimators': [50, 100, 150], 'max_depth': [5, 10]}	{'max_depth': 5, 'n_estimators': 150}	0.752
RF+TFIDF	{'n_estimators': [50, 100, 150], 'max_depth': [5, 10]}	{'max_depth': 5, 'n_estimators': 150}	0.747
RF+Glove	{'n_estimators': [50, 100, 150], 'max_depth': [5, 10]}	{'max_depth': 10, 'n_estimators': 100}	0.741
RF+BERT	{'n_estimators': [50, 100, 150], 'max_depth': [5, 10]}	{'max_depth': 10, 'n_estimators': 100}	0.758
XGBoost+Ngrams	'n_estimators': [100, 150, 200], 'max_depth': [3, 5, 7]	{'max_depth': 3, 'n_estimators': 200}	0.733
XGBoost+TFIDF	'n_estimators': [100, 150, 200], 'max_depth': [3, 5, 7]	{'max_depth': 3, 'n_estimators': 200}	0.687
XGBoost+Glove	'n_estimators': [100, 150, 200], 'max_depth': [3, 5, 7]	{'max_depth': 7, 'n_estimators': 100}	0.759
XGBoost+BERT	'n_estimators': [100, 150, 200], 'max_depth': [3, 5, 7]	{'max_depth': 5, 'n_estimators': 200}	0.788
SVM+Ngrams	'C': [0.01, 0.1, 1, 10, 100],'gamma': ['scale', 0.1, 0.05], kernel='linear'	{'C': 1, 'gamma': 'scale', 'kernel'='linear'}	0.764
SVM+TFIDF	'C': [0.01, 0.1, 1, 10, 100],'gamma': ['scale', 0.1, 0.05], kernel='linear'	{'C': 1, 'gamma': 'scale', 'kernel'='linear'}	0.774
SVM+Glove	'C': [0.01, 0.1, 1, 10, 100],'gamma': ['scale', 0.1, 0.05], 'kernel'='linear'	{'C': 0.1, 'gamma': 'scale', 'kernel'='linear'}	0.776
SVM+BERT	'C': [0.01, 0.1, 1, 10, 100],'gamma': ['scale', 0.1, 0.05], 'kernel'='linear'	{'C': 0.01, 'gamma': 'scale', 'kernel'='linear'}	0.802

4.5.2. Deep Learning Models

Next, we investigated our four deep learning models to discern implicit, explicit, and non-abusive online language. We varied the {'batch_size'} to balance effective learning with the dataset's small size, ensuring efficient training. The {' epochs '} values were selected to strike a suitable balance between preventing underfitting and overfitting to the data. The {'dropout_rate'} was fine-tuned to manage overfitting, preventing the loss of essential information while maintaining model robustness. Additionally, we integrated the adaptive learning rate technique through the Adam optimizer to dynamically adjust learning rates during training. Taking inspiration from the findings of Priyadarshini and Cotton (2021), we explored their recommended combinations, as illustrated in Table 11. Employing a grid search based on validation AUC scores, we identified the best hyperparameters, leading to substantial enhancements in model performance and increased efficacy in abusive language detection tasks.

Table 11 Hyperparameter Tuning for Deep Learning Models

Models	**Param_Grid**	**Best Parameters**	**AUC Score Validation**
LSTM	{'batch_size': [10, 20, 40, 60], 'epochs': [8, 10, 20], 'dropout_rate': [0.1, 0.2, 0.3]}	{'batch_size': 64, 'epochs': 10, 'dropout_rate': 0.1}	0.799
Bi-LSTM	{'batch_size': [10, 20, 40, 60], 'epochs': [8, 10, 20], 'dropout_rate': [0.1, 0.2, 0.3]}	{'batch_size': 32, 'epochs': 10, 'dropout_rate': 0.2}	0.802
CNN	{'batch_size': [10, 20, 40, 60], 'epochs': [8, 10, 20], 'dropout_rate': [0.1, 0.2, 0.3]}	{'batch_size': 32, 'epochs': 20, dropout_rate': 0.2}	0.804
CNN+Bi-LSTM	{'batch_size': [10, 20, 40, 60], 'epochs': [8, 10, 20], 'dropout_rate': [0.1, 0.2, 0.3]}	{'batch_size': 64, 'epochs': 20, 'dropout_rate': 0.1}	0.773

4.5.3. Transfer Learning Models

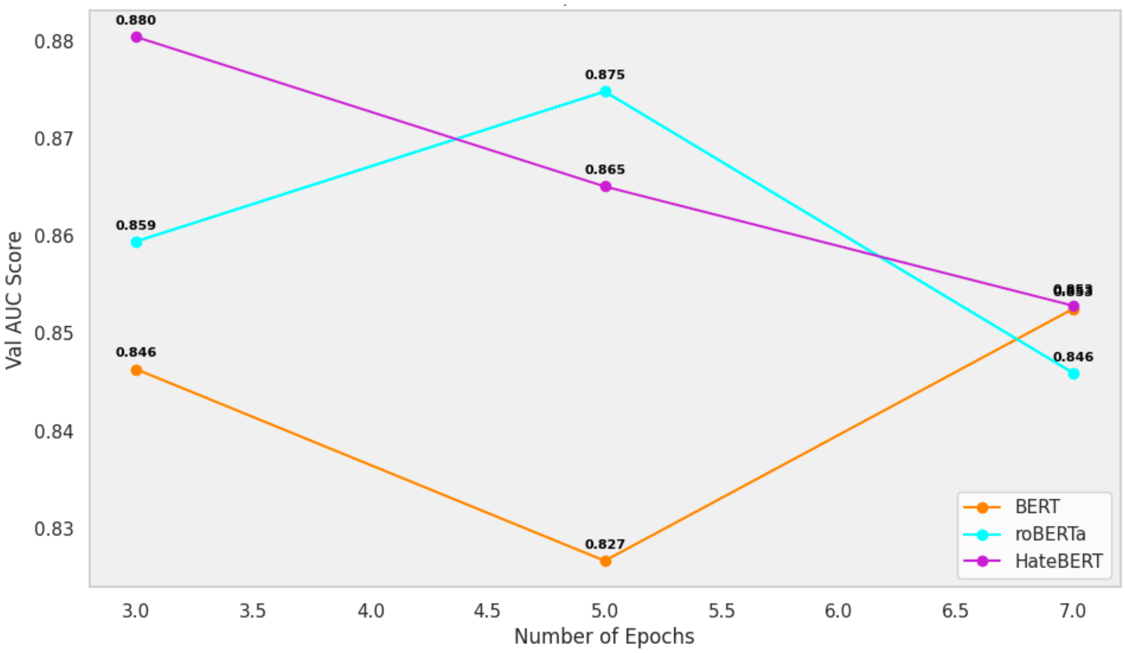
We fine-tuned three transfer learning models, BERT, RoBERTa, and HateBERT, to detect abusive language in online content. The Adam optimizer was selected for its effectiveness in fine-tuning BERT-based models, and we utilized categorical cross-entropy for multi-class classification (Sun et al. 2019). Our hyperparameter exploration encompassed varying the {'batch_size'}, which determines the number of training examples used in each iteration, with a particular emphasis on smaller batches tailored to our small dataset. We conducted a comprehensive grid search over the {' epochs '}, which represent the number of complete passes through the training dataset, and {'batch_size'}. Additionally, we fine-tuned the BERT model's hyperparameters, specifically the {'learning_rate'} and {'dropout_rate'}, to mitigate overfitting. The details of these hyperparameter configurations and the corresponding validation AUC scores are provided in Table 12. This meticulous approach enhances the reliability and robustness of our results by ensuring that our models are finely tuned for the task at hand.

Table 12 Hyperparameter Tuning for Transfer Learning Models

Models	**Param_Grid**	**Best Parameters**	**AUC Score Validation**
BERT	{'batch_size': [16, 32, 64], 'epochs': [3, 5, 7], 'learning_rate': [1e-5, 2e-5, 5e-5], 'dropout_rate': [0.2, 0.8]}	{'batch_size': 16, 'epochs': 7, 'learning_rate': 2e-5, 'dropout_rate': 0.2}	0.852
RoBERTa	{'batch_size': [16, 32, 64], 'epochs': [3, 5, 7], 'learning_rate': [1e-5, 2e-5, 5e-5], 'dropout_rate': [0.2, 0.8]}	{'batch_size': 16, 'epochs': 5, 'learning_rate': 1e-5, 'dropout_rate': 0.2}	0.875
HateBERT	{'batch_size': [16, 32, 64], 'epochs': [3, 5, 7], 'learning_rate': [1e-5, 2e-5, 5e-5], 'dropout_rate': [0.2, 0.8]}	{'batch_size': 16, 'epochs': 3, 'learning_rate': 2e-5, 'dropout_rate': 0.2}	0.880

Figure 3 displays validation AUC scores across the epochs listed in Table 12, highlighting the models' training performance and the influence of tuned epochs on AUC.

Figure 3 Impact of Epochs on Transfer Learning Models' Performance



We fine-tuned our BERT, RoBERTa, and HateBERT models with the best hyperparameters. Subsequently, we created an ensemble model by averaging predictions from all three models. The 'Ensemble_BERT' function converted the input data into PyTorch tensors, concatenated them to obtain predictions, and applied SoftMax to generate probabilistic outputs. The ensemble model, which utilizes averaging, harnessed the collective strengths of the individual models, resulting in a more robust and accurate abusive language detection system.

4.6. Multi-Class Evaluation

In our endeavor to establish a robust online abusive language detection system, our study employed a multifaceted approach focused on comprehensive evaluation metrics. This section is dedicated to elaborating on our evaluation methodology, beginning with the segmentation of the multiclass challenge, and culminating in an extensive exploration of the used evaluation metrics.

4.6.1. Segmentation Strategies: One-vs-Rest (OvR) and One-vs-One (OvO)

To address the inherent complexity of multiclass classification, our study judiciously employed two cardinal strategies: the "One-vs-Rest (OvR)" and "One-vs-One (OvO)" paradigms[4]. The OvR strategy, also known as one-vs-all, dissects each class individually against the collective background of all other classes. In doing so, it casts the problem into a series of binary classification tasks, where each class is designated as the positive class, while the rest amalgamate to form the negative class. In our specific context, this yields three discerning binary tasks: ImpAbu vs. Rest, ExpAbu vs. Rest, and NonAbu vs. Rest.

In contrast, the OvO strategy dives into the intricate realm of pairwise comparisons. Here, every unique combination of classes forms the basis of binary classification tasks. By unraveling the specific dynamics between class pairs, this strategy offers a more nuanced evaluation of discriminatory power. Within our study, three distinctive pairwise tasks emerge: ImpAbu vs. ExpAbu, ImpAbu vs. NonAbu, and ExpAbu vs. NonAbu.

Table 13 Multiclass Classification Task Segmentation

	Task Label	**Task Description**
OvR	ImpAbu Vs. Rest	Implicit Abusive Language Vs. Rest *(Rest = Explicit+ None Abusive)*
	ExpAbu Vs. Rest	Explicit Abusive Language Vs. Rest *(Rest = Implicit+ None Abusive)*
	NonAbu Vs. Rest	None Abusive Language Vs. Rest *(Rest = Implicit+ Explicit)*
OvO	ImpAbu Vs. ExpAbu	Implicit Abusive Language Vs. Explicit Abusive Language
	ImpAbu Vs. NonAbu	Implicit Abusive Language Vs. None Abusive Language
	ExpAbu Vs. NonAbu	Explicit Abusive Language Vs. None Abusive Language

[4] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

This stratified approach, founded upon OvR and OvO paradigms, ensures a comprehensive examination of model performance across various facets of the multiclass challenge.

4.6.2. Evaluation Metrics

In tandem with this segmentation approach, our study integrated a diverse suite of evaluation metrics to comprehensively assess classifier performance. This comprehensive set of metrics offers a nuanced perspective on the efficacy of our models in classifying data across multiple classes.

4.6.2.1. Confusion Matrix

The confusion matrix is a performance measurement tool for machine learning classifiers. The matrix organizes predicted classes against the actual classes, revealing a matrix of true positives (*tp*), true negatives (*tn*), false positives (*fp*), and false negatives (*fn*) across the various classes.

Table 14 Confusion Matrix

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	True Positive (*tp*)	False Negative (*fn*)
Predicted Negative Class	False Positive (*fp*)	True Negative (*tn*)

4.6.2.2. Evaluation Metrics Derived from the Confusion Matrix

A range of commonly used metrics can be computed from the confusion matrix, as presented in Table 15. These metrics serve to assess the performance of the classifier, with each metric offering insights into specific aspects of evaluation focus.

Table 15 Evaluation Metrics

Metrics	Formula	Evaluation Focus
Accuracy (Acc)	$\frac{tp + tn}{tp + fp + tn + fn}$	Ratio of correct predictions over total instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp + fn}$	Fraction of positive patterns that are correctly classified.
Specificity (sp)	$\frac{tn}{tn + fp}$	Fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp + fp}$	Positive patterns correctly predicted over total predicted positive patterns.
Recall (r)	$\frac{tp}{tp + tn}$	Fraction of positive patterns correctly classified.
F1-Score	$\frac{2 * p * r}{p + r}$	Harmonic mean between recall and precision values
Matthews correlation coefficient (MCC)	$\frac{tp * tn - fp * fn}{\sqrt{}(tp + fp)(tp + fn)(tn + fp)(tn + fn)}$	Balanced evaluation considering true positives, true negatives, false positives, and false negatives.

4.6.2.3. Receiver Operating Characteristic Curve (ROC Curve)

The ROC Curve is a graphical tool commonly used to evaluate binary classifiers. It presents the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) as the classification threshold varies. TPR reflects the ability to correctly identify implicit instances, while FPR indicates the rate of misclassifying all other as positives. The curve maps TPR against FPR for different threshold values, visualizing how the classifier performs across various thresholds. It starts at (0, 0) and ends at (1, 1), with each point denoting a specific threshold's performance. In our research, we employ the ROC Curve to evaluate six binary tasks using both One-vs-Rest (OvR) and One-vs-One (OvO) strategies, providing insights into classifier discrimination within the multiclass context.

4.6.2.4. Area Under the ROC Curve (AUC)

The Area Under the Curve (AUC) is a quantitative performance metric commonly derived from Receiver Operating Characteristic (ROC) Curves, as illustrated in figure 4. It measures a classifier's ability to rank instances correctly, regardless of the specific classification threshold used. AUC values range from 0.5 (random chance) to 1 (perfect ranking), with higher values indicating superior overall discrimination capability. Researchers use AUC to assess and compare classifier performance, providing a single numerical summary of its effectiveness in distinguishing between classes.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents the results of the preceding experiments and provides a thorough analysis of the impact of preprocessing steps and the various traditional ML models, deep learning models, and transfer learning models in online abusive language detection. Notably, transfer learning models outperformed both traditional and deep learning counterparts, showcasing exceptional proficiency, especially in detecting implicit abuse.

5.1. Performance of Preprocessing Steps

This section explores the impact of distinct preprocessing steps on the performance of our baseline model, which employs LR with N-Grams. Table 16 presents an overview of performance results across multiple classification tasks, each corresponding to one of the three preprocessing experiments outlined in Chapter 4.

Table 16 Preprocessing Techniques Evaluation Results

LR + N-Grams		**Preprocessing 1**			**Preprocessing 2**			**Preprocessing 3**		
		F1	AUC	**Mcc**	**F1**	AUC	**Mcc**	**F1**	AUC	**Mcc**
OvR	ImpAbu Vs. Rest	.65	.71	.34	.65	.74	.33	**.66**	**.75**	**.36**
	ExpAbu Vs. Rest	.66	.70	.32	.73	.78	.46	**.75**	**.80**	**.49**
	NonAbu Vs. Rest	**.66**	**.73**	**.34**	.63	.70	.28	.66	.70	.32
OvO	ImpAbu Vs. ExpAbu	.68	.71	.35	.69	.79	.39	**.69**	**.80**	**.37**
	ImpAbu Vs. NonAbu	**.65**	**.72**	**.30**	.63	.69	.27	.63	.68	.29
	ExpAbu Vs. NonAbu	.66	.71	.33	.71	.75	.42	**.74**	**.76**	**.48**

N.B: The bold font indicates the best reported results.

Our analysis underscores the efficacy of Preprocessing 3, which includes lemmatization, emoji removal, stop word elimination, and seven other preprocessing steps detailed in the previous chapter. The incorporation of lemmatization within

Preprocessing 3 leads to significant improvements in our results by simplifying and standardizing word forms, enhancing the model's ability to detect and extrapolate patterns from textual data. Consequently, preprocessing 3 emerges as the preferred choice for subsequent study phases, consistently exhibiting superior performance in online abusive language detection and outperforming alternative preprocessing methods.

5.2. Performance of Traditional Machine Learning Models

Our investigation into various traditional machine learning models and feature extraction techniques reveals a range of performance metrics as summarized in Table 17.

Table 17 Traditional Machine Learning Models Evaluation Results

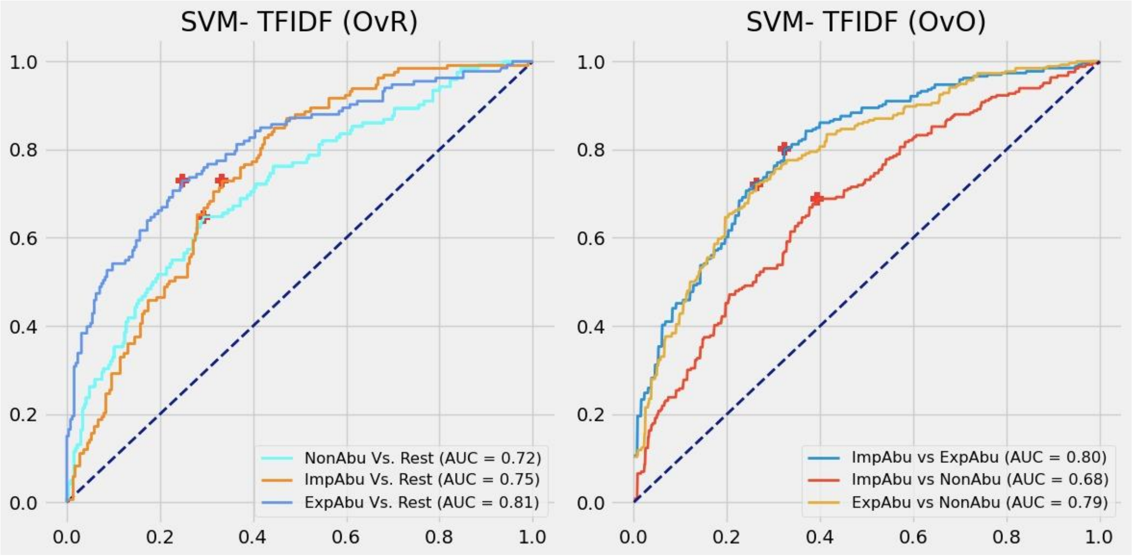
Logistic Regression + Extracted Features:		**N-Grams**			**TF-IDF**			**GloVe**			**BERT**		
		F1	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**
OvR	ImpAbu Vs. Rest	.66	.75	.36	**.67**	**.74**	**.39**	.64	.69	.28	.64	.69	.32
	ExpAbu Vs. Rest	.75	.80	.49	.72	.80	.44	.71	.79	.44	**.76**	**.82**	**.53**
	NonAbu Vs. Rest	.66	.70	.32	.63	.71	.28	.67	.72	.36	**.68**	**.77**	**.39**
OvO	ImpAbu Vs. ExpAbu	.69	.80	.37	**.74**	**.80**	**.47**	.66	.76	.33	.67	.76	.33
	ImpAbu Vs. NonAbu	.63	.68	.29	**.64**	**.67**	**.28**	.62	.65	.23	.62	.67	.25
	ExpAbu Vs. NonAbu	.74	.76	.48	.73	.78	.45	.74	.79	.48	***.81***	***.84***	***.61***
Random Forest + Extracted Features:		**N-Grams**			**TF-IDF**			**GloVe**			**BERT**		
		F1	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**
OvR	ImpAbu Vs. Rest	**.68**	**.74**	**.36**	.63	.71	.29	.63	.72	.29	.60	.65	.24
	ExpAbu Vs. Rest	.72	.81	.45	.69	.76	.38	.67	.75	.37	**.71**	**.79**	**.43**
	NonAbu Vs. Rest	.65	.72	.32	.63	.68	.26	**.67**	**.74**	**.36**	.66	.74	.36
OvO	ImpAbu Vs. ExpAbu	**.70**	**.78**	**.40**	.64	.75	.29	.64	.74	.29	.64	.72	.28
	ImpAbu Vs. NonAbu	.58	.70	.23	.61	.66	.23	**.63**	**.70**	**.26**	.60	.67	.20
	ExpAbu Vs. NonAbu	.72	.79	.45	.69	.74	.37	.70	.77	.41	**.75**	**.80**	**.50**
XGBoost + Extracted Features:		**N-Grams**			**TF-IDF**			**GloVe**			**BERT**		
		F1	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**
OvR	ImpAbu Vs. Rest	**.66**	**.74**	**.34**	.62	.67	.23	.64	.70	.29	.61	.67	.23
	ExpAbu Vs. Rest	**.76**	**.81**	**.51**	.68	.73	.36	.69	.77	.38	.71	.78	.42
	NonAbu Vs. Rest	.66	.72	.33	.60	.65	.23	.64	.73	.31	**.68**	**.76**	**.38**
OvO	ImpAbu Vs. ExpAbu	**.71**	**.80**	**.42**	.63	.72	.25	.66	.74	.31	.63	.71	.25
	ImpAbu Vs. NonAbu	.62	.68	.23	.60	.62	.21	**.64**	**.67**	**.27**	.60	.68	.20
	ExpAbu Vs. NonAbu	.73	.79	.45	.66	.71	.33	.71	.78	.41	**.76**	**.81**	**.51**
SVM + Extracted Features:		**N-Grams**			**TF-IDF**			**GloVe**			**BERT**		
		F1	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**
OvR	ImpAbu Vs. Rest	.66	.74	.36	***.68***	***.75***	***.37***	.64	.69	.28	.63	.70	.32
	ExpAbu Vs. Rest	.70	.78	.42	.73	.81	.46	.73	.80	.47	***.76***	***.83***	***.52***
	NonAbu Vs. Rest	.64	.70	.30	.66	.72	.33	.66	.72	.36	***.72***	***.78***	***.44***
Ov	ImpAbu Vs. ExpAbu	.71	.78	.41	***.74***	***.80***	***.49***	.67	.77	.33	.68	.77	.36
	ImpAbu Vs. NonAbu	.65	.68	.30	***.65***	***.68***	***.31***	.59	.64	.18	.61	.69	.22

ExpAbu Vs. NonAbu	.68	.76	.38	.73	.79	.45	.74	.80	.48	**.80**	**.85**	**.59**

N.B: The bold font indicates the best reported results

Impact on Implicit Abusive Language: SVM coupled with TF-IDF excels in identifying implicit abusive language, achieving F1 scores: 0.68 for ImpAbu Vs. Rest, 0.65 for ImpAbu Vs. NonAbu, and 0.74 for ImpAbu Vs. ExpAbu. This success can be attributed to TF-IDF's capacity to highlight rare, discriminative terms that help the SVM model detect subtle, context-dependent cues. Distinguishing implicit abuse from non-abusive content is notably challenging, this can be because implicit abuse relies on non-explicit words with hidden meanings, typically discernible within a broader context, involving elements like sarcasm and irony. ROC curve analysis in Figure 4 supports SVM+TFIDF's strong performance, with AUC scores of 0.75 for ImpAbu vs. Rest and 0.68 for ImpAbu vs. NonAbu, affirming its discriminative power.

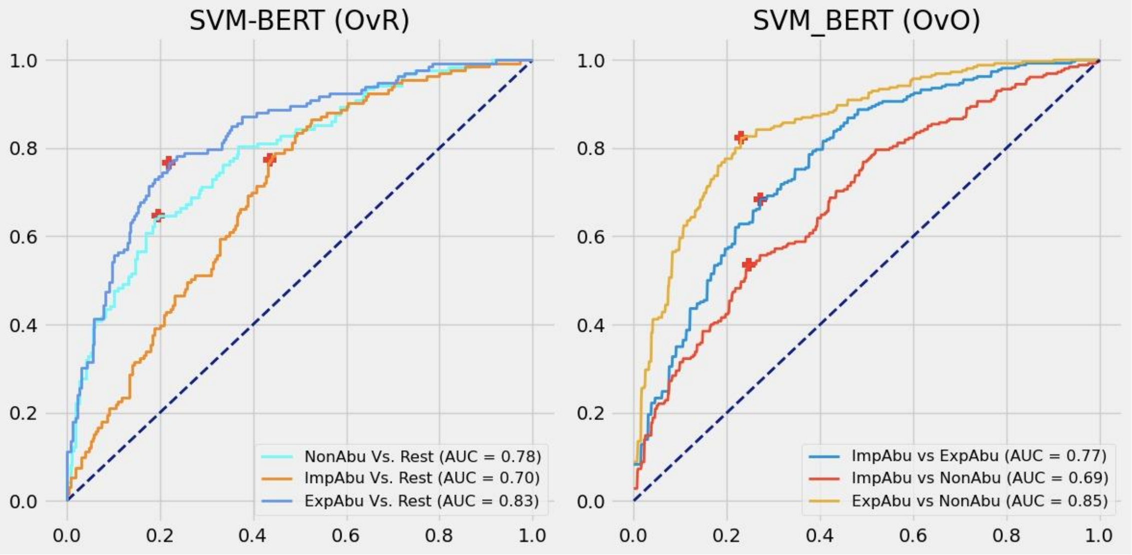
Figure 4 ROC Curve of SVM-TFIDF Model



Impact on Explicit and Non- Abusive Language: SVM coupled with BERT embeddings excels, achieving an F1 score of 0.76 for ExpAbu Vs. Rest and 0.72 for

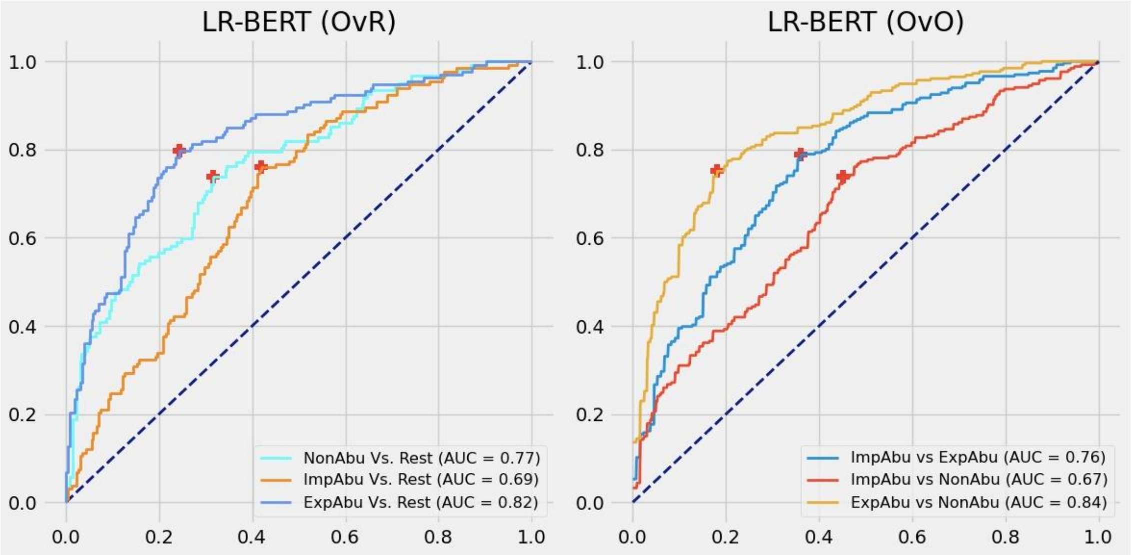
NonAbu Vs. Rest. It effectively identifies explicit abusive content with overt language patterns, leveraging BERT's contextual embeddings for precise linguistic cues. The accompanying ROC curve analysis in Figure 5 supports these findings with AUC scores of 0.83 for ExpAbu Vs. Rest and 0.78 for NonAbu Vs. Rest, ensuring high true positive rates while maintaining a low false positive rate.

Figure 5 ROC Curve of SVM+ BERT embeddings Model



In contrast, LR combined with BERT embeddings performs exceptionally well in distinguishing explicit abuse from non-abusive content, achieving an impressive F1 score of 0.81. The corresponding ROC curve in Figure 6 illustrates an AUC of 0.84, further validating the model's efficacy in this task.

Figure 6 ROC Curve of LR + BERT embeddings Model



The results of the traditional machine learning models demonstrate moderate performance, with implicit abusive language being more challenging to detect, particularly the task of implicit language vs. non-implicit. This underscores the necessity for exploring more robust models, driving our subsequent research phase, where we investigate the impact of deep learning models on enhancing online abusive language detection across our tasks.

5.3. Performance of Deep Learning Models

In this section, we extend our evaluation to deep learning models, unveiling their potential contributions to abusive language detection, as detailed in table 18.

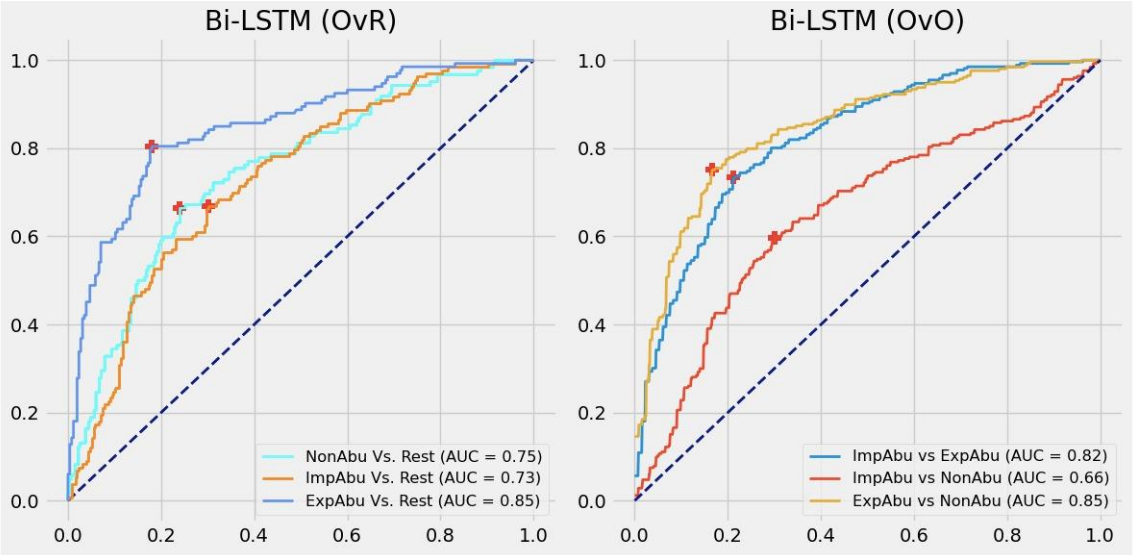
Table 18 Deep Learning Models Evaluation Results

		LSTM			Bi-LSTM			CNN			CNN+Bi-LSTM		
		F1	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**	**F1**	**AUC**	**Mcc**
OvR	ImpAbu Vs. Rest	.66	.72	.37	***.69***	***.75***	***.38***	.68	.74	.36	.68	.76	.38
	ExpAbu Vs. Rest	.77	.83	.53	.77	.84	.54	***.77***	***.86***	***.55***	.74	.81	.48
	NonAbu Vs. Rest	.67	.74	.38	***.70***	***.76***	***.40***	.68	.76	.36	.68	.75	.38
OvO	ImpAbu Vs. ExpAbu	.71	.81	.42	***.73***	***.82***	***.74***	.71	.82	.42	.71	.80	.43
	ImpAbu Vs. NonAbu	.61	.66	.21	.65	.70	.30	.65	.68	.30	***.68***	***.73***	***.36***
	ExpAbu Vs. NonAbu	.79	.82	.57	.77	.82	.53	**.79**	**.86**	**.58**	.72	.79	.44

N.B: The bold font indicates the best reported results.

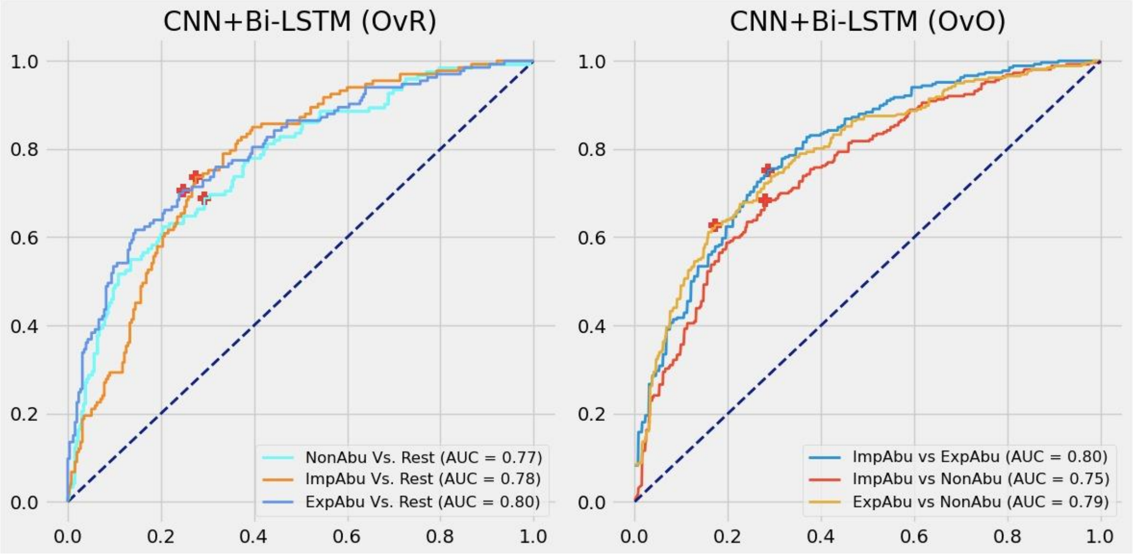
Impact on Implicit Abusive Language: The Bi-LSTM model stands out with F1 Scores of 0.69 for ImpAbu Vs. Rest and 0.73 for ImpAbu Vs. ExpAbu. Bi-LSTM networks excel in contextual understanding, a key aspect in distinguishing subtle nuances between implicit and explicit abusive language. The ROC curve analyses in figure 7 further validates this performance, with the Bi-LSTM model achieving AUC scores of 0.75 for ImpAbu vs. Rest and 0.82 for ImpAbu vs. ExpAbu, ensuring high true positive rates while effectively reducing false positives.

Figure 7 ROC CURVE of Bi-LSTM Model



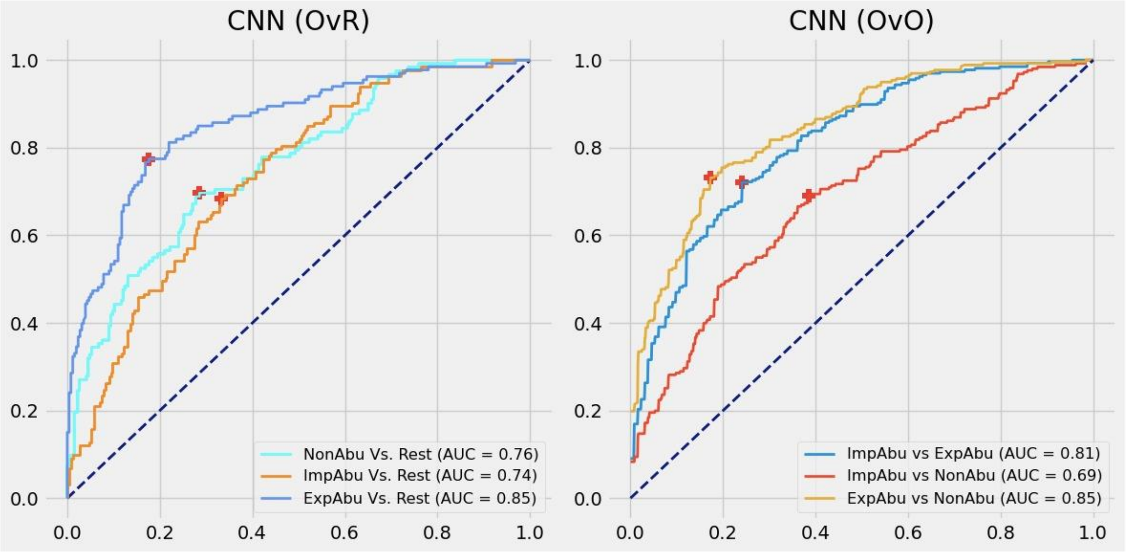
Furthermore, the CNN+Bi-LSTM model proves highly effective in detecting ImpAbu Vs. NonAbu, an F1 Score of 0.68. Combining CNN and Bi-LSTM networks, it excels in capturing both local and long-range linguistic patterns, making it well-suited for implicit abuse detection. The ROC curve of the CNN+Bi-LSTM model in figure 8 exhibits an AUC score of 0.75, emphasizing its ability to distinguish ImpAbu vs. NonAbu.

Figure 8 ROC Curve of CNN+ Bi-LSTM Model



Impact on Explicit and Non- Abusive Language: the CNN model leads in detecting both ExpAbu Vs. Rest and ExpAbu Vs. NonAbu, achieving notable F1 scores of 0.77 and 0.79, respectively, owing to its adeptness in capturing local linguistic patterns and effectively detecting explicit abuse. The ROC curve of the CNN model in Figure 9 displays AUC scores of 0.86 for ExpAbu vs. Rest and 0.86 for ExpAbu vs. NonAbu. These scores underscore the model's ability to maintain high true positive rates while minimizing false positives in explicit and non-abusive language detection tasks.

Figure 9 ROC Curve of CNN Model



On the other hand, the Bi-LSTM model excels in detecting NonAbu vs. Rest, with an F1 score of 0.70, as evident in the model's ROC curve in Figure 7 with an AUC score of 0.76 for NonAbu Vs. Rest, further underscoring the Bi-LSTM effectiveness in this specific context. In summary, deep learning models show significant progress in distinguishing ImpAbu vs. NonAbu compared to traditional models. These results motivate our future research to investigate the potential of transfer learning models in further enhancing abusive language detection across various tasks.

5.4. Performance of Transfer Learning Models

In this section, we explore transfer learning models and their impact on the abusive language detection task. Table 19 showcases the evaluation results of various transfer learning models, across our classification tasks. It is evident that leveraging transfer learning leads to superior results compared to both traditional machine learning models and deep learning, showcasing their efficiency across various classification tasks.

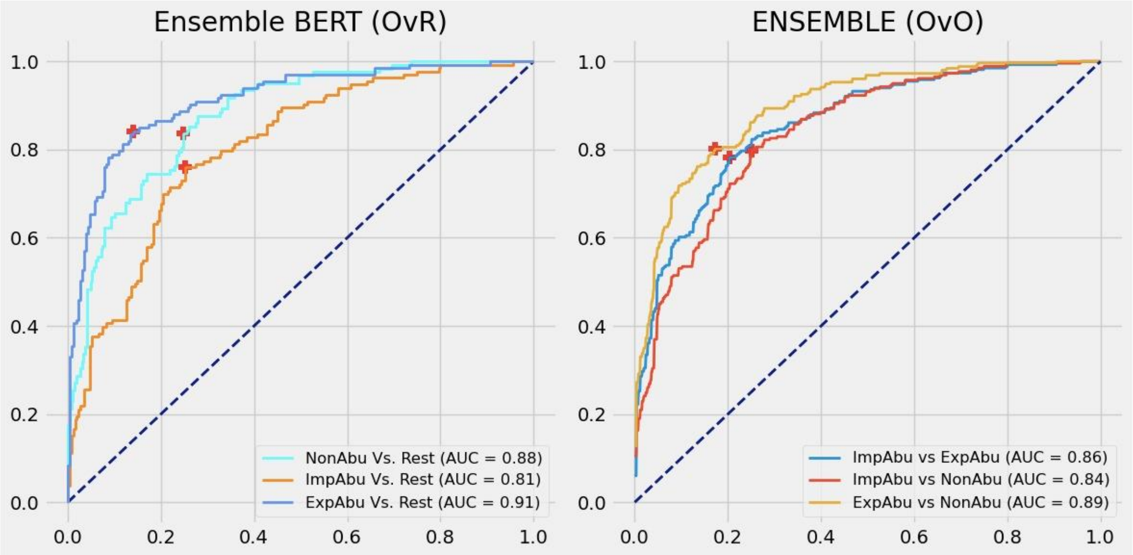
Table 19 Transfer Learning Models Evaluation Results

		BERT			RoBERTa			HateBERT			Ensemble BERT		
		F1	AUC	Mcc	F1	AUC	Mcc	F1	AUC	Mcc	F1	AUC	Mcc
OvR	ImpAbu Vs. Rest	.69	.77	.41	.69	.77	.38	.70	.78	.41	***.72***	***.80***	***.46***
	ExpAbu Vs. Rest	***.84***	***.90***	***.67***	.82	.91	.65	.81	.88	.62	.82	.91	.65
	NonAbu Vs. Rest	.71	.79	.44	***.81***	***.86***	***.63***	.76	.86	.52	.80	.88	.61
OvO	ImpAbu Vs. ExpAbu	.76	.86	.51	.67	.85	.40	.73	.83	.46	***.78***	***.86***	***.55***
	ImpAbu Vs. NonAbu	.66	.73	.34	.65	.74	.38	***.68***	***.78***	***.37***	.67	.84	.33
	ExpAbu Vs. NonAbu	.85	.88	.71	***.87***	***.94***	***.74***	.80	.91	.60	.85	.90	.71

N.B: The bold font indicates the best reported results.

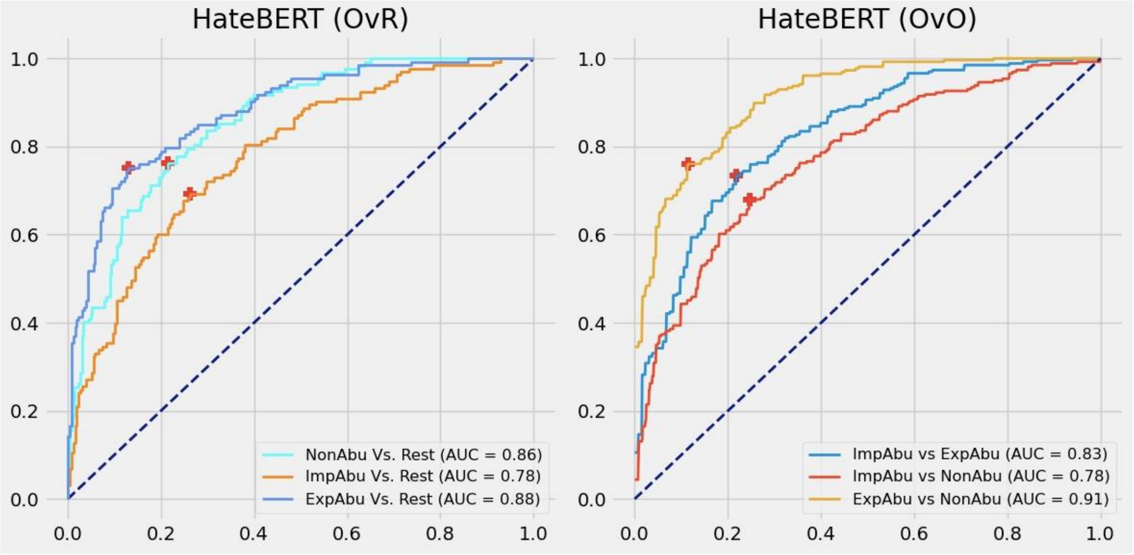
Impact on Implicit Abusive Language: Ensemble BERT excels as the top model for ImpAbu Vs. ExpAbu and ImpAbu vs. Rest tasks, achieving substantial F1 scores of 0.78 and 0.72, respectively. This success is attributed to BERT's pre-trained contextual embeddings, which capture intricate nuances of implicit abuse. The ROC curve analysis confirms Ensemble BERT's (figure 10) performance with AUC scores of 0.86 for ImpAbu vs. ExpAbu and 0.80 for ImpAbu vs. Rest.

Figure 10 ROC Curve of Ensemble BERT Model



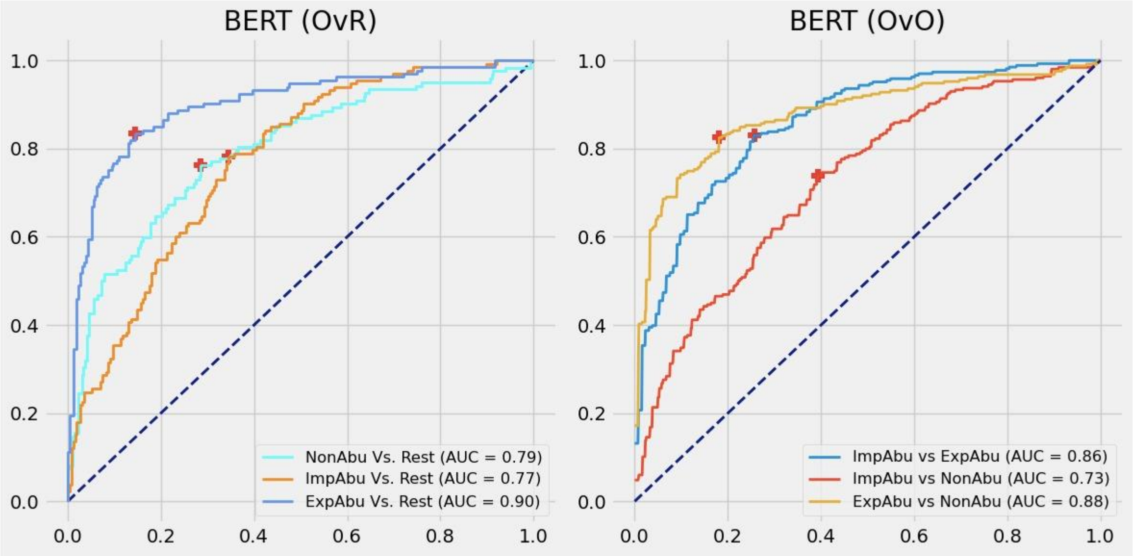
In the ImpAbu Vs. NonAbu task, HateBERT distinguishes itself with an F1 score of 0.68, indicating its effectiveness in discerning implicit abusive language from non-abusive content. Furthermore, the AUC score of 0.78, as demonstrated by its ROC curve (Figure 11), provides additional evidence of HateBERT's strong discriminatory power.

Figure 11 ROC Curve of HateBERT Model



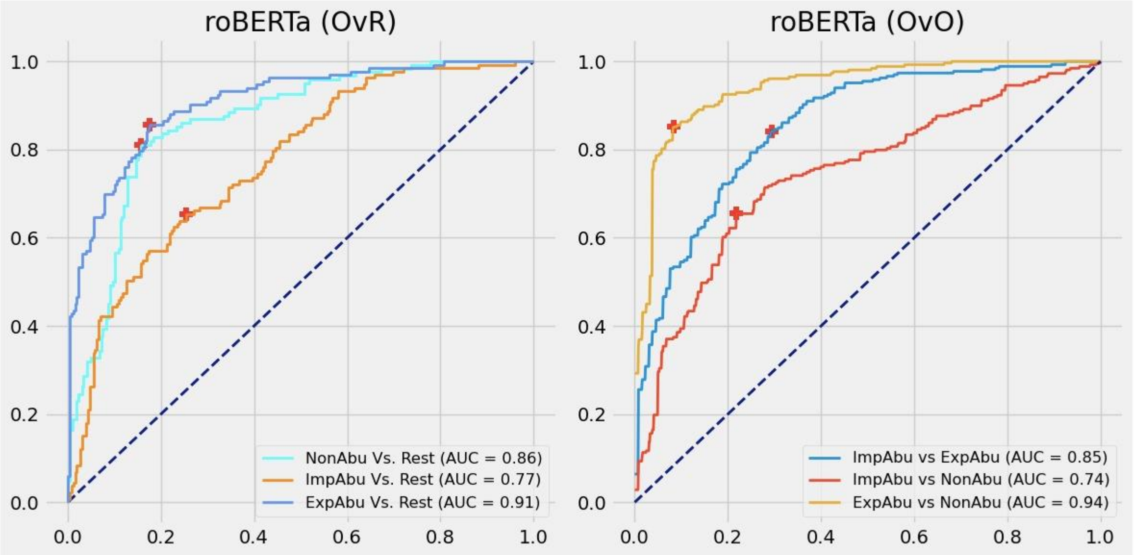
Impact on Explicit and Non-Abusive Language: BERT emerges as the top-performing model for ExpAbu vs. Rest, achieving an impressive F1 score of 0.84. The ROC curve of the BERT model exhibits (figure 12) an impressive AUC score of 0.90 for ExpAbu vs. Rest, confirming its ability to maintain a high true positive rate while minimizing false positives.

Figure 12 ROC Curve of BERT Model



Similarly, RoBERTa takes the lead in detecting NonAbu vs. Rest and ExpAbu vs. NonAbu tasks, achieving remarkable F1 scores of 0.81 and 0.86, respectively. RoBERTa's strong performance is attributed to its advanced training techniques, enabling precise differentiation between non-abusive and explicit abusive language. RoBERTa's ROC curve (figure 13) showcases AUC scores of 0.86 for NonAbu vs. Rest and an outstanding 0.94 for ExpAbu vs. NonAbu. These AUC scores underscore the robust discriminatory power of transfer learning models in explicit and non-abusive language detection tasks.

Figure 13 ROC Curve of RoBERTa Model



In summary, transfer learning models have notably enhanced abusive language detection across various tasks, particularly excelling in the detection of implicit abusive language when compared to deep learning and traditional models.

5.5. Discussion

The discussion of our findings begins with the acknowledgment of the significant challenge presented by the detection of implicit abuse. This challenge has been consistently highlighted in our comprehensive literature review and has been empirically validated through our experiments. The most notable achievement is the robust performance of the ensemble BERT model in distinguishing Implicit Abuse from the Rest category, with an impressive F1 score of 0.72. In juxtaposition, our results exhibit a favorable comparative advantage over prior work, particularly a study by El Sherief et al. (2021), wherein a BERT-based model coupled with augmentation techniques achieved an F1 score of 0.70 in binary classification for implicit hate vs. not hate. Notably, Implicit Abuse vs. Rest shares conceptual proximity with El Sherief et al.'s task, establishing a meaningful baseline for benchmarking and underscoring the incremental value of our contributions.

Expanding our perspective beyond the Implicit Abuse vs. Rest task, the diverse landscape of task performance within various implicit abuse-related contexts demands a nuanced examination. We find that there is no universally optimal model and that detecting implicit abusive language is task specific. For instance, Ensemble BERT excels in distinguishing ImpAbu Vs. ExpAbu and ImpAbu vs. Rest, while HateBERT proves more effective in detecting ImpAbu Vs. NonAbu. This highlights the need for tailored approaches for specific abusive language detection tasks, underscoring the critical significance of our research, as it fills a crucial gap in the realm of abusive language content moderation.

Moreover, among the tasks we investigated, it's crucial to emphasize that our empirical results highlight that distinguishing Implicit Abuse from Non-Abuse is the most challenging task. We posit that factors such as linguistic intricacies and complex contextual dependencies underpin the reasons for the heightened difficulty in detecting implicit abuse versus non-abusive content. Implicit abuse often hinges on seemingly non-explicit words that appear harmless yet carry concealed meanings. These meanings typically reveal themselves only within a broader context, encompassing elements such as sarcasm, irony, or nuanced language usage. To address this challenge effectively, models must not only rely on the recognition of individual words but also possess the capacity to discern nuanced linguistic cues within a broader context.

In conclusion, implicitly abusive content, given its concealed nature, is a persistent threat to online discourse and user well-being. Our research represents a significant advancement in the realm of online abuse detection, offering a vital step towards the development of more robust and proactive content moderation systems. As we continue to push the boundaries of implicit abuse detection, our contributions serve as tangible assets in the collective endeavor to maintain secure digital environments for users worldwide.

5.6. Error Analysis

In this section, we conduct a comprehensive error analysis to assess how well our models categorize tweets into implicit, explicit, or non-abusive content, covering both traditional machine learning and transfer learning. This analysis offers insights into our models' strengths and weaknesses, guiding future research for online abusive language detection.

Table 20 Comparison of Model Predictions for Tweet Classification

	Tweets	**True Label**	**Predicted Label**	
			Traditional Model	**Transfer Learning**
1	@USER Aha, and it is a vacuum that fills the void in her life. If she didn't have her hate, what would she be since a decent human being is not what Karen aspires to.	***ImpAbu***	NonAbu	***ImpAbu***
2	@USER Yeah, right, and on what planet do you come from??	***ImpAbu***	NonAbu	***ImpAbu***
3	@USER You were there Vicky. Tell us the rest of story since your Meghann's best friend.	***ImpAbu***	NonAbu	***ImpAbu***
4	@USER hell i don't even want them on the same continent i am on...what say we send them to a deserted island.	***ImpAbu***	NonAbu	***ImpAbu***
5	@USER Nonsensical bullshyt (…) Pretty shitty parenting.	***ExpAbu***	NonAbu	***ExpAbu***
6	@USER She won't even sue people who call her a yacht ho because she would have to come to court to prove she wasn't. Lol	***ExpAbu***	NonAbu	ImpAbu
7	@USER She is a beautiful woman.	***NonAbu***	ImpAbu	***NonAbu***
8	@USER There's a lot of actors who have jobs they look back on with regret.	***NonAbu***	ImpAbu	ImpAbu

This section explains instances where the model's predictions diverge from actual labels. For example, in the first tweet, the true label is "Implicit Abusive" (ImpAbu). The traditional model classifies it as non-abusive due to the absence of explicit abusive language or threats. In contrast, transfer learning accurately identifies it as implicit abuse, recognizing the negative tone. Notably, the term "Karen[5]" carries evolving internet slang connotations, implying attributes like vileness, racism, and entitlement, even though the model may not specifically understand this evolving definition, it captures the overall underlying sarcasm.

In the fourth tweet, the traditional model incorrectly labels it as non-abusive, while transfer learning accurately detects the underlying negative tone. The phrases used in the tweet, such as *"I don't even want them on the same continent"* and the suggestion to *"send them to a deserted island"*, strongly convey disapproval and a desire to isolate or exclude individuals. Remarkably, the transfer learning model successfully captures this implied negativity, even without the presence of explicit offensive language or direct threats.

Our transfer learning model excels in detecting implicit abuse without relying on explicit markers and effectively capturing sarcasm and negative connotation. However, it exhibits varying performance in different scenarios. For example, it successfully identifies misspelled profane words, as in the fifth tweet where it correctly labels *"bullshyt"* as ExpAbu, surpassing the traditional model's performance. Nevertheless, its consistency in capturing all misspelled explicit expressions is not guaranteed. In the sixth tweet, it erroneously classifies tweets containing explicit content as implicit abuse, highlighting the challenge of recognizing explicit expressions that deviate from

[5] https://www.urbandictionary.com/define.php?term=Karen

conventional spellings or have alternative spellings, such as the term "ho," a derogatory word for women that can be written as 'ho' or 'hoe.' This emphasizes the significance of language understanding, training data quality, and contextual awareness in identifying misspelled explicit expressions.

In addition, the last two tweets, both with true non-abusive labels, highlight the variation in model performance. The seventh tweet is correctly identified as non-abusive by our transfer learning model. However, the eighth tweet is misclassified as implicit abuse by transfer learning despite the absence of harmful content. This misclassification indicates the model's challenge in distinguishing between tweets with implicit expressions of negativity or criticism and those devoid of abusive or harmful content. This challenge likely arises from the nuanced and intricate nature of language, as well as the diverse spectrum of implicit expressions.

In conclusion, the error analysis underscores the effectiveness of Ensemble BERT in identifying implicit, explicit, and non-abusive content, surpassing the traditional model. However, it also reveals instances where the model's classification is incorrect, emphasizing the need for further refinement and a deeper understanding of the complexities inherent in natural language.

5.7. GPT-3.5's Performance in Detecting Abusive Language

In this section, we introduce GPT-3.5's zero-shot evaluation, a method that assesses GPT-3.5's performance without any prior training on our specific dataset or the abusive language detection task. Zero-shot learning demonstrates GPT-3.5's capability to make predictions in tasks it has not encountered before, utilizing its general language understanding. Table 21 presents classification results for tweets, used in our error analysis, serves as a benchmark for comparing GPT-3.5's performance to traditional and transfer learning models in identifying abusive language.

Table 21 Comparison of GPT 3.5 Predictions for Tweet Classification

	Tweets	**True Label**	**Predicted Label**		
			Traditional Model	**Transfer Learning**	**GPT**
1	@USER Aha, and it is a vacuum that fills the void in her life. If she didn't have her hate, what would she be since a decent human being is not what Karen aspires to.	***ImpAbu***	NonAbu	***ImpAbu***	***ImpAbu***
2	@USER Yeah, right, and on what planet do you come from??	***ImpAbu***	NonAbu	***ImpAbu***	***ImpAbu***
3	@USER You were there Vicky. Tell us the rest of story since your Meghann's best friend.	***ImpAbu***	NonAbu	***ImpAbu***	***ImpAbu***
4	@USER hell i don't even want them on the same continent i am on...what say we send them to a deserted island.	***ImpAbu***	NonAbu	***ImpAbu***	*NonAbu*
5	@USER Nonsensical bullshyt (…) Pretty shitty parenting.	***ExpAbu***	NonAbu	***ExpAbu***	***ExpAbu***
6	@USER She won't even sue people who call her a yacht ho because she would have to come to court to prove she wasn't. Lol	***ExpAbu***	NonAbu	ImpAbu	ImpAbu
7	@USER She is a beautiful woman.	***NonAbu***	ImpAbu	***NonAbu***	***NonAbu***
8	@USER There's a lot of actors who have jobs they look back on with regret.	***NonAbu***	ImpAbu	ImpAbu	***NonAbu***

The analysis reveals GPT-3.5's effectiveness in understanding subtle linguistic cues, evident in the classification of the first three tweets. However, a limitation appears in the fourth tweet, where the model struggles to capture negative connotations, in phrases like "I don't even want them on the same continent" and the suggestion to "send them to a deserted island." Notably, our transfer learning model outperforms GPT-3.5 in classifying this instance. This emphasizes a need for improvement in GPT 3.5 to comprehend all nuances and implicit expressions.

Examining GPT-3.5's proficiency in detecting misspelled explicit expressions, it demonstrates success in recognizing specific words like 'bullshyt' but exhibits limitations with others, notably 'ho.' Parallel findings are observed in our transfer learning model. This highlights a targeted area for enhancement in unconventional spelling recognition, emphasizing the need for refinement to achieve a more comprehensive understanding of language.

Moreover, GPT-3.5 encounters similar difficulties as other models, facing challenges in precisely differentiating tweets with implicit negativity or criticism from those devoid of abusive content. This intricacy underscores the imperative for continuous research and model refinement, aiming to augment language comprehension and secure more precise outcomes in content moderation.

In the following chapter, we will explore potential avenues for future research, investigating methods to enhance GPT-3.5's capabilities, ultimately contributing to more effective content moderation solutions.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Our research examined the intricacies of identifying online abuse, aiming to deepen our understanding of abusive language. We concentrated on implicit forms of online abuse and crafted a framework that goes beyond simplistic categorization of abusive language. This nuanced approach considers subtleties often overlooked by existing studies and recognizes the crucial distinction between implicit, explicit, and non-abusive language. The significance of our findings lies in our ability to successfully detect implicit forms of abuse, overcoming the limitations associated with methods solely focused on explicit content. This not only strengthens digital spaces but also acts as a proactive defense against potential harms stemming from unnoticed language nuances. In essence, our approach contributes to creating a safer online environment by addressing a broader spectrum of potentially harmful content.

Despite our promising findings, we acknowledge certain limitations. Constraints such as resource limitations, encompassing computational infrastructure and access to diverse data sources, have influenced the scope of our work. The project timeline has also placed constraints on the depth and breadth of our investigations. The limited number of annotators and potential bias from a single gender perspective further underscore the need to address these limitations to realize the full potential of our work.

Moving forward, we are considering the exploration of fine-tuning GPT-3.5 specifically for the task of implicit abuse detection. The unique adaptability and generative capabilities inherent in GPT models provide a distinctive perspective for effectively identifying implicit abuse. Additionally, a potential avenue for future research involves the development of an ensemble model that leverages the strengths of both BERT and GPT, with the goal of enhancing accuracy and robustness in detecting implicit abusive content. Furthermore, it is imperative to expand the scope of our research by incorporating larger and more diverse datasets. This expansion is crucial for enhancing our model's accuracy and ensuring its adaptability to the dynamic landscape of online communication, including emerging internet jargon and trends within our dataset.

In conclusion, our research has not only made strides in online abuse detection but has also laid the foundation for a more profound understanding of the linguistic intricacies that define online communication. By addressing implicit abuse and differentiating between the various forms of online abuse, our work contributes to a safer and inclusive digital landscape. The impact of our research extends beyond the immediate, paving the way for a future where online conversations are safeguarded with a depth that mirrors the richness of human expression.

APPENDIX

ANNOTATION GUIDELINE

I- Overview:

This annotation project aims to provide precise and standardized annotations for a Twitter dataset that contains 2,000 tweets, with a specific focus on the detection of online abusive language. The dataset includes tweets that may contain implicit or explicit abusive language, as well as non-abusive content. The annotated data will be used to train and evaluate machine learning models that detect and classify instances of online abusive language. *Please be aware that this task may involve adult content, and workers are advised to exercise discretion while working on it.*

II- Definitions:

Term	Definition
Abusive Language	• Abusive language refers to expressions of hatred, derogation, humiliation, or insult directed toward an individual or a targeted group, with the intention to demean or degrade them.
Non-Abusive Language	• A message that does not contain any behavior that meets the definition of abusive language. • Irrelevant content that does not attack a person or a group of people. • The content might still use offensive terms but does not attack others (e.g., that is a F**king beautiful idea).
Explicit Abusive Language	• Explicit abusive language refers to direct and unambiguous expressions of abusive intent. • It employs clear and unequivocal words or phrases that convey derogatory, offensive, or harmful messages. • Often relies on the usage of specific keywords such as name-calling, insults, threats, or other aggressive language.
Implicit Abusive Language	• Implicit abusive language refers to expressions of abusive intent that are conveyed indirectly or through veiled and subtle means. • Constitutes more complex, abstract, coded, and indirect expressions and attitudes towards other individuals without the reliance on explicit profane language. • Implicit abusive language can come in many shapes and forms such as sarcasm, stereotyping, degrading and misinformation. • Recognizing the abusive nature of the language may require contextual interpretation.

III- Annotation tasks:

Before proceeding, please familiarize yourself with the provided definitions above, of abusive online language. For each tweet, follow these two annotation tasks:

- **Task 1: Abusive or non-abusive Tweets:**
    - Your goal is to determine whether the tweet exhibits abusive behavior or not.
    - Carefully read the entire tweet and consider its overall tone and content.
    - Look for any language that aligns with the provided definition of abusive language.
    - If the tweet contains any elements of abusive behavior, label it as abusive.
    - Otherwise, label it as non-abusive.

- **Task 2: implicit or explicit abusive Tweets:**

- Your focus is on the tweets that have been categorized as abusive in Task 1.
- Now, you need to distinguish between explicit and implicit abusive language.
- Read the previously labeled abusive tweets and assess whether the abusive language used in each tweet is explicit or implicit, based on the definitions provided.

IV- Annotation Examples:

Tweet Type	**Text**	**Label**	**Explicit/Implicit**
Offensive Tweet	• "@USER you are such a stupid idiot! I can't believe anyone takes you seriously. #pathetic" • "@USER you are a narcissist " • "@USER she is stupid" • "*@USER* what a B***"	Abusive	Explicit
Implicitly Abusive Tweet	• "@USER your opinion is interesting...for someone who knows absolutely nothing about the subject." • "@USER If you have read the actual complaint, you will find it is a complete joke. A bit like you. " • "@USER hell i don't even want them on the same continent i am	Abusive	Implicit

Tweet Type	Text	Label	Explicit/Implicit
	on...what say we send them to a deserted island. "		
Non-Abusive Tweet	• "@ USER I disagree with your viewpoint, but I respect your right to express it."	Not Abusive	-

IV. Quality Control

- To ensure consistency and accuracy, a subset of annotated tweets will be reviewed by a second annotator and any discrepancies will be resolved.
- Annotators should provide feedback or ask questions if they encounter any issues or uncertainties during the annotation process.

Thank you for your participation in this annotation project. Your work is valuable in helping to improve the accuracy of machine learning models for detecting cyberbullying in social media data.

REFERENCES

Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433–443. doi: 10.1016/j.chb.2016.05.051

Ali, I., Asif, M., Hamid, I., Sarwar, M. U., Khan, F. A., & Ghadi, Y. (2022). A word embedding technique for sentiment analysis of social media to understand the relationship between Islamophobic incidents and media portrayal of Muslim communities. *PeerJ Computer Science*, *8*. doi: 10.7717/PEERJ-CS.838

Aroyehun, S. T., & Gelbukh, A. (2018). *Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling*. Retrieved from www.gelbukh.com

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 759–760. doi: 10.1145/3041021.3054223

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., & Sanguinetti, M. (n.d.). *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*. Retrieved from http://evalita.org

Bertie Vidgen, Dong Nguyen, Rebekah Tromble, Alex Harris, Scott Hale, & Helen Margetts. (2019). *Challenges and frontiers in abusive content detection*.

Breitfeller, L. M., Ahn, E., Jurgens, D., & Tsvetkov, Y. (n.d.). *Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts*. Association for Computational Linguistics. Retrieved from https://bit.ly/2lVv3BG.

Caselli, T., Basile, V., Mitrovi´c, J. M., & Granitzer, M. (2021). *HateBERT: Retraining BERT for Abusive Language Detection in English*. Retrieved from https://en.wikipedia.org/wiki/

Caselli, T., Basile, V., Mitrovi´cmitrovi´c, J., Kartoziya, I., & Granitzer, M. (2020). *I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language*. Retrieved from https://sites.google.com/view/alw3/

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW). doi: 10.1145/3274301

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017a). *Mean Birds: Detecting Aggression and Bullying on Twitter*. Retrieved from http://arxiv.org/abs/1702.06877

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017b). Mean birds: Detecting aggression and bullying on Twitter. *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*, 13–22. doi: 10.1145/3091478.3091487

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012a). Detecting offensive language in social media to protect adolescent online safety. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 71–80. doi: 10.1109/SocialCom-PASSAT.2012.55

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012b). Detecting offensive language in social media to protect adolescent online safety. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 71–80. doi: 10.1109/SocialCom-PASSAT.2012.55

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (n.d.). *Antisocial Behavior in Online Discussion Communities*. Retrieved from www.aaai.org

Dadvar, M., De, F., Roeland, J., & Dolf Trieschnigg, O. (2012). *Improved Cyberbullying Detection Using Gender Information*. Retrieved from http://www.noswearing.com/dictionary

Dadvar, M., Trieschnigg, D., & De Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8436 LNAI*, 275–281. doi: 10.1007/978-3-319-06483-3_25

Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7814 LNCS*, 693–696. doi: 10.1007/978-3-642-36973-5_62

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. Retrieved from www.aaai.org

Davison, B., Edwards, A., Yin, D., Xue, Z., Liangjie, H., Davison, B. D., Kontostathis, A., Edwards, L., & Edu, L. (2009). *Detection of harassment on Web 2.0*. Retrieved from https://www.researchgate.net/publication/228978102

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from https://github.com/tensorflow/tensor2tensor

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, 29–30. doi: 10.1145/2740908.2742760

Duggan, M. (2017). *Research Tom Caiazza, Communications Manager 202.419.4372 www.pewresearch.org RECOMMENDED CITATION Pew Research Center*. Retrieved from www.pewresearch.org.

Elsherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (n.d.). *Latent Hatred: A Benchmark for Understanding Implicit Hate Speech*. Retrieved from https://github.com/

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). *Latent Hatred: A Benchmark for Understanding Implicit Hate Speech*. Retrieved from http://arxiv.org/abs/2109.05322

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. In ACM Computing Surveys (Vol. 51, Issue 4). Association for Computing Machinery. doi: 10.1145/3232676

Frenda, S. (n.d.). *The role of sarcasm in hate speech. A multilingual perspective*. Retrieved from http://alt.qcri.org/semeval2019/index.

Gao, L., Kuppersmith, A., & Huang, R. (2017). *Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach*. Retrieved from http://arxiv.org/abs/1710.07394

Huang, Q., Singh, V. K., & Atrey, P. K. (2014). Cyber bullying detection using social and textual analysis. *SAM 2014 - Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, Workshop of MM 2014*, 3–6. doi: 10.1145/2661126.2661133

Kennedy, B., Atari, M., Davani, A. M., & Yeh, L. (n.d.). *The Gab Hate Corpus: A collection of 27k posts annotated for hate speech Identity-based motivation View project Designing a Meta User Interface to Support the Interaction with Ambient Intelligence View project*. doi: 10.31234/osf.io/hqjxn

Kshirsagar, R., Cukuvac, T., Mckeown, K., & Mcgregor, S. (2018). *Predictive Embeddings for Hate Speech Detection on Twitter*.

Kumar Mishra, A., Saumya, S., & Kumar, A. (2020). *IIIT_DWD@HASOC 2020: Identifying offensive content in Indo-European languages*. Retrieved from http://ceur-ws.org

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). *Benchmarking Aggression Identification in Social Media* (Issue 1). Retrieved from https://competitions.codalab.org/

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, *14*(8). doi: 10.1371/journal.pone.0221152

Mandl, T., Modha, S., Shahi, G. K., Kumar Jaiswal, A., Nandini, D., Patel, D., & Schäfer, J. (2020). *Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages under Creative Commons License Attribution 4.0 International (CC BY 4.0)*. Retrieved from http://ceur-ws.org

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, *3*. doi: 10.3389/frai.2020.00055

Mishra, P., Tredici, M. Del, Yannakoudakis, H., & Shutova, E. (n.d.). *Author Profiling for Abuse Detection*. Retrieved from https://github.com/pushkarmishra/AuthorProfilingAbuseDetection

Mishra, S. (n.d.). *Grammatical gender as the basis to create gender metaphors in Indian political discourse*.

Mohammad, S. M., Shutova, E., & Turney, P. D. (n.d.). *Metaphor as a Medium for Emotion: An Empirical Study*. Retrieved from www.crowdflower.com

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). *A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media*. Retrieved from http://arxiv.org/abs/1910.12574

Munro, E., & Great Britain. Department for Education. (2011). *The Munro review of child protection : final report : a child-centred system*. TSO.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *25th International World Wide Web Conference, WWW 2016*, 145–153. doi: 10.1145/2872427.2883062

Park, J. H., & Fung, P. (n.d.). *One-step and Two-step Classification for Abusive Language Detection on Twitter*.

Poletto, F., Acmos, M. S., Sanguinetti, M., Patti, V., & Bosco, C. (n.d.). *Hate Speech Annotation: Analysis of an Italian Twitter Corpus*. Retrieved from http://ec.europa.eu/justice/

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. In Language

Resources and Evaluation (Vol. 55, Issue 2, pp. 477–523). Springer Science and Business Media B.V. doi: 10.1007/s10579-020-09502-8

Ptaszynski, M., Masui, F., Nitta, T., Hatakeyama, S., Kimura, Y., Rzepka, R., & Araki, K. (2016). Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *International Journal of Child-Computer Interaction*, *8*, 15–30. doi: 10.1016/j.ijcci.2016.07.002

Raisi, E., & Huang, B. (2016). *Cyberbullying Identification Using Participant-Vocabulary Consistency*. Retrieved from http://arxiv.org/abs/1606.08084

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6085 LNAI*, 16–27. doi: 10.1007/978-3-642-13059-5_5

Risch, J., Stoll, A., Wilms, L., Wiegand, M., & Ai, J. R. (n.d.). *Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*. Retrieved from https://germeval2021toxic.github.

Safaya, A., Abdullatif, M., & Yuret, D. (2020). *KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media*. Online. Retrieved from https://github.com/nlpaueb/greek-bert

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2019). Online hate ratings vary by extremes: A statistical analysis. *CHIIR 2019 - Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 213–217. doi: 10.1145/3295750.3298954

Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., & Russo, I. (2020). *HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task*. Retrieved from https://www.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., Choi, Y., & Allen, P. G. (n.d.). *SOCIAL BIAS FRAMES: Reasoning about Social and Power Implications of Language*. Retrieved from http://tinyurl.com/social-bias-frames.

Schmidt, A., & Wiegand, M. (n.d.). *A Survey on Hate Speech Detection using Natural Language Processing*. Association for Computational Linguistics. Retrieved from https://en.wikipedia.org/wiki/List_

Sood Sara Owsley, Antin Judd, & Churcill Elizabeth F. (n.d.). *Profanity Use in Online Communities*.

Spertus, E. (1997). *Smokey: Automatic Recognition of Hostile Messages*. Retrieved from www.aaai.org

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. In Computers in Human Behavior (Vol. 26, Issue 3, pp. 277–287). doi: 10.1016/j.chb.2009.11.014

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). *Challenges for Toxic Comment Classification: An In-Depth Error Analysis*. Retrieved from http://www.perspectiveapi.com/

Vidgen, B., & Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. In PLoS ONE (Vol. 15, Issue 12 December). Public Library of Science. doi: 10.1371/journal.pone.0243300

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (n.d.). *Introducing CAD: the Contextual Abuse Dataset*. Retrieved from https://github.com/dongpng/cad_

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). *Understanding Abuse: A Typology of Abusive Language Detection Subtasks*. Retrieved from http://arxiv.org/abs/1705.09899

Waseem, Z., & Hovy, D. (n.d.). *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter*. Retrieved from http://github.com/zeerakw/hatespeech

Wiegand, M., Ruppenhofer, J., & Eder, E. (n.d.). *Implicitly Abusive Language-What does it actually look like and why are we not getting there?* Retrieved from http://thelawdictionary.org/abusive-language

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (n.d.). *Detection of Abusive Language: the Problem of Biased Datasets*. Retrieved from www.kaggle.com/c/jigsaw-toxic-comment-

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (n.d.). *Inducing a Lexicon of Abusive Words-A Feature-Based Approach*. Association for Computational Linguistics. Retrieved from www.urbandictionary.com

Wojcik, S., & Hughes, A. (2019). *Sizing Up Twitter Users FOR MEDIA OR OTHER INQUIRIES*. Retrieved from www.pewresearch.org.

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (n.d.). *Learning from Bullying Traces in Social Media*.

Xu, Z., & Zhu, S. (n.d.). *Filtering Offensive Language in Online Communities using Grammatical Relations*. Retrieved from http://www.icra.org/sitelabel/

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (n.d.). *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*. Retrieved from http://competitions.codalab.org/

Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *ACM International Conference Proceeding Series, 04-07-January-2016*. doi: 10.1145/2833312.2849567