LANGUAGE, IDENTITY, AND BIAS: INVESTIGATING AAVE IN HATE SPEECH DETECTION

by

DAVIS DEES

(Under the Direction of KIMBERLY VAN ORMAN)

ABSTRACT

This thesis investigates how hate speech detection models misclassify African American Vernacular English (AAVE) on social media, leading to disproportionate false positives and algorithmic bias. Many systems struggle to distinguish between culturally significant language and harmful content, resulting in the over-flagging of Black speech. The study evaluates models including GloVe + LSTM, TF-IDF + SVM, and fine-tuned DistilBERT across datasets with varying class distributions. A hand-labeled AAVE subset is used to examine false positives and highlight model shortcomings. Results show that even widely used models consistently underperform on AAVE tweets, with low F1 scores and poor generalization. These findings reveal how training data composition and linguistic bias shape detection outcomes. Ultimately, the work calls for more inclusive datasets and fairness-aware model design to reduce disproportionate harm and better support the complexity of online Black language.

INDEX WORDS: Hate Speech Detection, Algorithmic Bias, African American Vernacular English (AAVE), Ethical AI Development

LANGUAGE, IDENTITY, AND BIAS: INVESTIGATING AAVE IN HATE SPEECH DETECTION

by

DAVIS DEES

B.A. Linguistics, University of Florida, 2021

B.A. German Language and Literature, University of Florida, 2021

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2025

LANGUAGE, IDENTITY, AND BIAS: INVESTIGATING AAVE IN HATE SPEECH DETECTION

by

DAVIS DEES

Major Professor:	Kimberly Van Orman
Committee:	Frederick Maier
	Lewis C. Howe

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2025

DEDICATION

To the version of me who thought this day would never come.

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my fiancé, who has been my unwavering support throughout this entire journey. I truly could not have done this without him.

To my family – thank you for believing in me, even when this process stretched longer than expected. Your constant support, understanding, and gentle reminders that I was capable of finishing kept me going.

I'd also like to extend my sincere appreciation to my advisors, whose guidance and patience helped shape this work. In particular, I am grateful to Dr. Kimberly Van Orman, for being endlessly patient with me throughout this project and believing in me when I felt like I couldn't finish.

This thesis is the result of so much more than my own effort. To everyone who stood by me – thank you.

TABLE OF CONTENTS

Page

CHAPTER ONE

Introduction

1.1 Motivation

As online platforms increasingly shape public discourse, the need for accurate and equitable hate speech detection systems has never been greater. These automated systems are tasked with the difficult role of moderating harmful language in real time—but they often fail when faced with linguistic variation, particularly from marginalized communities [1]. One such variety is African American Vernacular English (AAVE), a culturally and historically rich dialect that differs from Standard American English in grammar, vocabulary, and usage.

Despite its legitimacy, AAVE is regularly misclassified by hate speech detection models, which are typically trained on datasets that center dominant language norms. This misclassification disproportionately affects Black users – especially Black men[2] – whose use of reclaimed, in-group terms such as the n-word are often flagged as hateful or abusive, even when used non-pejoratively. These errors do more than undermine model performance; they represent a pattern of algorithmic censorship that disproportionately silences marginalized voices.

This thesis investigates how and why these systems fail, focusing on the misclassification of male AAVE speech. By identifying consistent trends in model error, it contributes to the

growing body of work aimed at creating more context-aware and fair systems. This research also gestures toward broader ethical concerns around algorithmic injustice and the need for reparative approaches to fairness in AI.

1.2 Research Objective/Question

The central aim of this research is to expose and understand the linguistic and systemic factors that contribute to the misclassification of AAVE. To do so, this study:

1. Evaluates model performance across several benchmark hate speech datasets.
2. Identifies consistent patterns of misclassification, especially related to reclaimed language like the n-word.

Together, these goals are guided by the following research question: *How do hate speech detection models misclassify male AAVE speech, and what linguistic features contribute to these misclassifications?*

1.3 Thesis Outline

The remainder of this thesis is organized as follows:

- Chapter 2 examines AAVE's online presence, its appropriation, and the persistent marginalization of Black linguistic expression
- Chapter 3 examines hate speech detection, including model architectures, dataset construction, and the ethical implications of biased algorithms.
- Chapter 4 outlines the methodology, detailing our dataset analysis and model selection.
- Chapter 5 presents and discusses experimental results, analyzing misclassification patterns and linguistic influences.
- Chapter 6 concludes with key findings, implications, and directions for future research.

CHAPTER TWO

African American Vernacular English

2.1 Overview

Because this work focuses on improving hate speech detection models for the minority language variety of African American Vernacular English (AAVE), it is essential to understand AAVE's historical and contemporary context. Moreover, understanding the ways in which AAVE differs from Standard American English (SAE) highlights the challenges in accurately processing minority language varieties online. Examining its linguistic evolution, social significance, and structural features provides insight into why hate speech models struggle with AAVE and why addressing this issue is crucial.

African American Vernacular English is a dialect of the English language spoken in the United States, historically associated with Black communities, but not exclusively limited to them. Since the early 1960s, when attention to African American Vernacular English (AAVE) first gained momentum, the terminology used to describe this linguistic variety has shifted over time, often reflecting broader societal trends. For instance, during the era when African Americans were commonly called "Negroes," the language was referred to as "Negro dialect" or "Negro English." These changes in terminology are closely linked to evolving social contexts and perceptions of race [3]. The variety has been known to go over many different names,

including “Black communications”, “Black street speech”, “Nonstandard Negro English”, “Afro American English”, and more [3]. For the purposes of this work, it will be referred to as AAVE.

Research into the origins of AAVE, draws on comparative data from other non-standard English varieties, English varieties within the African diaspora, and Caribbean Creoles. With the increasing availability of sources like ex-slave narratives and hoodoo texts, perspectives on the origins of AAVE have broadened [3]. The origin of AAVE has various theories on its formation as a language, with historical accounts often highlighting the linguistic challenges faced by African slaves when learning English. One major theory, the substratist hypothesis, posits that AAVE shares structural similarities with West African languages, such as Kikongo and Mande, which influenced its development [4-6]. Another prominent view is the creolist hypothesis, which suggests that AAE developed from creole languages like Gullah – a creole spoken among the African American Geechee communities of the Southeastern coastal low country – or Jamaican Creole, shaped by sociohistorical conditions on plantations [7, 8]. Understanding the origins of AAVE is essential for appreciating its linguistic complexity and sociocultural significance.

While the term 'nonstandard' is often used in academic discussions to describe AAVE, it is important to recognize that this classification is based on social and linguistic norms rooted in standard English, not on the inherent value or complexity of the language itself. AAVE, like all dialects, is a fully developed and rule-governed variety of English with its own rich history and structure. Referring to AAVE as 'nonstandard' is not meant to diminish its legitimacy or cultural significance, but rather to align with the terminology commonly used in sociolinguistic studies. Despite what the name might imply, AAVE is not exclusively spoken by Black individuals, nor

do all Black people use AAVE. Today, it is estimated that over 30 million people across the United States speak this variety [9].

Table 1 [3] displays some distinctive features in AAVE. One of the most well-known and discussed is the use of the habitual ‘be.’

Table 1: Comparisons of African American Vernacular English and Standard American English language use

FEATURE	AAVE	SAE	EXPLANATION
HABITUAL "BE"	"I be eating."	"I am always/usually eating."	The use of "be" refers to habitual or repeated actions.
REMOTE PAST (STATE, HABIT)	"I been eating.”	"I have been eating for a long time.”	Denotes a continual action that began in the past.
REMOTE PAST (COMPLETION)	“I been ate.”	“I ate a long time ago.”	Denotes an action that began and ended in the past.
AUXILIARY OMISSION	"They got everything they need.”	"They have got everything they need."	Auxiliaries may be omitted before the main verb in a sentence.
DOUBLE NEGATION	"I don’t ever have no problems.”	"I never have any problems.”	AAVE uses multiple negation to emphasize negation, whereas SAE uses single negation.
GENETIVE MARKING	"That’s the church responsibility."	“That’s the church’s responsibility.”	The possessive marker (‘s) may be omitted.
PREVERBAL MARKING “FINNA”	"I’m finna leave."	"I’m getting ready/about to leave.”	“Finna” is an example of a preverbal marker which indicates the event is imminent.

2.2 The Sociolinguistics of Social Media

Language is inherently dynamic and socially situated, which poses significant challenges for hate speech detection systems. Sociolinguistics, the study of language in its social context, provides critical insights into how variations in language – such as dialectal differences, reclaimed terms, and contextual nuances – can affect the accuracy and fairness of these systems. Understanding the interconnectivity of language use and social identity is crucial for developing hate speech detection models that can navigate the complexities of real-world communication. This section explores how AAVE evolves in digital spaces, focusing on its cultural significance,

reclamation, and appropriation. It highlights the need for models that can navigate the linguistic complexities of marginalized communities in online environments.

2.2.1 AAVE in Digital Spaces: Evolution, Appropriation, and Cultural Significance

African American Vernacular English (AAVE) has evolved rapidly in digital spaces, where social media platforms serve as hubs for cultural expression, linguistic innovation, and identity formation. Online environments like Twitter, TikTok, and Instagram provide spaces where AAVE speakers creatively adapt the dialect to fit various digital contexts[10], such as character-limited posts, video captions, memes, and comment threads. Phonetic features, including consonant dropping and alternative spellings, are frequently mirrored in written forms. Additionally, users often engage in code-switching between AAVE and Standard American English (SAE), depending on the platform and audience. This linguistic flexibility allows speakers to navigate different social spheres while maintaining cultural ties [10].

Social media also plays a crucial role in cultural reclamation. AAVE speakers use digital platforms to resist appropriation by non-Black individuals and to assert ownership over linguistic expressions that hold cultural significance. As noted by Rickford and Rickford[11], “The reasons for the persistence and vitality of Spoken Soul [AAVE] are manifold: It marks Black identity; it is the symbol of a culture and a lifestyle that have had and continue to have a profound impact on American popular life; it retains the associations of warmth and closeness for the many Blacks who first learn it from their mothers and fathers and other family members; it expresses camaraderie and solidarity among friends; it establishes rapport among Blacks; and it serves as a creative and expressive instrument in the present and as a vibrant link with this nation’s past.” This connection to cultural identity and ancestry makes AAVE not just a means of communication, but a powerful tool for cultural resistance and resilience. Through digital

platforms, AAVE serves not only as a linguistic expression but as a form of cultural solidarity, reinforcing bonds within the community while resisting efforts to strip it of its meaning and significance.

While AAVE plays a crucial role in cultural reclamation and solidarity within the Black community, it is often co-opted in online spaces by non-Black individuals, frequently without proper recognition of its origins and significance. For instance, AAVE expressions are commonly used as humor devices in meme creation, where their cultural context is overlooked or diminished [12]. As noted by Hill[13], 'African American English is the single most important source for new slang (and, eventually, unmarked everyday colloquial usage) in White American English.' Such widespread adoption of AAVE's linguistic innovations in mainstream culture occurs without acknowledging their cultural roots or the community from which they originate.

The intersection of reclamation and appropriation of language online is further examined, particularly in relation to the racial epithet "nigga"[14]. This exploration delves into the dynamics of how the term is used across both Black and non-Black digital spaces. It highlights the nuanced reclamation of this term within the African American community, where it has evolved from a racial slur to a term of camaraderie and solidarity among Black individuals. However, the paper also underscores how the term’s usage in online spaces often blurs the lines between reclamation and appropriation. Non-Black individuals, particularly in digital spaces, frequently use the term without understanding its historical context or cultural significance, which leads to its commodification and diminishment. This phenomenon is emblematic of a broader trend where linguistic expressions rooted in AAVE are stripped of their cultural weight when adopted by mainstream culture. Understanding the evolution and appropriation of AAVE expressions is crucial for recognizing how digital platforms both empower and exploit

marginalized linguistic communities. This awareness is key to developing fair and accurate language technologies that respect the cultural significance of AAVE and do not marginalize the communities they aim to protect.

Chapter 3

Background and Related Works

3.1 Hate Speech

Addressing issues in hate speech detection begins with defining what constitutes hate speech and understanding how it is categorized. Especially considering its possible subjectivity. Generally, “hate speech is any form of expression through which speakers intend to vilify, humiliate, or incite hatred against a group or a class of persons on the basis of race, religion, skin color sexual identity, gender identity, ethnicity, disability, or national origin [15].” While offensive language may be permissible in the eyes of the First Amendment, hate speech crosses a legal and moral line that cannot be ignored. Under current laws, hate speech can only be criminalized when it directly incites imminent criminal activity or consists of specific threats of violence targeted against a person or group [15]. Laws surrounding hate speech extend to online spaces as well, and as our online presence continues to expand across a wide range of platforms, the need for effective hate speech detection has become increasingly important.

Conflating hate speech with offensive language can undermine the accuracy of hate speech detection studies, potentially weakening the effectiveness of models in identifying genuinely harmful actors [16]. These systems are implemented to safeguard users in diverse environments, from social media and discussion forums to customer support and dating

sites. Their aim is to prevent the spread of harmful language that can incite violence, harassment, or discrimination. However, despite the growing need for them and their growing presence, hate speech detection systems face significant challenges.

The presence of racial epithets and other charged language presents a challenge in detecting hate speech online. While such terms can sometimes indicate hateful intent, they do not fully define hate speech. Training models to identify hate speech without relying solely on these words could improve both accuracy and precision [16]. Furthermore, racial epithets and similar language are often reclaimed or used in positive, affirming ways by the communities they have historically been used against, complicating their role as reliable indicators of harmful speech. This distinction is critical for hate speech detection, as models that overemphasize individual words risk misclassifying neutral or even supportive statements as offensive. Moreover, bias in these models can lead to the over-flagging of language associated with marginalized identities – such as references to being Black, Muslim, or LGBTQ+ – even when those references are not used in a derogatory manner [17]. Given that hate speech comprises only a small fraction of online discourse, developing models that can accurately distinguish between harmful and benign language remains a persistent challenge [18].

3.2 Hate Speech Detection Models Overview

In order to develop more effective hate speech detection models, it is crucial to examine existing research. This section will explore the biases and limitations present in hate speech data, the interpretations of what constitutes hate speech, and how these factors shape model performance. Additionally, it will provide an overview of the datasets used to train these models and the various approaches employed in their development.

Bias and Limitations in Hate Speech Data

The data used to train hate speech detection models is imperfect. The basis of a hate speech dataset almost always begins with a manually annotated set – leaving them open to inconsistencies and biases. Human annotators may have different interpretations of what constitutes hate speech, leading to subjective disagreements – especially when labeling content that relies on sarcasm, or cultural references unfamiliar to them [19]. Even among trained annotators, judgments can vary significantly, resulting in inconsistencies across datasets Additionally, there is a need to balance the number of annotators per instance, the cost of crowdsourcing, and the time required to complete the task, as increasing one of these factors often comes at the expense of the others [18]. This trade-off affects dataset quality and, by extension, the performance of hate speech detection models. Bias in annotated training data and the tendency of machine learning algorithms to amplify this has shown definitively that AAVE text is often mislabeled with a high positive rate by current hate speech detection models [20, 21].

AAVE will often be misclassified as "bad" or "non-standard" English on social media. Blodgett and O'Connor [22], highlight this issue, demonstrating how bias in natural language processing (NLP) models leads to inaccurate language classification and reinforces linguistic discrimination [22]. This misclassification of AAVE reflects a broader issue of racial bias in NLP systems. Many NLP models, trained primarily on Standard American English (SAE) datasets, often misinterpret AAVE, associating it with negative traits like toxicity or inappropriateness. As a result, AAVE content is disproportionately flagged as harmful or abusive, even when it contains no derogatory language. Incorporating diverse linguistic varieties, especially those prevalent in social media contexts, and improving the representativeness of training datasets are

crucial steps toward creating more equitable and accurate text analysis tools. Research[1] underscores the need for NLP systems to account for linguistic diversity to prevent biased outcomes and ensure fair treatment of all users. These under-representative models are often trained homogeneously. A prime example is the Common Crawl dataset [23], which is frequently criticized within ethical data representation discourse. Despite its immense size – spanning "over 250 billion pages collected over 17 years"[23], its sheer volume does not guarantee fairness or inclusivity. Instead, the lack of equal linguistic and historical data perpetuates biases and marginalizes nonstandard dialects and underrepresented communities [24].

Data collection practices heavily influence how well automated hate speech detection systems function. Unfortunately, much of the data used to train these systems comes from platforms like Reddit and Wikipedia, which are predominantly used by younger, male demographics [25, 26]. This lack of diverse representation skews the training data, making it difficult for models to accurately process content from underrepresented groups. Blodgett et al. [1] describe how language variations tied to different social and cultural groups, specifically AAVE, are treated on social media platforms [1]. Their research discusses how automated text analysis systems, including hate speech detectors, often struggle with accurately processing AAVE. This struggle arises because these systems are commonly trained on data that predominantly features Standard American English (SAE), leading to a lack of understanding and misclassification of AAVE's unique linguistic features. As evidenced by Table 1, AAVE includes distinct grammatical structures, vocabulary, and phonological patterns that differ markedly from SAE. Without sufficient exposure to these linguistic features, automated systems frequently misclassify AAVE expressions as offensive or harmful. This misclassification is more than a technical error – it disproportionately affects speakers of AAVE by failing to recognize the

context and intent behind their communication. As hate speech detection is largely based on human interpretation – a topic further explored in Chapter 3 – these language differences often results in AAVE being unfairly flagged as offensive or harmful by automated hate speech detection systems because of a lack of linguistic diversity in the training data [1]. Such misclassifications highlight a paradox: models designed to protect marginalized groups end up censoring them instead.

Biased Interpretations of Hate Speech

Along with the lack of linguistic diversity in model training data, defining what constitutes hate speech presents a significant challenge to ensuring model inclusivity during the data collection process. The criteria for what is considered "harmful" or "unintelligible" content can introduce bias. For example, when filtering out certain words believed to be offensive, data-cleaning processes may inadvertently remove discussions critical to marginalized communities, such as LGBTQ groups reclaiming slurs. Similar issues were observed during the training of GPT-3 [24].

The challenge of establishing clear and universally accepted definitions of hate reflects a broader issue: distinctions in what is considered hate speech vary significantly depending on cultural and contextual factors. Without recognizing these differences, hate speech detection systems risk erasing important voices under the guise of moderation.

Building on the factors discussed above, it becomes clear that effective hate speech detection requires models that are not only sensitive but fundamentally aware of the complexities of different dialects. While it may not be feasible for a model to grasp the full societal dynamics that contribute to bias, curating diverse and representative datasets is a critical step. By capturing the linguistic and cultural nuances of marginalized communities, these systems can become more

equitable, explainable, and accurate. Numerous efforts have been made to tackle these challenges and reduce bias in hate speech detection models. Datasets, such as Davidson et al.'s [16], and Waseem and Hovy [27], provided foundational resources for hate speech detection, but primarily relied on binary or three-way classification without deeper insight into annotator reasoning. The HateXplain dataset builds upon these earlier efforts by introducing word and phrase-level span annotations, which document the specific parts of a text that led annotators to classify it as hate speech, offensive, or neither [28]. This level of detail in annotation offers greater transparency into how annotators justify hate speech classification, allowing for more nuanced model training and evaluation. As a result, the annotations are more context-aware and better consider intent. This methodological approach has also been applied in other studies focusing on how models interpret and reason through text-based tasks that require general knowledge and human-like inference [29]. This dataset includes content from Twitter and Gab, with MTurk workers assigned to annotate it across three categories: hate, offensive, and normal. What sets this dataset apart is its requirement for annotators to provide justifications for their choices based on the text, enhancing the rationale behind their decisions.

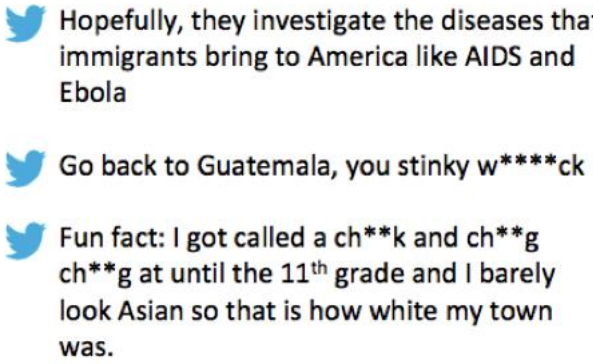
The HateXplain dataset represents a significant advancement for transparency by requiring annotators to justify their decisions. Building on this approach, another strategy to reduce bias involves making annotators aware of nonstandard dialects, such as AAVE, and considering race in their assessments [21]. Research indicates that when annotators are made explicitly aware of the dialect of an AAVE tweet, they are significantly less likely to label it as offensive. In practice, this process involves providing annotators with data that has been analyzed using a dialect classifier. Based on the results of this classifier, inferences can be drawn about the speaker's race; for example, if a tweet is identified as being written in AAVE, it can be

inferred that the speaker is likely Black for the purposes of this study. By equipping annotators with this contextual information, they can make more informed and sensitive judgments about the hatefulness of a tweet, ultimately leading to a reduction in bias.

3.2.1 Composition of Hate Speech Datasets

Online hate speech datasets can be categorized based on the social media platforms from which they originate, and the labeling practices used to classify hate speech. The structure and user behavior of different social media sites – such as Twitter, Reddit, or Gab – shape the nature of the data collected, meaning that hate speech on one platform may look different from hate speech on another. For example, Twitter's character limit often results in concise, coded language, while Gab, known for its permissive moderation policies, may contain more explicit hate speech. Similarly, the categories used to label hate speech vary widely across datasets, creating additional challenges for comparison. For example, one dataset may classify content as racist, sexist, or normal [27] while another may use categories like abusive, hateful, normal, or spam [18], and yet another may opt for a simpler binary classification of hateful vs. non-hateful [30]. These differences in labeling make it difficult to compare model performance across datasets without some form of normalization.

Figure 1: Examples of tweets illustrating that hatefulness or abuse does not directly correlate with the presence of profanity: the first tweet is hateful without using profanity; the second is hateful with the use of a slur; the third is not hateful itself but describes being the target of hate.



Given that language often carries nuance, one-dimensional labeling techniques cannot always fully encompass its complexity. Just as posts online containing slurs, this does not necessarily preclude the content as being hate speech. As illustrated in Figure 1, profanity alone is not a definitive indicator of hate speech [31]. Figure 1 shows three tweets in which (1) accuses immigrants of harming society without using any direct insult; (2) insults a Hispanic person using a slur; and (3) uses slurs to give a personal account of discrimination. Recognizing these complexities, the work of Ousidhoum et al. seeks to address the limitations of one-dimensional labeling by classifying hate speech data along five distinct attributes: directness, hostility, target, group, and annotator. This multidimensional approach, conducted across three different languages (English, French, and Arabic), provides a more nuanced framework for identifying and categorizing hate speech [31]. **Directness:** Within directness, a tweet can either be direct or indirect. This is based on whether or not a target group is named in the tweet. **Hostility:** The hostility attribute categorizes the type of harmful language used. Tweets can be labeled as abusive if they convey dangerous rhetoric, hateful or offensive depending on the intensity of hate or disrespect, and fearful if they express fear rooted in ignorance. Tweets that do not express hostility are marked as normal. **Target:** This attribute identifies the basis on which individuals or

groups are attacked. Annotators classify tweets according to six categories: origin (encompassing race, ethnicity, and nationality), religious affiliation, gender, sexual orientation, special needs, or other. **(Target) Group:** Here, annotators determine the specific group being targeted, choosing from 16 predefined groups such as women, people of African descent, Muslims, immigrants, and political ideologies. When multiple groups are targeted, annotators select the one deemed most affected. **(Sentiment of the) Annotator:** This attribute captures the annotator's emotional response to the tweet using a range of negative and neutral sentiments, including shock, sadness, disgust, anger, fear, confusion, and indifference. This multi-attribute framework provides a comprehensive approach to understanding hate speech, offering nuanced insights into its manifestation across linguistic and cultural contexts.

The choice of labels goes beyond simply describing language – it directly influences how well a model learns to detect and classify hate speech. If the labels are too broad or ambiguous – for example in a binary classification – the model may struggle to distinguish between different types of harmful speech, leading to over- or under-detection. On the other hand, a dataset that distinguishes between "abusive" and "hateful" speech allows for a more nuanced understanding of harmful language. The process of assigning these labels is inherently subjective – with a reliance on human annotators who may disagree on what constitutes hate speech versus offensive or controversial language This subjectivity can introduce inconsistencies, particularly when annotators come from different backgrounds or hold different cultural perspectives on harmful speech. These biases may be mitigated by assigning multiple annotators to each data point in order to get multiple perspectives [28], or by priming annotators on the data beforehand [21].

Table 2: Hate Speech Dataset Labels

Dataset	Labels	Total Size
Founta et al. (2018) [18]	Abusive, Hateful, Normal, Spam	80,000
Waseem and Hovy (2016) [27]	Racist, Sexist, Normal	16,914
Davidson et al. (2017) [16]	Hate Speech, Offensive, Normal	24,802
Ousidhoum et al. (2019) [31]	Directness, Hostility, Target, Group, Annotator (multi-attribute)	13,000
HateXplain (2021) [28]	Hate Speech, Offensive, Normal	20,148
HateCheck (2021) [30]	Hateful, Non-hate	3,728

3.2.2 Hate Speech Detection Approaches

Numerous approaches have been explored in the field of hate speech detection. This section provides an overview of three key groups: traditional classification methods, deep learning techniques based on word embeddings, and deep learning methods leveraging transformer architectures.

Traditional Machine Learning Approaches

Traditional approaches to hate speech detection rely on well-established text representation techniques combined with classical machine learning classifiers. These models typically encode text using methods such as TF-IDF [32] and n-grams [16, 33], which have proven effective when paired with classifiers like support vector machines (SVM), naive Bayes, logistic regression, and decision trees [34-36].

To improve the ability to identify hateful or offensive content, researchers have incorporated additional linguistic features, such as sentiment lexicons and polarity scores [33]. Part-of-speech tagging and syntactic structures [33], as well as dependency parsing [36], have also been explored to improve classification accuracy. Among machine learning models, SVM remains one of the most widely used techniques for hate speech detection [16, 35-37], alongside

other classifiers like naive Bayes [16, 36], logistic regression [16], random forests [16], and gradient boosting decision trees [38].

Deep Learning Approaches

Deep learning-based hate speech detection relies on neural network models that learn complex representations from text data. Unlike traditional methods, which require manually engineered features, deep learning models can automatically learn relevant patterns from text. These models can be trained on various forms of text representation, including traditional encodings like TF-IDF and more advanced word embeddings. Neural architectures such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and bidirectional LSTMs (Bi-LSTMs)[37] have been widely used in this domain. CNNs excel at capturing local word or character-level patterns [37, 39], while LSTMs are particularly effective at modeling long-range dependencies in text [37, 40].

Word Embedding-Based Methods

Word embeddings provide a distributed representation of words, enabling models to capture semantic relationships between terms. Techniques such as Word2Vec [41], GloVe [37], and FastText [37] generate vectorized word representations that position similar words closer together in a multi-dimensional space. These embeddings have been extensively applied in hate speech detection and related tasks such as sentiment analysis [39, 42]. Hate speech classifiers often integrate word embeddings with traditional machine learning models [37] or deep neural networks such as recurrent neural networks (RNNs) [42], LSTMs [37], and CNNs [39]. The ability of word embeddings to capture both semantic and syntactic relationships enhances the detection of hateful speech across different datasets.

Transformer-Based Methods

Recent advances in natural language processing have introduced transformer-based models, which have outperformed earlier deep learning approaches in many text classification tasks. Unlike LSTMs and CNNs, transformers leverage self-attention mechanisms to process entire sequences of text in parallel, capturing contextual relationships more effectively. State-of-the-art transformer models such as BERT [43] and ELECTRA [44] have achieved significant improvements in hate speech detection. Transformers can be integrated with architectures like CNNs, LSTMs, and multi-layer perceptrons (MLPs) [28] to enhance performance.

3.3 Ethics of Hate Speech Detection

The limitations of hate speech detection systems are not merely technical; they raise important ethical questions about the role of AI in shaping online discourse. As we continue to rely on automated systems to monitor and moderate speech, we must confront the ethical challenges that arise from their design and implementation. These systems, in their current form, often reflect the biases inherent in the data they are trained on, disproportionately impacting marginalized communities. This section will explore these ethical concerns, specifically examining the distinction between descriptive and normative accuracy, the role of bias in word embeddings, and the limitations of distributive fairness. It will also incorporate perspectives from fairness frameworks, such as non-distributive justice, to highlight how classification errors should be equitably distributed across demographic groups. Addressing these issues is crucial for creating systems that are not only effective but also just.

Understanding the distinction between descriptive and normative accuracy is important in discussions on ethical AI development [45]. Descriptive accuracy refers to how well a system captures and represents reality, even if it is not perfect, while normative accuracy concerns

whether something aligns with moral or ethical expectations. To illustrate this, consider an example of taking a photo of someone robbing a store. While the photo captures the event (descriptive accuracy), it may still have flaws, such as being underexposed due to incorrect settings, which impacts the accuracy of the representation. For the camera to perform its intended task successfully, it must meet a certain level of descriptive accuracy, meaning it must depict the scene accurately enough to be useful, even if it's not perfect. In contrast to descriptive accuracy, it is expected that people get things right in a moral sense. Continuing with the example, while it is morally expected for someone to pay for their groceries rather than steal them, the act of robbery represents a normative wrong – the individual deviates from what is ethically expected by choosing to steal instead of paying.

For a language model to be normatively correct, it must avoid reflecting morally relevant biases. A model that perpetuates bias is considered normatively incorrect, even if it may succeed in one aspect (e.g., avoiding sexist outputs) but fail in another (e.g., producing racist outputs). Normative correctness can encompass multiple objectives, such as eliminating ethically problematic biases or promoting social justice. However, the primary focus is on ensuring the model does not reflect bias [45].

Discussions around bias mitigation in NLP often begin with word embeddings, as they are foundational in shaping how language models understand and represent different words. Word embeddings are mathematical representations of words that capture the semantic relationships between them based on large language corpora [46]. Word embeddings excel in descriptive accuracy, meaning they capture language as it is used by people, often reflecting real-world associations and relationships between words. However, while they accurately capture language as it is used, they often reproduce societal biases present in the data. For instance,

research has shown that that word2vec embeddings can reinforce sexist stereotypes, producing analogies like "Man is to computer programmer as woman is to homemaker" and "Father is to doctor as mother is to nurse" [47]. As these models learn word relationships, they also inevitably pick up on any biases that may be within the corpora. Without proper care in algorithmic development, biases from the corpora can in fact be amplified by machine learning systems[47]. While an ideal solution might be to eliminate biases in language itself, that's unrealistic, leaving the challenging task of trying to debias the models, which raises both practical and ethical concerns.

These challenges in addressing bias within language models highlight the need for broader frameworks to evaluate fairness in order to create adequate solutions. Kong [48] presents an approach that involves assessing fairness through the lens of distributive and non-distributive. Distributive justice refers to theories that consider justice to involve the fair allocation of benefits and responsibilities among individuals within a society [48]. This approach posits that fairness in an algorithm is defined as the equal distribution of outcomes. Take for example a large tech company that uses an AI algorithm to decide who gets invited for interviews. According to distributive justice, the algorithm would be considered intersectionally fair if the rate of interview invitations is evenly distributed across intersectional subgroups of candidates, such as Latina women, Asian men, white women, and Black men. Each subgroup should have the same statistical likelihood of receiving an interview invitation, suggesting that equality in opportunity ensures fairness across the board [48]. Given the complexities of our current societal landscape, the notion of equality in opportunity – central to distributive justice, which emphasizes the equal distribution of outcomes – may not always be the most suitable approach. While it is often portrayed as an ideal in theoretical frameworks, there are numerous real-world situations where

an *unequal* distribution of resources may be more appropriate. AI-driven interview practices have historically struggled to achieve fairness, even with human intervention such as the removal of names and gender-identifying characteristics. The longstanding disparities in opportunities – particularly in fields like technology, where men have historically received more job prospects due to various societal factors – make it challenging to ensure a truly "fair" and equal distribution of opportunities. If left unchecked, these models would likely perpetuate existing biases and fail to perform effectively [49]. At leading companies like Google, Microsoft, and Facebook, Black employees make up only 5–7% of the workforce[1], which is barely half of the Black population's representation in the U.S. at 13.6%. This lack of proportional representation matters not only in terms of fairness but also because the inclusion of genuinely diverse perspectives is critical for identifying and addressing biases in AI systems. Given that approximately three-quarters of the U.S. population identifies as white, according to census data, it is expected that the majority of applicants will be white, particularly white men, while the number of Black and brown women who are applicants are likely to be fewer. If 30 white men and 3 Black women apply for a job posting, a truly equal probability across all applicants would mean that each individual, regardless of group, has the same chance of being invited for an interview. However, if the selection process aims to maintain proportional representation of each group, then the expected rate of interview invitations would reflect the initial applicant distribution. In this scenario, if interview slots are distributed proportionally based on group representation, we might expect that roughly 10 white men (out of 30) and 1 Black woman (out of 3) receive invitations. While this maintains distributive justice, it does not address the underlying structural barriers that led to the applicant pool being imbalanced in the first place. Black and brown women may have fewer

[1] According to the companies' 2023 diversity reports

opportunities to enter the hiring pool due to structural disadvantages such as access to quality education, networking opportunities, and mentorship. Without considering these underlying disparities, the model risks reinforcing existing inequalities rather than actively working to mitigate them[48].

Instead of focusing solely on equal representation, it can be more insightful to examine fairness in AI algorithms through the framework of non-distributive justice, which Kong[48] applies from Iris Young's critique of distributive justice. Young argues that social justice should be understood through the lens of domination and oppression rather than mere distribution, emphasizing the need to eliminate structural inequalities. This framework focuses less on equal numerical distribution and more on addressing the deeper biases that shape societal outcomes. Instead of just balancing statistics, it aims to challenge the unfair systems that create these inequalities. This is particularly relevant when considering biases that cannot be measured by simple distribution, such as the discrimination seen in Google search results. A search query on Google for "unprofessional hairstyles for work" predominantly shows images of Black women, while "professional hairstyles for work" primarily displays images of white women [50]. Such disparities highlight the limitations of distributive justice, which focuses on measurable outcomes, by demonstrating that fairness cannot solely be assessed through statistical equality. Even if search results were evenly distributed among racial groups, the underlying issue of misrepresentation and stereotype reinforcement would persist.

Assessing algorithmic impact through the lens of harm, rather than fairness alone, provides a clearer understanding of the ways marginalized groups are affected. Continuing to assess algorithms solely through fairness models focused on equal distribution risks further harm to marginalized groups. As described by Kate Crawford [51], this harm can manifest in two

distinct ways: allocative and representational harms. Allocative harms arise when certain groups are disproportionately penalized or flagged by the system, such as when language from marginalized communities is more likely to be marked as toxic or offensive due to dialect differences or reclaimed slurs. Representational harms, on the other hand, occur when these systems fail to recognize or appropriately classify hate speech directed at those same communities, allowing covert or coded harmful speech to go undetected.

Recognizing a problem largely depends on our ability to see it; without visibility, fixing it becomes challenging. Along this vein, model opacity presents a significant barrier in creating ethical algorithms. This opacity refers to the lack of transparency regarding the data, code, and foundational elements that make up the model. A model's construction directly influences its effectiveness in identifying bias and addressing fairness concerns and ability to mitigate bias. Although it may not always be feasible to design models with thoughtful precision due to various constraints, one way to address this challenge is through the use of data statements [52]. These statements offer detailed insights into the origin, composition, and collection methods of the data, with a focus on identifying potential biases. Applying this approach to hate speech detection could enhance bias mitigation efforts by making the training data more transparent and easier to scrutinize. Additionally, data statements encourage developers to reflect critically on the ethical ramifications of their datasets, particularly in how marginalized communities are represented—or misrepresented. By clearly documenting the social and demographic characteristics of the data, researchers can minimize the risk of perpetuating harmful biases, ultimately leading to more equitable, explainable, and accurate hate speech detection models. Increasing model explainability has also been done through specific data curation [28].

Failing to address the issues surrounding algorithmic bias risks complicity in perpetuating disparate mistreatment, which is the misclassification of certain groups more frequently than others [53]. This is particularly relevant to hate speech detection systems, which often misclassify content from marginalized communities, as toxic. The framework moves beyond the concepts of disparate treatment (differential treatment based on group membership) and disparate impact (unequal outcomes across groups). It emphasizes ensuring that classification errors, such as false positives and false negatives, are not disproportionately distributed among demographic groups [53]. In the context of hate speech detection – and specifically in the work outlined in this thesis – adopting this perspective can help reduce bias in the way systems handle content from underrepresented communities, ensuring they are not unfairly targeted by incorrect flagging, while still accurately identifying harmful speech.

In the pursuit of improving AI fairness, Kong[48] suggests tailoring specific questions to each problem. In the case of major tech companies' facial recognition systems, which displayed the highest levels of discrimination against Black women [54], the questions asked might be: "Through what process is the structure of racial patriarchy is being embedded into AI algorithms? How does the biased algorithm perpetuate the racial patriarchy of society? In order to resist this intersecting structure of racial and gender oppression, how should the entire development process be redesigned?[48]" These questions help guide research toward prioritizing non-distributive fairness rather than solely focusing on distributive fairness.

As we've seen through our discussion on ethics, not all fairness is created equal. Weak fairness focuses solely on debiasing flawed algorithms, while strong fairness involves leveraging algorithms to actively confront oppression and foster a more equitable society [48]. This perspective underscores the necessity of designing systems that not only strive for fair outcomes

but also prioritize equitable error distribution among diverse social groups. Understanding this distinction is particularly important in hate speech detection, where the consequences of misclassification are not evenly distributed. Without a shift toward strong fairness, these systems risk perpetuating the very inequalities they aim to mitigate.

Chapter 4

Methodology

4.1 Overview

Due to factors such as limited or unbalanced training data and bias in data annotation there are significant challenges in developing hate speech detection models that will perform well on nonstandard language. This presents an opportunity to contribute to the ongoing discourse on how to improve these models. This work aims to aid in better, more inclusive hate speech detection model design – specifically for African American Vernacular English (AAVE) on Twitter. To do this, we trained three model architectures – LSTM, SVM, and DistilBERT – on three benchmark hate speech datasets to compare their performance and generalizability, with the goal of identifying misclassification patterns when the best model is tested on AAVE tweets. To identify these misclassifications, we perform a corpus study – analyzing a subset of the model predictions by closely examining the language used in the tweets – to understand the linguistic factors contributing to misclassifications. This chapter discusses the methodology for our experimentation.

In their work, Kim et al. investigate the multifaceted biases present in hate speech detection models, revealing that these models not only misinterpret racial nuances in language but also exhibit significant gender biases [2]. Their study finds that African American men are

disproportionately flagged for hateful speech, even when their language is non-offensive. This work aims to understand what linguistic factors are at play here to inform more deliberate hate speech detection model creation.

Publicly available code and data were used throughout this project.

4.2 Data

Our experiments include three widely-used datasets: Davidson [16], Founta [18] and HateXplain [28]. Each tweet is assigned to only one category, meaning it is classified as either abusive, hateful, or another label, but not both. Below, we provide an overview of each dataset, with key statistics summarized in Table 3. To maintain user privacy, these datasets contain only textual content and their corresponding category labels, with all user-specific details removed in accordance with platform policies, such as those enforced by Twitter [55]. Further information on the data labeling process can be found in Appendix A.

The final stage of experimentation involved a corpus study using a fourth dataset, in which a subset of misclassified tweets was closely analyzed to explore the language patterns that may have influenced the models' errors. This dataset was created by Blodgett et al.[1], and is comprised solely of unlabeled AAVE tweets.

4.2.1 Dataset statistics and creation

Datasets for model fine-tuning

The following three datasets are used to simulate the current landscape of hate speech detection. It is possible that they may even offer an improved perspective since each was created with a specific, well-intentioned design. By training different model architectures on these datasets, we aim to understand why models that should, in theory, be performing well today might still be failing.

The Davidson et al. dataset [16] consists of approximately 25,000 tweets, categorized into three classes: neither (16.8%), offensive language (77.43%), and hate speech (5.77%). This dataset was constructed in 2017 using a hate speech lexicon compiled by Hatebase.org, which contains words and phrases identified by internet users as hate speech. Tweets containing terms from the lexicon were collected using the Twitter API, resulting in a sample from 33,458 Twitter users. The corpus was then expanded to 85.4 million tweets, from which a random sample of 25,000 tweets was extracted for manual annotation. The annotators were provided with a detailed definition of hate speech and were asked to consider the context in which the words appeared, ensuring that the presence of offensive words alone did not necessarily indicate hate speech. Each tweet was labeled by three or more workers, with a resulting intercoder-agreement score of 92%. The majority label was used for each tweet, and some tweets were excluded from the final dataset due to a lack of consensus.

The Founta et al. dataset [18] consists of approximately 100,000 annotated tweets, categorized into four classes: normal (56%), abusive (24%), hateful (5%), and spam (15%). The tweets were collected via the Twitter Stream API. The process included extracting metadata from each tweet, such as URLs, hashtags, mentions, emojis, and numerals. Sentiment analysis was also applied to measure polarity and subjectivity, and tweets were tagged with offensive terms using dictionaries like HateBase3 and a general offensive words dictionary. Due to the class imbalance inherent in abusive language detection, a boosted sampling procedure was employed, particularly to increase the presence of inappropriate content (abusive and hateful tweets) within the dataset. Tweets showing negative sentiment and containing offensive language were prioritized for inclusion. Their subset of 80,000 tweets underwent five judgments per tweet for the annotation task. Statistical analysis of annotation agreement reveals that over 55% of the

tweets reached an overwhelming agreement (4 out of 5 annotators), reinforcing the reliability of the labeled data. The annotation process involved demographic analysis of annotators to better understand label biases and ensure diverse perspectives.

HateXplain[28] is our third dataset for model fine-tuning. The total size of this dataset is about 20,000 tweets, categorized into three classes: normal (40.47%), offensive (28.59%), hate speech (30.94%). This dataset was created using posts from two social media platforms: Twitter and Gab. The dataset construction follows a method consistent with prior studies, collecting posts using lexicons provided by previous works like Davidson et al. [16], Ousidhoum et al. [31], and Mathew et al. [56]. For Twitter, the tweets were collected between January 2019 and June 2020, while for Gab, posts were sourced from the dataset shared by Mathew et al. [56]. The annotation of the dataset was conducted using Amazon Mechanical Turk (MTurk) workers, with each post receiving three types of annotations: (1) classification of the text as either hate speech, offensive speech, or normal, (2) identification of the target community of the speech, and (3) identification of words or phrases that could explain why the post was categorized as hate speech or offensive. Annotators were instructed to consider target groups such as race, religion, gender, and sexual orientation when identifying which communities were being targeted by the speech. The workers were also provided with definitions for each category, clear instructions on annotating spans, and examples to guide the classification task. To ensure high-quality results, MTurk qualifications were applied, requiring annotators to have an approval rate of 95% and at least 5,000 approved HITs [2]. Of the three datasets, HateXplain is the most balanced in terms of class distributions. Their work presents a hate speech dataset that captures multiple dimensions

[2] HITs stands for Human Intelligence Tasks. In the context of Amazon Mechanical Turk (MTurk), a HIT refers to a task that is designed for workers (known as Turkers) to complete. These tasks can range from simple data entry to more complex tasks like data labeling or transcription. The number of approved HITs indicates how many tasks a worker has successfully completed and had approved by the requester.

of each tweet. Not only are tweets labeled for their type of speech, but they are also labeled for who the target community of a hateful tweet is directed toward. Dataset statistics can be found in Table 3.

Table 3: Class Statistics for Each Datasets

Dataset	Class and Statistics
Davidson	Neither – 4,163 Offensive – 19,190 Hate Speech – 1,430 **Total ~ 25k**
Founta	Normal – 53,560 Abusive – 22,766 Hateful – 4,496 Spam – 13,996 **Total ~ 95k**
HateXplain	Normal – 8,153 Offensive – 5,761 Hate Speech – 6,234 **Total ~ 20k**

Dataset for linguistic study

For the relevant corpus study, we used the Blodgett et al. dataset, which contains tweets likely authored by African American users [1]. This dataset was specifically curated to reflect a diverse range of AAVE usage and facilitate detailed linguistic analysis. These tweets are not labeled for hate speech detection. The authors created their AAVE dataset by collecting geo-located tweets from U.S. users in 2013, primarily sent from mobile phones. To infer user demographics, they mapped each tweet to a U.S. Census block group – a small geographic area with 600 – 3,000 people – and averaged the demographic data of all tweets from each user. They then identified words strongly associated with African American demographics and used a seedlist approach to gather tweets from users who frequently used these terms. A seedlist approach builds a dataset by starting with key terms linked to a group, identifying users who

frequently use them, and collecting their tweets to capture relevant language patterns. To validate the dataset, they analyzed lexical, orthographic, and phonological variations, comparing AAVE-aligned text to Standard American English (SAE) to ensure it reflected known linguistic patterns. It is unlabeled as hate speech or otherwise. The resulting dataset is made up of tweets that have an over 80% confidence of being AAVE – according to the results of their model. It is made up of 1.1 million tweets.

Gender Classification

To test the ways in which – if any – AAVE from men is more likely to be falsely flagged as hateful, it was necessary to classify these tweets as being either written by a man or woman. This has been done previously in hate speech detection [2] in order to show the ways in which hate speech detection models are disproportionately biased against Black male speech. This classification was done using Kim et al.'s[2] gender classifier. It is based on labeled data from Kaggle's website. The gender data was originally provided by the Data for Everyone Library on CrowdFlower [57].

Their classification didn't reflect the authors' actual gender identities. Rather, it aimed to identify whether the linguistic features of a tweet aligned more closely with those commonly associated with a particular gender group. The classifier assigned one of three labels: female, male, or brand (referring to accounts associated with organizations, businesses, or public figures rather than individual users).

4.2.2 Dataset Standardization and Similarity Analysis

To prepare the datasets for training and analysis, several preprocessing steps were applied: tweets were converted to lowercase, punctuation and special characters were removed, stop words were filtered out, and lemmatization or stemming was used to reduce words to their

root form (e.g., 'running' → 'run'). Additionally, URLs, Twitter mentions (e.g., @user), and numbers were removed. For model training and evaluation, all datasets were split into train, test, and validation sets.

4.2.2.1 Label Normalization

Since we are comparing the performance of three different datasets on the same model, some analysis into the similarity of the datasets was necessary to aid in qualifying model results. By measuring cosine similarity, we identified similarities across datasets within the same classes.

This similarity across dataset labels led us to remove Founta's [18] additional label of "spam." Once the "spam" label was removed, labelling was normalized across the datasets. Unified labels across all datasets are "normal", "offensive", and "hateful." This made computation tasks easier and will make it easier to compare model results. Updated labels and class distributions are pictured in Table 4. Further comparison on dataset labelling is outlined in Appendix A. Cosine similarity and Euclidean distances between labels and across datasets can also be found in Appendix A.

Table 4: Unified labeling across datasets

Datasets with Unified Labeling			
	Founta [18]	Davidson [16]	HateXplain [28]
Normal	53,560	4,163	8,153
Offensive	22,766	19,190	5,761
Hateful	4,496	1,430	6,234
Total	**80,822**	**24,783**	**20,148**

4.3 Models

For our study, we tested three distinct model architectures that have previously been used in hate speech detection.

For baseline result comparison, we use a SVM model, as seen in previous studies [34, 35, 58]. Text data is first preprocessed using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer, which converts the raw text into numerical features based on term frequency and inverse document frequency. Our model primarily relies on unigrams (n-gram range: (1,1)), indicating that individual words contribute the most to classification performance. Unlike some feature-restricted approaches, we found that using all available features (max_features=None) yielded the best results. A linear Support Vector Machine (SVM) classifier is applied to these features, trained on the labeled dataset using a train-test split (80%-20%). The model utilizes a regularization parameter of C=1. The model also incorporates probability estimation for additional interpretability and use in downstream applications.

Secondly, we trained our datasets on a Long Short-Term Memory (LSTM) model created by Badjatiya et al.[37] in 2017. Their model architecture for hate speech detection in this study uses an LSTM network to capture sequential dependencies in text. It consists of an embedding layer initialized with pre-trained GloVe embeddings, a single LSTM layer with 50 units, and a dense output layer with three units for classification into hateful, offensive, or normal. Dropout layers with rates of 0.25 and 0.5 are applied after the embedding and LSTM layers to reduce overfitting. The model is trained using categorical cross-entropy loss and the Adam optimizer, with class-weighted loss scaling to address class imbalance. Training occurs over 10 epochs with a batch size of 512 and 10-fold cross-validation.

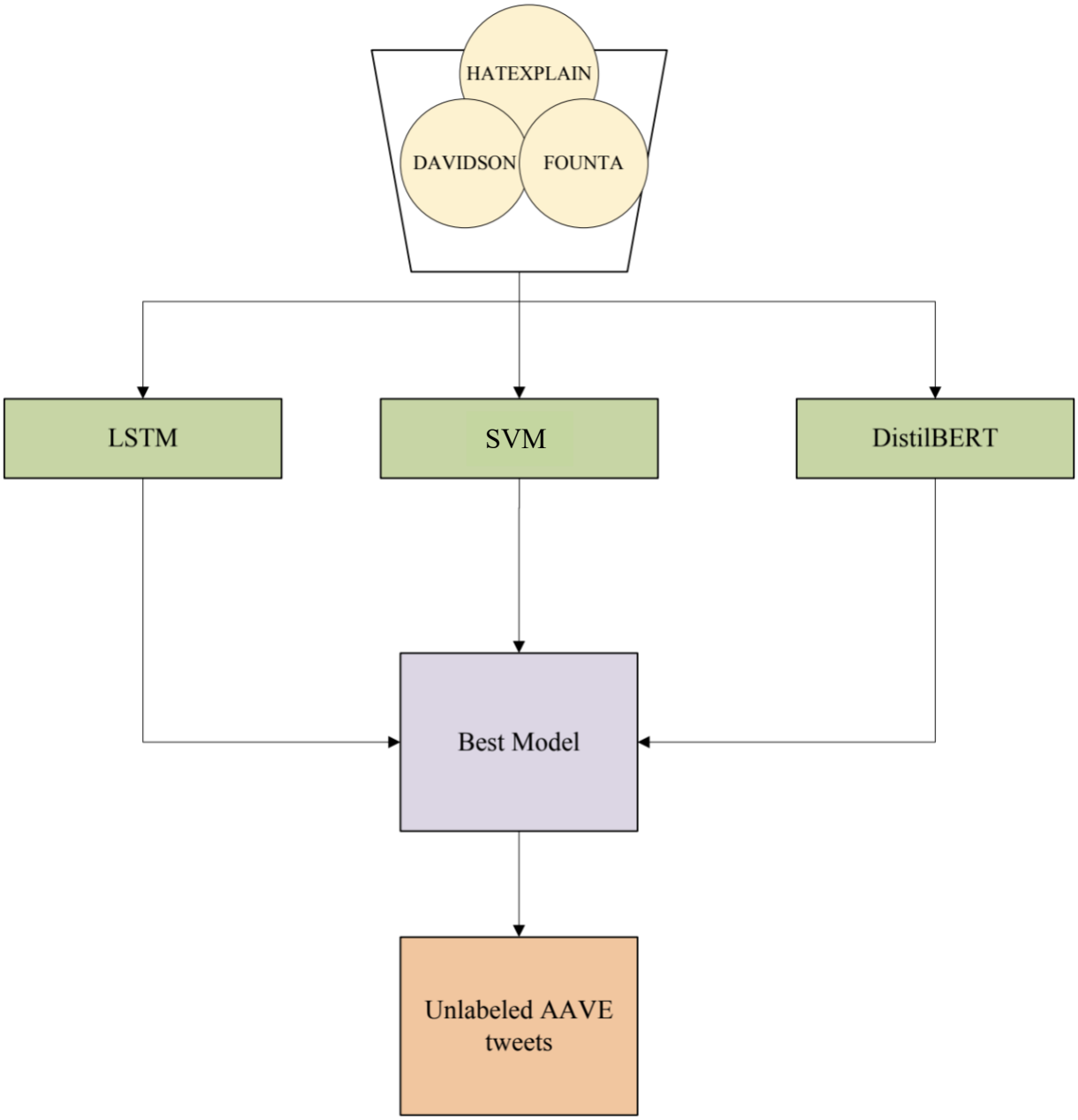
Lastly, we fine-tune a transformer-based model – as seen in previous hate speech detection studies[58] – in DistilBERT. The DistilBERT model is fine-tuned for hate speech detection using a custom pipeline with Hugging Face's *Trainer* class. The training dataset is tokenized using the *DistilBertTokenizer*, with truncation applied to ensure input sequences do not

exceed the model's maximum length. During training, the model is configured with *TrainingArguments* to train for 1 epoch, apply weight decay of 5e-4, and a learning rate of 1e-5. The *Trainer* is set up to use the tokenized datasets, custom metrics, and *DataCollatorWithPadding* to ensure proper batching.

4.4 Experimental Overview

- **Clean and preprocess datasets:** Remove irrelevant content, normalize text, and split the data into training, testing, and validation sets.
- **Fine-tune models using the cleaned training data:** Adjust hyperparameters and perform encoding for optimal feature representation based on the preprocessed data.
- **Evaluate model performance on validation set:** Once model has been adequately altered, test the trained model on the validation data to assess accuracy and generalization ability.
- **Apply best model to unlabeled AAVE data**
- **Hand-label a subset of AAVE data:** Manually annotate a sample of AAVE data from each class to further validate the model's performance on non-standard language use.
- Analyze linguistic patterns in the model’s false positive hateful results.

Figure 2: Chart of model training and evaluation process.



Chapter 5

Experimental Results and Analysis

5.1 Introduction

The objective of this thesis is to highlight and analyze patterns of misclassification in hate speech detection systems when applied to African American Vernacular English (AAVE), particularly in the context of reclaimed, in-group language. By surfacing these patterns, the goal is to expose the limitations of current model and dataset designs and advocate for more inclusive, context-aware approaches to hate speech detection. This chapter presents a detailed discussion of the results obtained from training and evaluating three different model architectures across three benchmark hate speech datasets. After determining which model architecture and dataset combination generalizes best, the analysis then shifts to linguistic observations on the top-performing model's results on unlabeled AAVE tweet data, exploring the linguistic patterns that may have contributed to misclassification.

Further experiments, confusion matrices, and tables can be found in Appendix B; however, these results offer only minimal additional insight beyond what is discussed in the main text.

5.2 Model Performance

Before training, all datasets were split to include a validation set for evaluation after the models were run. Each model architecture was trained using a 60:20:20 split for training, testing, and validation. All reported results reflect the models' performance on the hidden validation set. Our model training process began with using a Glove+LSTM hate speech detection model that was claimed to have "state-of-the-art" performance [37]. However, further review of the literature uncovered a critique of the original authors' work [42], which conducted experiments using the authors' original training data and found evidence of overfitting. This suggested that the model's reported performance was likely inflated. While we include its accuracy results across our three datasets for reference, its inconsistent performance rendered it unsuitable for our analysis.

5.2.1 Model Performance on Original Dataset Class Distribution

Table 5 reports findings from our initial experimentation with original class distributions in the datasets. Findings and comparisons of model performance on balanced class distributions, as well as per-class performance can be found in Appendix B.

Table 5: Imbalanced Dataset Results. Accuracy results from the initial Glove+LSTM model on three hate speech datasets. Although originally reported to achieve state-of-the-art performance, later findings suggest the model was overfit to its training data [42].

Model	Macro			Weighted Average			Overall Accuracy	AUC Average
	Precision	Recall	F1	Precision	Recall	F1		
Results on Davidson Dataset								
TF-IDF + SVM	0.79	0.68	0.70	0.89	0.91	0.89	0.91	0.92
Glove + LSTM	0.26	0.33	0.29	0.60	0.77	0.68	0.77	0.50
DistilBERT	0.76	0.71	0.72	0.90	0.91	0.90	0.91	0.81
Results on Founta Dataset								
TF-IDF + SVM	0.81	0.70	0.73	0.90	0.91	0.90	0.91	0.91
Glove + LSTM	0.02	0.33	0.04	0.01	0.06	0.01	0.06	0.50
DistilBERT	0.80	0.71	0.73	0.90	0.91	0.90	0.91	0.82
Results on HateXplain								
TF-IDF + SVM	0.42	0.40	0.39	0.43	0.44	0.40	0.44	0.79
Glove + LSTM	0.10	0.33	0.16	0.10	0.31	0.15	0.31	0.50
DistilBERT	0.65	0.65	0.65	0.66	0.66	0.66	0.66	0.74

In table 5, we see that the TF-IDF + SVM model consistently outperforms the other models in terms of overall accuracy and AUC average, as well as precision, recall, and F1 scores, especially in the Davidson and Founta datasets, where it achieves a high weighted average F1 score of 0.89 and 0.90, respectively. This suggests that TF-IDF combined with SVM is effective at identifying positive cases while minimizing false positives and negatives. However, the LSTM model struggles across all datasets, with particularly poor performance in the Founta dataset, where its macro averages for precision and F1 are almost zero, indicating its inability to effectively classify the data. Additionally, its AUC scores remain at 0.50 across all datasets, suggesting it is performing no better than random chance. DistilBERT, on the other hand, is more consistent, achieving reasonable weighted average F1 scores of 0.90 in Davidson and Founta datasets and 0.66 on the HateXplain dataset. Although its precision and recall scores are competitive, its AUC scores are notably lower than those of TF-IDF + SVM, particularly in the

HateXplain dataset (0.74 vs. 0.79 for SVM), indicating that its overall ranking of predicted probabilities is less reliable. Overall, TF-IDF + SVM emerges as the strongest model for these datasets, while Glove + LSTM requires further optimization. DistilBERT offers a solid middle ground, demonstrating stable performance across different datasets but not quite matching the precision of TF-IDF + SVM.

The best dataset-model combination appears to be the TF-IDF + SVM model on the Founta dataset, where it achieves the highest weighted average F1 score of 0.90. The Davidson dataset also shows strong performance with TF-IDF + SVM, where it achieves a weighted average F1 score of 0.89, making it the second-best combination overall. The following figures provide a more detailed breakdown of these results.

Figure 3: Confusion Matrix for TF-IDF+SVM trained on imbalanced Founta Dataset. Results show that the model reliably identifies normal tweets but struggles to distinguish hateful speech from both offensive and normal content.

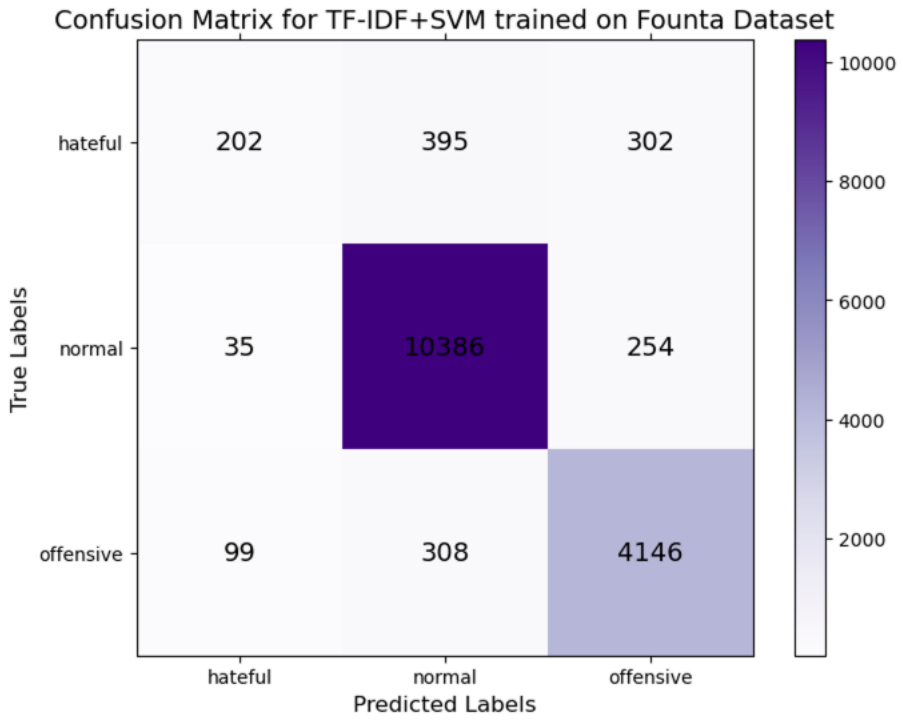


Figure 3's confusion matrix represents results from the TF-IDF+SVM model trained on the Founta dataset. The model performs well in classifying normal tweets, correctly identifying 10,386 instances, with only 35 misclassified as hateful and 254 misclassified as offensive. This suggests that the model has a strong bias towards correctly identifying normal tweets, but may still struggle with some edge cases. However, when it comes to hateful tweets, the model performs much worse. Out of all hateful tweets, only 202 were correctly classified, while 395 were misclassified as normal and 302 as offensive. This indicates that the model has difficulty distinguishing hateful speech from both normal and offensive language, potentially due to overlapping linguistic patterns. Similarly, offensive tweets are somewhat well classified, with 4,146 correct predictions, but 99 instances were misclassified as hateful and 308 as normal, suggesting some level of confusion between offensive and normal speech.

5.3 Cross-Domain Testing

In this section, we evaluate the generalizability of our models by testing them across different datasets. Models trained on one dataset are assessed on the other two to determine how well they perform when applied to data from different domains. This approach helps us understand the robustness of each model, given that linguistic variation and contextual nuances may differ across datasets. By examining the performance of models outside of their training environments, we aim to assess their ability to generalize and identify potential weaknesses that could hinder their effectiveness in real-world applications. Ultimately, the most robust model will likely be the best choice for our task of labeling unlabeled AAVE tweets to determine whether they are hateful or not. Results are shown in Table 6.

Table 6: Cross-Domain Results. Accuracy results from the initial Glove+LSTM model on three hate speech datasets. Although originally reported to achieve state-of-the-art performance, later findings suggest the model was overfit to its training data [42].

Model	Macro			Weighted Average		
	Precision	Recall	F_1	Precision	Recall	F_1
CROSS-DOMAIN EXPERIMENTAL RESULTS WITH THE DAVIDSON DATASET AS THE SOURCE DOMAIN						
Testing on Founta						
TF-IDF + SVM	0.58	0.54	0.49	0.77	0.71	0.70
Glove + LSTM	0.09	0.33	0.15	0.08	0.28	0.12
DistilBERT	0.69	0.62	0.61	0.85	0.88	0.86
Testing on HateXplain						
TF-IDF + SVM	0.48	0.42	0.40	0.48	0.41	0.40
Glove + LSTM	0.10	0.33	0.15	0.08	0.29	0.13
DistilBERT	0.45	0.45	0.45	0.46	0.45	0.45
CROSS-DOMAIN EXPERIMENTAL RESULTS WITH THE FOUNTA DATASET AS THE SOURCE DOMAIN						
Testing on Davidson						
TF-IDF + SVM	0.53	0.63	0.54	0.83	0.74	0.76
Glove + LSTM	0.02	0.33	0.04	0.01	0.06	0.01
DistilBERT	0.53	0.64	0.54	0.83	0.70	0.72
Testing on HateXplain						
TF-IDF + SVM	0.43	0.41	0.40	0.43	0.44	0.41
Glove + LSTM	0.10	0.33	0.16	0.10	0.31	0.15
DistilBERT	0.47	0.46	0.46	0.47	0.48	0.47
CROSS-DOMAIN EXPERIMENTAL RESULTS WITH THE HATEXPLAIN DATASET AS THE SOURCE DOMAIN						
Testing on Founta						
TF-IDF + SVM	0.40	0.40	0.39	0.60	0.62	0.60
Glove + LSTM	0.02	0.33	0.04	0.01	0.06	0.01
DistilBERT	0.51	0.40	0.39	0.67	0.70	0.62
Testing on Davidson						
TF-IDF + SVM	0.47	0.53	0.46	0.74	0.57	0.61
Glove + LSTM	0.02	0.33	0.04	0.01	0.06	0.01
DistilBERT	0.48	0.50	0.38	0.75	0.42	0.45

TF-IDF + SVM

When trained on Davidson, it achieves reasonable performance on Founta (macro F1 = 0.49, weighted F1 = 0.70) but significantly drops on HateXplain (macro F1 = 0.40, weighted F1 = 0.40). A similar trend appears when trained on Founta, performing moderately well on Davidson (macro F1 = 0.54, weighted F1 = 0.76) but weakly on HateXplain (macro F1 = 0.40, weighted F1 = 0.41). When trained on HateXplain, it performs the worst, particularly on Founta (macro F1 = 0.39, weighted F1 = 0.60), indicating that HateXplain's linguistic diversity does not translate well to other datasets. Overall, TF-IDF + SVM generalizes moderately well but loses effectiveness when tested on HateXplain.

Glove + LSTM

Glove + LSTM is the weakest performer across the board. Regardless of which dataset it is trained on, it consistently achieves poor performance on all target datasets. The macro F1 scores remain below 0.16 in every case, and weighted F1 scores never exceed 0.15.

DistilBERT

DistilBERT shows the strongest cross-dataset generalization. When trained on Davidson, it achieves an F1 score of 0.61 (macro) and 0.86 (weighted) on Founta, demonstrating strong transferability. It does show a performance drop when tested on HateXplain (macro F1 = 0.45, weighted F1 = 0.45), but it still outperforms TF-IDF + SVM in every case. When trained on Founta, it maintains a macro F1 of 0.54 and weighted F1 of 0.72 on Davidson, similar to the SVM model but with better contextual understanding. Even when trained on HateXplain, which is the most challenging dataset to generalize from, it outperforms TF-IDF + SVM on Founta (macro F1 = 0.39 vs. 0.39, weighted F1 = 0.62 vs. 0.60) and Davidson (macro F1 = 0.38 vs. 0.46, weighted F1 = 0.45 vs. 0.61), though its results are more unstable.

Figure 4: Confusion Matrix for DistilBERT trained on the Davidson dataset and tested on the Founta dataset. Results show that the model performs well on normal and offensive tweets, but consistently misclassifies hateful tweets as either offensive or normal.

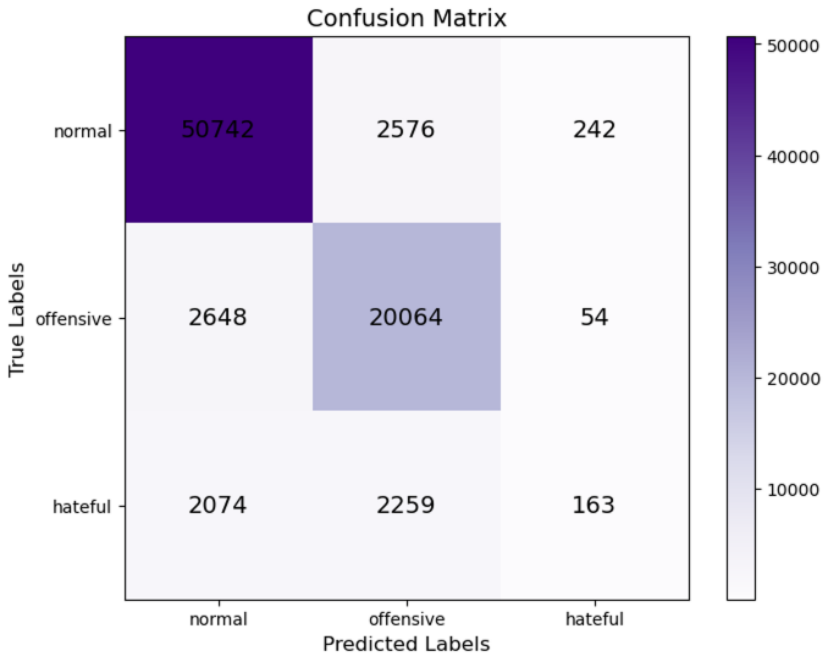


Figure 4 is a confusion matrix representing the results from testing the DistilBERT model (trained on the Davidson dataset) on the Founta validation set. It shows that the model does well in classifying normal and offensive tweets, but not on hateful tweets. Most hateful tweets are instead labeled as offensive or normal by the model.

Figure 5: Confusion Matrix for DistilBERT trained on the Davidson dataset and tested on the HateXplain dataset. Results show strong performance on normal and offensive tweets, but the model consistently misclassifies hateful tweets as either offensive or normal.

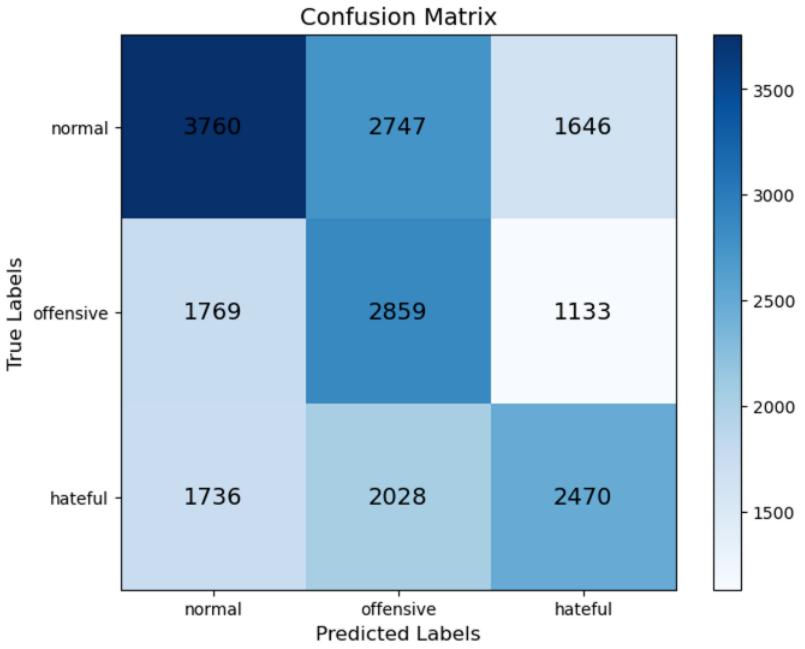


Figure 5 is a confusion matrix representing the results from testing the DistilBERT model (trained on the Davidson dataset) on the HateXplain dataset. It shows that the model classifies normal tweets best and does worse on offensive and hateful tweets. However, false positives are apparent in each class, with normal even having 2,747 misclassified as offensive. The hateful class is most likely to be misclassified as offensive, and the offensive class is most likely to be classified as normal.

The DistilBERT model consistently delivers the best results across datasets, making it the most reliable choice for handling unlabeled AAVE data. Its consistent F1 scores across different datasets suggest that it most effectively captures contextual relationships. TF-IDF + SVM is a secondary option, as it shows moderate transferability, but its performance degrades significantly on HateXplain. Glove + LSTM is unsuitable for generalization and should not be used. DistilBERT trained on Davidson performs best overall, achieving the highest average macro F1 score of 0.53 across both Founta and HateXplain. It transfers best to Founta (0.61 macro

F1), which suggests that Davidson's data provides good generalizable features for this dataset. Although performance on HateXplain is lower (0.45 macro F1), it is still better than models trained on Founta or HateXplain. Training on HateXplain results in the worst generalization, likely because its data is more different than between Founta and Davidson.

5.4 Aggregate Dataset Performance

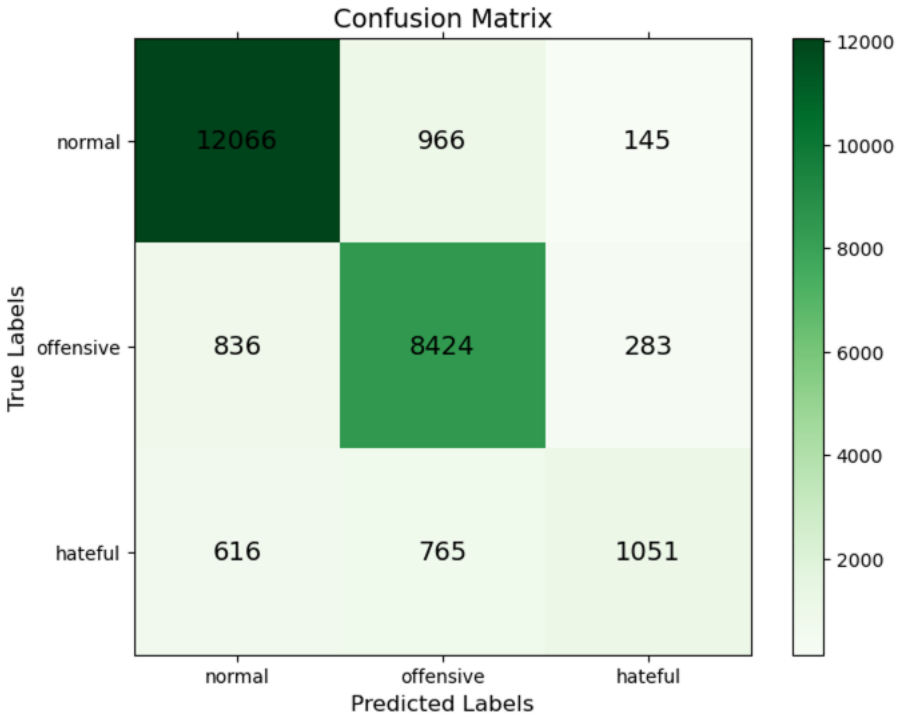
Each dataset demonstrates strengths in different aspects of hate speech detection, raising the question of whether combining them could lead to a more robust and generalizable model. By training on a merged dataset and evaluating performance on an aggregate of the validation sets, we aim to assess whether this approach enhances overall effectiveness. The results of this experiment are presented in Table 7.

Table 7: Aggregated Dataset Performance. Accuracy results from the initial Glove+LSTM model on three hate speech datasets. Although originally reported to achieve state-of-the-art performance, later findings suggest the model was overfit to its training data [42].

Model	Macro			Weighted Average			Macro Average AUC	Overall Accuracy
	Precision	Recall	F_1	Precision	Recall	F_1		
TF-IDF + SVM	0.76	0.66	0.68	0.79	0.80	0.78	0.87	0.80
Glove + LSTM	0.03	0.33	0.06	0.01	0.10	0.02	0.50	0.10
DistilBERT	0.81	0.74	0.77	0.85	0.86	0.85	0.83	0.86

TF-IDF + SVM demonstrated solid overall performance with a balanced weighted average F1 score of 0.78 and an AUC of 0.87, suggesting that it was effective at identifying the relevant features for classification. However, its macro-average scores (precision of 0.76, recall of 0.66, and F1 of 0.68) indicate a slightly less consistent performance across different classes, with potential struggles in correctly identifying certain labels. In contrast, GloVe + LSTM performed poorly across all metrics, with an F1 score of only 0.06 for the macro-average and a dismal AUC of 0.50, suggesting that this model fails to capture meaningful patterns in the data.

Figure 6: Confusion Matrix for DistilBERT Trained on Combined Datasets. Results show strong overall performance, particularly in classifying normal and offensive tweets. However, despite high macro and weighted average scores, the model continues to struggle with accurately identifying hateful tweets, which are frequently misclassified as either normal or offensive.



DistilBERT, however, outperformed both of the other models by a significant margin. It achieved strong macro-average results (precision of 0.81, recall of 0.74, and F1 of 0.77), and its weighted average scores (precision of 0.85, recall of 0.86, and F1 of 0.85) indicate exceptional performance in classifying the dataset with high accuracy across all classes. Additionally, its AUC of 0.83 suggests it is highly effective at distinguishing between different classes. Overall, DistilBERT is the best-performing model, when trained on the aggregated datasets. Figure 6 is the confusion matrix of this model's results. It shows that the model does best in classifying normal tweets, with 12,066 true positives, indicating that it correctly identifies most normal instances. For the offensive class, the model correctly predicts 8,424 instances, but it also mislabels normal and hateful instances as offensive, with 836 false positives and 283 false negatives. Although it correctly predicts 1,051 hateful instances, there are 616 instances where hateful tweets are classified as normal and 765 instances where hateful instances are mistaken

for offensive. This higher number of misclassifications in the hateful class suggests that the model struggles to identify hateful content with the same accuracy as the other classes.

Combining all of the datasets may not necessarily lead to a more powerful model due to several factors that can negatively impact performance. First, the datasets likely have different characteristics, including varying levels of noise, and domain-specific language use. When these datasets are merged, a model trained on the combined data might struggle to generalize well across all classes, as it may be forced to learn patterns that do not necessarily align across the datasets. This could result in a model that overfits certain data types while underperforming on others. Variations in annotation techniques across datasets may contribute to these inconsistencies, posing challenges for training a unified model. Since each dataset was created using distinct methodologies, the model may struggle to reconcile these variations, potentially hindering its ability to learn consistent patterns across datasets.

5.5 AAVE Study

To examine how AAVE is misclassified by current hate speech detection models, we apply our best-performing and most generalizable model – trained using the original dataset's class distribution, without balancing – to the AAVE dataset created by Blodgett et al [1]. This is the DistilBERT model trained on the Davidson dataset.

5.5.1 Model Performance

The Blodgett dataset consists of 1.1 million tweets, which we classified using our DistilBERT model trained on the Davidson dataset with original class distributions. We then manually annotate 500 instances from each predicted class to examine linguistic patterns and trends in misclassified data points. Our labeling approach defined offensive language as tweets containing words or expressions that, while disrespectful or inappropriate, do not necessarily

intend to harm or target a specific group. Such language may cause discomfort or annoyance but is not typically intended to incite harm or violence. We classified hateful tweets as more severe, typically containing expressions that deliberately target individuals or groups with the intent to cause harm, incite violence, or convey animosity based on inherent characteristics such as race, ethnicity, gender, religion, or sexual orientation. After creating these ground truth labels, we were able to generate a report on the model's performance – shown in Table 8.

Table 8: AAVE Results from classification by DistilBERT run on Davidson Dataset. Results indicate limited generalizability, with an overall accuracy of 64%.

Class	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Normal	1.00	0.61	0.76	0.99
Offensive	0.70	0.61	0.65	0.70
Hateful	0.22	0.98	0.36	0.22
Macro Avg	0.64	0.73	0.59	
Weighted Avg	0.82	0.64	0.69	
Overall Accuracy: 0.64				

From our sample of 1500 data points, the normal class performs very well with a precision of 1.00, indicating that every instance predicted as normal is indeed normal. This suggests that the model is highly accurate in predicting instances of the normal class. However, the recall for normal is 0.61, meaning that the model identifies only 61% of all true normal instances. This indicates that the model misses a significant portion of normal instances, especially those that may have been misclassified as other categories. The offensive class presents a different challenge for the model. While the precision for offensive is 0.70, meaning that 70% of instances predicted as offensive are truly offensive, this is lower than that of the normal class. The recall for offensive is 0.61, indicating that the model identifies 61% of true offensive instances, suggesting that the model is relatively effective in detecting Offensive instances but still misses about 39%. Of the three, performance for the hateful class is the lowest.

The precision for hateful is very low at 0.22, meaning that only 22% of instances predicted as hateful are actually hateful. This low precision indicates that the model is often misclassifying other classes as hateful. However, the recall for hateful is exceptionally high at 0.98, meaning that the model correctly identifies 98% of all true hateful instances. This suggests that the model is highly sensitive to the presence of hateful instances, but it struggles with accurately predicting them without incorrectly labeling other instances as hateful.

It is important to note that these results are derived from a subset of the entire dataset, which may not fully capture the diversity of the entire dataset. As such, while these results provide valuable insights into the model's performance, they may not be entirely representative of how the model would perform on the full dataset. Variations in the distribution of classes, the characteristics of the data, or other external factors could lead to different results if the model were tested on a broader, more diverse set of data. Therefore, while the analysis offers useful information, further testing on a more comprehensive dataset would be necessary to ensure that these findings are fully generalizable.

Figure 7: Confusion Matrix for AAVE Results from fine-tuned DistilBERT trained on Davidson Dataset. Results show that while the model performs moderately well on normal and offensive tweets, it struggles to accurately distinguish hateful content.

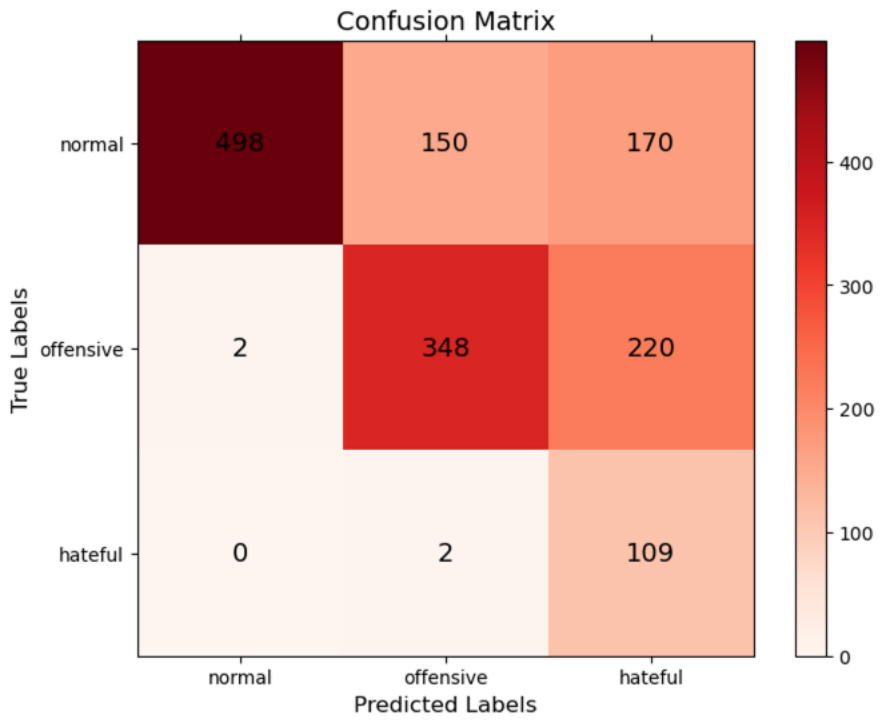


Figure 7's confusion matrix gives us insight into where the model's misclassifications are occurring. The normal class has 498 true positives, indicating that the model successfully classifies normal instances. However, it also misclassifies 150 normal instances as offensive and 170 as hateful, suggesting that the model struggles with distinguishing normal from the other two categories. For the offensive class, the model correctly identifies 348 instances, but it misclassifies 2 normal instances as offensive and 220 offensive instances as hateful. These results indicate some difficulty in separating it from the hateful class, leading to confusion between these two categories. The hateful class, while showing a relatively high number of true positives (109), suffers from a significant misclassification issue, with 220 offensive and 170 normal instances incorrectly predicted as hateful. The small number of true positives and relatively low misclassification from other classes suggest that the model is struggling to accurately identify hateful content.

5.5.2 Gender and Linguistic Trends

This section focuses on the gendered and linguistic trends observed in the misclassification of AAVE in hate speech detection. Given the challenges models face in accurately interpreting the diverse linguistic features of nonstandard languages, this analysis seeks to identify specific aspects of AAVE that may contribute to its disproportionate misclassification. Additionally, in line with the findings of Kim et al[2], which highlight the tendency for African American male speech to be disproportionately flagged as hateful, this analysis will investigate whether this pattern is evident in our own dataset. This analysis will contribute to a deeper understanding of the complexities in hate speech detection on nonstandard language.

Table 9: AAVE misclassifications by gender, according to classifications made using Kim's [2] gender classifier.

Percentage of false positive hateful predictions by gender	
Female	0.35
Both	0.27
Neither	0.23
Male	0.15

To start our analysis, we examine how the misclassifications are reflected in the gender associations of the tweets. Table 9 shows these findings. In our sample of 1,500 tweets, those classified as female were most frequently misclassified as hateful. However, since these results are based on a subset of the dataset, we cannot definitively conclude that they are fully representative of the entire dataset.

Figure 8: Word Cloud visualization of tweets incorrectly labeled as hateful. Word size corresponds to the frequency of terms in misclassified examples.

Since this dataset consists entirely of AAVE data, we begin our linguistic analysis by examining word usage, as it is the most relevant factor in this context, given that all tweets share a similar linguistic framework. Among the 20 most common words in tweets misclassified as hateful, the top two were variations of the n-word: "niggas" (misclassified 87 times) and "nigga" (misclassified 67 times). Overall, the n-word appeared in 38% of false predictions in our sample. Figure 8 presents a WordCloud highlighting the words most frequently misclassified as hateful, with "nigga" appearing most prominently, indicating its strong association with misclassification. It is important to note that some tweets misclassified as hateful were actually labeled as offensive, which explains the presence of certain offensive words in the WordCloud.

Table 10: N-word Usage in Tweets – illustrating different contexts and highlighting how the word can be used in both normal and hateful contexts.

Tweet	Predicted Label	Actual Label
dont play wit my feelings your liable to get your ass whooped or shot if u a nigga	Offensive	Hateful
i know who that is nigger boy	Hateful	Hateful
dese young niggas out here hungry	Hateful	Normal

To highlight the varying ways in which the n-word is used and how these differences impact classification, we present a small case study examining three distinct examples from the dataset. Table 10 showcases these instances, illustrating how context and tone influence whether a tweet is truly hateful, offensive, or simply a neutral usage of the word. By analyzing these cases, we aim to better understand the linguistic patterns that may contribute to misclassification and the challenges that arise in distinguishing between harmful and non-harmful language.

Tweet 1: "dont play wit my feelings your liable to get your ass whooped or shot if u a nigga"

This tweet was predicted as offensive but actually labeled as hateful. The language used here includes a direct threat ("liable to get your ass whooped or shot"), paired with the term "nigga," which is often flagged as hateful due to its association with racial hostility. However, the model misclassifies it as offensive, possibly because it overlooks the threatening nature of the message or struggles to interpret the intensity of aggression when combined with AAVE linguistic features.

Tweet 2: "i know who that is nigger boy"

This tweet was correctly predicted as hateful. The use of the "hard R" variation of the n-word is widely recognized as a slur, historically and presently used by outsiders to demean and oppress the Black community. "Nigger boy" is an explicitly racial slur, making it easy for the model to classify as hateful. This kind of direct derogatory language often falls under well-established patterns of hate speech, making it less susceptible to misclassification.

Tweet 3: "dese young niggas out here hungry"

The third tweet was predicted as hateful but is actually normal. The term "niggas" appears in this tweet, which might trigger the model's hate speech classification. However, "niggas" can be used as a term of camaraderie or to refer to individuals in a non-hostile way. The tweet's overall tone does not contain hate but is more about describing a situation involving "young niggas" who are "hungry." The misclassification highlights how the model may struggle to understand contextual usage, especially with a reclaimed slur which can have different meanings based on social and cultural factors.

The n-word has undergone a process of reclamation within the Black community, transforming from a term historically used to oppress into one that can express solidarity, familiarity, and cultural identity. In many contexts, it is used casually among peers, as a term of endearment, or as a marker of in-group belonging. However, despite this reclamation, the word is not universally benign – its meaning is highly dependent on context, tone, and the speaker's identity. While it can serve as an expression of camaraderie, it can also still be weaponized in hateful or derogatory ways, particularly when used by those outside the community. It is this distinction – between reclaimed, non-hateful uses and those that perpetuate harm – that hate speech detection models must navigate carefully.

Future studies should focus on refining hate speech detection models to better recognize these nuanced distinctions, ensuring that tweets containing the n-word are not automatically flagged as hateful without considering context. By incorporating linguistic and sociocultural understanding into model training, researchers can work toward reducing the unnecessary censorship of Black voices while still identifying genuinely harmful speech.

Chapter 6

Conclusion and Future Work

This thesis has explored the complexities of hate speech detection, particularly focusing on how these models struggle to accurately classify AAVE text. As online platforms increasingly serve as primary spaces for social interaction, the need for robust and inclusive automated systems for detecting harmful language is more pressing than ever. Through this work, we have examined the limitations of existing methodologies, the challenges posed by diverse linguistic forms like AAVE, and the ethical problems inherent in these systems. In this concluding chapter, we reflect on the contributions of this research, discuss its limitations, and consider future directions.

6.1 Contributions

This research makes a key contribution by demonstrating a consistent pattern of misclassification in hate speech detection models, specifically in relation to the use of the n-word within African American Vernacular English (AAVE). This pattern reveals how reclaimed, in-group terms used by Black speakers are disproportionately flagged as hateful, pointing to a trend of algorithmic censorship of Black voices. By surfacing this trend, the thesis sheds light on the

linguistic biases embedded in current detection systems and contributes to ongoing efforts to make hate speech detection more context aware.

6.2 Limitations

While this research offers valuable insights into the misclassification of AAVE in hate speech detection, it is not without limitations. First, the hand-labeling of the AAVE dataset as hateful/normal/offensive introduces subjectivity and potential biases into the labeling process, as these classifications were made based on personal judgment rather than automated or widely accepted standards. Additionally, the labeled subset of the Blodgett AAVE dataset represents only a small portion of over one million tweets and may not accurately reflect the broader dataset. Finally, the study's reliance on a binary gender classification framework limits the scope of analysis, excluding the full spectrum of gender identities and expressions that exist in real-world discourse. These limitations should be considered when interpreting the findings, as they may influence the generalizability and robustness of the results.

6.3 Future Directions

Looking ahead, there are several promising avenues for future research in the development of more inclusive models for hate speech detection. Future work could focus on augmenting existing datasets with a more diverse range of linguistic samples, particularly those reflecting how marginalized communities communicate online. Enhancing linguistic diversity in training data can help models perform more equitably and more accurately reflect the complexities of real-world language. This process should be paired with annotator priming strategies, such as those proposed by Sap et al. [21]. This would involve informing annotators about the cultural and contextual significance of reclaimed or in-group terms – such as the n-

word in Black communities or the f-word in queer spaces. Together, these efforts can lead to more representative datasets and, ultimately, fairer hate speech detection systems.

The algorithmic biases explored in this work are a small part of a larger, widespread pattern of algorithmic injustice present across multiple domains. Studies have shown that women's voices are often underrepresented and misclassified in speech recognition and automated systems. YouTube's auto-captioning system, for instance, performs better on male speech than female speech, largely due to differences in voice pitch [59]. Biased algorithms have also been known to create obstacles in access to housing and loan opportunities. All of these seemingly small and isolated issues compound into a broader system of algorithmic marginalization. This work aims to contribute to the ongoing discussion on how we can address and rectify these injustices.

As hate speech detection systems continue to evolve, their development must strike a balance between technical advancement and ethical responsibility, ensuring that these tools are not only effective but also just. Given the persistent social inequalities that shape outcomes across demographic groups, a commitment to non-distributive forms of justice is essential to achieving genuine fairness – one that accounts for context, identity, and historical marginalization. By integrating both cultural and linguistic sensitivity as well as reparative justice into their design, we can mitigate harm while fostering a more inclusive and equitable digital space. This research aims to support a more holistic approach to hate speech detection – one that acknowledges cultural, sociolinguistic, and ethical contexts to foster fairer and more inclusive systems.

REFERENCES

[1] Blodgett, S. L., Green, L. and O'Connor, B. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[2] Kim, J. Y., Ortiz, C., Nam, S., Santiago, S. and Datta, V. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921* (2020).

[3] Green, L. J. *African American English: a linguistic introduction*. Cambridge University Press, 2002.

[4] Dalby, D. The African element in American English. *Rappin'and stylin'out: Communication in urban Black America* (1972), 170-186.

[5] Dunn, E. F. *Black-Southern white dialect controversy*, 1976.

[6] DeBose, C., Faraclas, N. and Africanisms, C. G. An Africanist approach to the linguistic study of Black English: Getting to the roots of the tense-aspect-modality and copula systems in Afro-American English. *1993* (1993), 364-387.

[7] Rickford, J. R. *The creole origins of African-American Vernacular English: Evidence from copula absence*. Routledge, 2013.

[8] Edwards, W. F. and Winford, D. Verb phrase patterns in Black English and Creole. *(No Title)* (1991).

[9] Wolfram, W. *Urban African American Vernacular English*, 2020.

[10] Ilbury, C. "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of sociolinguistics*, 24, 2 (2020), 245-264.

[11] Rickford, J. R. and Rickford, R. J. *Spoken soul: The story of black English*. Turner Publishing Company, 2007.

[12] Ushiyama, N., Oganesyan, S., Boehm, A., Lee, R. and Vaughn, A. *Fun, Cool, Hip Title Here: AAVE Usage in Twitter Memes*. Retrieved from Languaged Life: https://languagedlife. humspace. ucla. edu …, 2021.

[13] Hill, J. H. *The everyday language of white racism*. John Wiley & Sons, 2024.

[14] Smith, H. L. Has nigga been reappropriated as a term of endearment? A qualitative and quantitative analysis. *American Speech: A Quarterly of Linguistic Usage*, 94, 4 (2019), 420-477.

[15] Association, A. L. *Hate Speech and Hate Crime*, 2017.

[16] Davidson, T., Warmsley, D., Macy, M. and Weber, I. *Automated hate speech detection and the problem of offensive language*, 2017.

[17] Borkan, D., Dixon, L., Sorensen, J., Thain, N. and Vasserman, L. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Proceedings of the Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA, 2019). Association for Computing Machinery, 2019.

[18] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M. and Kourtellis, N. *Large scale crowdsourcing and characterization of twitter abusive behavior*, 2018.

[19] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G. and Vakali, A. *Mean birds: Detecting aggression and bullying on twitter*, 2017.

[20] Xia, M., Field, A. and Tsvetkov, Y. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246* (2020).

[21] Sap, M., Card, D., Gabriel, S., Choi, Y. and Smith, N. A. *The risk of racial bias in hate speech detection*, 2019.

[22] Blodgett, S. L. and O'Connor, B. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).

[23] *Common Crawl*, 2025.

[24] Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □ü¶ú. In *Proceedings of the Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada, 2021). Association for Computing Machinery, 2021.

[25] Wagner, C., Graells-Garrido, E., Garcia, D. and Menczer, F. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ data science*, 5 (2016), 1-24.

[26] Massanari, A. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, 19, 3 (2017), 329-346.

[27] Waseem, Z. and Hovy, D. *Hateful symbols or hateful people? predictive features for hate speech detection on twitter*, 2016.

[28] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P. and Mukherjee, A. *Hatexplain: A benchmark dataset for explainable hate speech detection* 2021.

[29] Rajani, N. F., McCann, B., Xiong, C. and Socher, R. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361* (2019).

[30] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H. and Pierrehumbert, J. B. HateCheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606* (2020).

[31] Ousidhoum, N., Lin, Z., Zhang, H., Song, Y. and Yeung, D.-Y. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049* (2019).

[32] Tang, Z., Li, W., Li, Y., Zhao, W. and Li, S. Several alternative term weighting methods for text representation and classification. *Knowledge-Based Systems*, 207 (2020), 106399.

[33] Burnap, P. and Williams, M. L. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7, 2 (2015), 223-242.

[34] Schmidt, A. and Wiegand, M. *A survey on hate speech detection using natural language processing*, 2017.

[35] Greevy, E. and Smeaton, A. F. *Classifying racist texts using a support vector machine*, 2004.

[36] Chen, Y., Zhou, Y., Zhu, S. and Xu, H. *Detecting offensive language in social media to protect adolescent online safety*. Ieee, 2012.

[37] Badjatiya, P., Gupta, S., Gupta, M. and Varma, V. *Deep learning for hate speech detection in tweets*, 2017.

[38] Saroj, A., Mundotiya, R. K. and Pal, S. *IRLab@ IITBHU at HASOC 2019: Traditional Machine Learning for Hate Speech and Offensive Content Identification*, 2019.

[39] Gambäck, B. and Sikdar, U. K. *Using convolutional neural networks to classify hate-speech*, 2017.

[40] Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M. and Tesconi, M. *Hate me, hate me not: Hate speech detection on facebook*, 2017.

[41] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26 (2013).

[42] Arango, A., Pérez, J. and Poblete, B. *Hate speech detection is not as easy as you may think: A closer look at model validation*, 2019.

[43] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019.

[44] Clark, K., Luong, M.-T., Le, Q. V. and Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[45] Deery, O. and Bailey, K. The bias dilemma: the ethics of algorithmic bias in natural-language processing. *Feminist Philosophy Quarterly*, 8, 3/4 (2022), 1-28.

[46] Mikolov, T. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[47] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29 (2016).

[48] Kong, Y. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *Proceedings of the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea, 2022). Association for Computing Machinery, 2022.

[49] Dastin, J. *Insight - Amazon scraps secret AI recruiting tool that showed bias against women*, 2018.

[50] Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. New York university press, 2018.

[51] Crawford, K. *The Trouble with Bias - NIPS 2017 Keynote* YouTube, 2017.

[52] Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6 (2018), 587-604.

[53] Zafar, M. B., Valera, I., Gomez Rodriguez, M. and Gummadi, K. P. *Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment*, 2017.

[54] Buolamwini, J. and Gebru, T. *Gender shades: Intersectional accuracy disparities in commercial gender classification*. PMLR, 2018.

[55] *Developer Agreement and Policy*, 2024.

[56] Mathew, B., Dutt, R., Goyal, P. and Mukherjee, A. *Spread of hate speech in online social media*, 2019.

[57] Eight, F. *Twitter User Gender Classification*, 2017.

[58] Malik, J. S., Qiao, H., Pang, G. and Hengel, A. v. d. Deep learning for hate speech detection: a comparative study. *arXiv preprint arXiv:2202.09517* (2022).

[59] Tatman, R. *Gender and dialect bias in YouTube's automatic captions*, 2017.

APPENDIX A

A.1 Justifications Behind Annotation and Labeling Techniques

This section provides an overview of the annotation and labeling techniques employed in the datasets used in this study. The justifications for these techniques were drawn from the original papers and methodologies provided by the creators of each dataset. By reviewing these sources, this section outlines the rationale behind the chosen labeling conventions, offering insight into the specific decisions made during the annotation process. Understanding these foundations helps clarify how and why certain categories were defined, ensuring transparency and a clear understanding of the dataset preparation for the study.

A.1.1 Davidson Dataset

Davidson et al. used Hatebase.org to create a lexicon of hateful terms, which they applied to the Twitter API to collect tweets from 33,458 users, resulting in a corpus of 85.4 million tweets. From this, they randomly sampled 25,000 tweetscontaining lexicon terms and had them manually labeled by CrowdFlower workers.

Tweets were categorized as:

- **Hate speech**
- **Offensive** but not hate speech
- **Neither** offensive nor hate speech

Annotators were given detailed definitions and instructed to consider context, not just word presence. Each tweet was labeled by at least three annotators, with a 92% intercoder

agreement. Labels were assigned based on majority vote; tweets without a majority were excluded, resulting in 24,802 labeled tweets.

A.1.2 Founta Dataset

Label Definitions: Annotators were provided with definitions for each category, drawn from existing literature and dictionaries:

- **Offensive Language**: Profanity or vulgar language used to insult an individual or group.
- **Abusive Language**: Strongly impolite or hurtful language using profanity that may express intense emotion or debasement.
- **Hate Speech**: Language expressing hatred or aiming to humiliate based on attributes such as race, religion, ethnicity, gender, disability, or sexual orientation.
- **Aggressive Behavior**: Angry or violent online communication intended to harm or upset others.
- **Cyberbullying**: Repeated, hostile behavior via electronic means intended to intimidate, abuse, or embarrass others.
- **Spam**: Promotional or malicious content, including advertising, phishing, or repeated unwanted posts.
- **Normal**: Tweets not fitting any of the above categories.

Annotation Process:

The Founta annotation process was conducted in three phases:

1. **Exploratory Round:** Annotators labeled tweets as *normal*, *spam*, or *inappropriate*. Inappropriate tweets were further classified into one of five categories: *offensive*, *abusive*, *hateful*, *aggressive*, or *cyberbullying*.

2. **Refinement Round:** Based on low annotation agreement and significant category overlap, the labeling scheme was revised. *Cyberbullying* was removed due to its reliance on repeated behavior (which is hard to assess in isolated tweets), and *offensive*, *abusive*, and *aggressive* were collapsed into a single category. *Hateful* remained separate due to its distinct definition and relevance. This resulted in four final labels: Abusive, Hateful, Normal, and Spam.
3. **Validation and Final Annotation:** We validated the simplified scheme with further rounds on the smaller dataset. These showed improved annotator agreement and clearer label separation. The final 80K tweets were then annotated using this scheme, with each tweet receiving five independent labels.

Annotator Demographics: Annotations were crowdsourced globally. Annotators came from over 100 countries, with the majority from Venezuela, the U.S., Egypt, and India. Most had at least a secondary education, and income levels skewed low. Two-thirds of annotators identified as male.

A.1.3 HateXplain Dataset

Annotation Process: Annotation was conducted via Amazon Mechanical Turk (MTurk) with each post labeled across three dimensions:

1. **Type of speech**:
    - Hateful
    - Offensive
    - Normal
2. **Target community** (e.g., Race, Religion, Gender, Sexual Orientation).

3. **Rationale spans**: Annotators highlighted the words/phrases responsible for their classification.

Instructions:

Annotators were warned about exposure to offensive content, given detailed guidelines, definitions, and examples.

Target Communities:

Included communities: African, Islam, Jewish, LGBTQ, Women, Refugee, Arab, Caucasian, Hispanic, Asian. A group is considered a target if at least 2 of 3 annotators agree and the group appears in ≥100 posts.

Top Targets Identified:

- Hate speech: African, Islam, Jewish
- Offensive speech: Women, African, LGBTQ

Rationale Annotation:

Each hateful/offensive post received 2–3 rationale annotations.

- Avg. tokens per rationale: ~5.48
- Avg. tokens per post: 23.42
- Top hateful terms: *nigger*, *kike*, *moslems* (30.02% of hateful posts)
- Top offensive terms: *retarded*, *bitch*, *white* (47.36% of offensive posts)

A.2 Text Cleaning Procedure

To prepare the training datasets for modeling, I implemented a text preprocessing function in Python using several tools from the regular expression, string, and nltk (Natural Language Toolkit) libraries, as well as pandas for data handling. The cleaning function was

designed to reduce noise and standardize text input for downstream natural language processing tasks. The steps are as follows:

1. **Lowercasing**: All text was converted to lowercase using Python's built-in str.lower() method to ensure uniformity and reduce duplication due to case differences.
2. **URL and Handle Removal**: URLs (e.g., strings beginning with http, https, or www) were removed using regular expressions (re.sub). Twitter-specific elements such as mentions (@username) and hashtags (#topic) were also removed to eliminate social media artifacts that do not contribute to semantic meaning.
3. **Digit and Punctuation Removal**: All numeric characters were removed, including long strings of digits (e.g., phone numbers or IDs), along with punctuation using string.punctuation and Python's translate() function.
4. **Tokenization**: Text was split into individual words (tokens) using word_tokenize() from the nltk.tokenizemodule.
5. **Stop Word Removal**: Common English stop words (e.g., "the", "is", "and") were removed using the list provided by nltk.corpus.stopwords to focus on semantically meaningful words.
6. **Lemmatization**: Words were reduced to their base (or lemma) forms using WordNetLemmatizer from nltk.stem, helping to consolidate variations of the same word (e.g., "running" → "run").
7. **Whitespace Normalization**: Finally, extraneous spaces were removed using regular expressions to ensure clean, readable output.

Before running the function, I downloaded the required NLTK resources (punkt, stopwords, and wordnet) using nltk.download().

A.3 Dataset Label Comparisons

This section presents an exploration of the relationship between dataset labels using Euclidean distance and cosine similarity. The goal was to assess how the tweets in the datasets are separated in feature space according to their assigned labels. By calculating these distances, the aim was to identify any patterns or groupings that could inform understanding of the label distributions. However, the results did not yield significant insights due to the nature of the data, and as such, they are included here in the appendix for reference.

A.3.1 Intra-Class Label Similarity

After preprocessing the datasets, we analyzed class statistics and measured cosine similarity to evaluate the separation of tweets within each class. Cosine similarity is a measure of similarity between two vectors, with values ranging from 0 (no similarity) to 1 (identical). In this case, the cosine similarity within each class label indicates how similar the instances of that class are to each other in terms of their feature representations. Lower values suggest more variability within the class, while higher values imply more consistency. Additionally, we calculated the Euclidean distance between tweet vectors within each class, which measures the straight-line distance between instances in a multidimensional space. Smaller Euclidean distances indicate that the tweets within a class are closer to each other in feature space, suggesting a higher degree of similarity. Larger distances indicate more spread out instances within a class. Table A shows these values for each class in each dataset.

Table A: Intra-Class cosine similarity and Euclidean distance. These values indicate how semantically similar or dispersed tweets are within classes.

Label	Cosine similarity within class	Euclidean distance within class
Davidson Dataset		
normal	0.52	3.09
offensive	0.56	2.85
hateful	0.50	3.10
Founta Dataset		
normal	0.63	2.82
offensive	0.58	2.88
hateful	0.61	2.74
HateXplain Dataset		
normal	0.65	2.71
offensive	0.64	2.70
hateful	0.64	2.59

Davidson Dataset

The cosine similarity within the Davidson dataset remains relatively low across all classes, indicating that the feature representations of tweets within each category (normal, offensive, hateful) are still fairly diverse. The *hateful* class now shows the lowest similarity (0.50), suggesting that hateful tweets in this dataset may vary widely in how they are expressed linguistically. The *normal* class follows closely at 0.52, continuing the trend of higher intra-class variability. Interestingly, the *offensive* class now has the highest similarity (0.56), suggesting that offensive tweets may exhibit more consistent patterns or phrasing. The Euclidean distances are fairly close across all classes, with *offensive* tweets being the most tightly clustered (2.85) and *hateful* tweets the most spread out (3.10), reinforcing the idea that hateful language in this dataset is especially diverse in its expression.

Founta Dataset

In the Founta dataset, the *normal* class now exhibits the highest cosine similarity (0.63), diverging from the patterns seen in Davidson. This suggests that tweets labeled as normal in this

dataset may follow more consistent linguistic structures. The *hateful* class follows closely with a similarity of 0.61, while the *offensive* class is somewhat lower at 0.58. This reversal of trends may reflect differences in labeling guidelines or tweet content across datasets. In terms of Euclidean distance, *hateful* tweets are the most compact (2.74), followed by *normal* (2.82), and *offensive* tweets are the most dispersed (2.88). The closeness of these values still suggests that all classes are moderately well-clustered but that hate speech in this dataset may be more narrowly defined or exhibit more consistent features.

HateXplain Dataset

All classes in the HateXplain dataset show high cosine similarity values, with *normal* at 0.65, and both *offensive* and *hateful* classes very close behind at 0.64. This overall consistency suggests that in this dataset, tweet representations within each class are highly aligned. The small variation between classes implies that the dataset might be more curated, or the language used in each class might conform to particular linguistic norms. Euclidean distances echo this tight clustering: *hateful* tweets have the lowest spread (2.59), while *normal* tweets are slightly more dispersed (2.71). These patterns could reflect clearer distinctions between categories and possibly more rigorous annotation, resulting in lower intra-class variability and more homogenous representations.

Across all datasets, the normal class consistently exhibits the lowest intra-class cosine similarity, particularly in the Davidson (0.0168) and Founta (0.0167) datasets. This suggests that tweets labeled as "normal" may be more diverse in their content and wording compared to offensive or hateful tweets, which often share common linguistic patterns. In contrast, offensive and hateful labels tend to have higher intra-class similarity, particularly in the Founta (0.0280 for offensive, 0.0229 for hateful) and HateXplain (0.0280 for offensive, 0.0282 for hateful) datasets.

The increased similarity indicates that offensive and hateful speech could follow more distinct patterns, possibly due to recurring slurs, insults, or syntactic structures that make them more uniform. HateXplain exhibits the highest intra-class similarity overall, suggesting it may have stricter labeling criteria or more homogenous representations of each category.

We might expect cosine similarity within labeled tweet classes for hate speech detection to be low because hate speech and abusive language do not follow a consistent linguistic structure. Hate speech, abusive language, and even neutral speech can take many different lexical, syntactic, and stylistic forms, depending on factors such as dialect, sarcasm, topic, and intended audience. Unlike sentiment analysis, where positive or negative language often relies on a shared set of words and expressions (e.g., "happy," "excited" for positive sentiment), hate speech can be conveyed through a wide range of words, tones, and rhetorical strategies. Some instances may include direct slurs, while others rely on euphemisms, sarcasm, or context-dependent implications. This variability means that even tweets labeled as the same category may have low cosine similarity, as they do not share a common linguistic pattern.

A.3.2 Inter-Class Similarity

To measure similarity between classes within each dataset, we calculated cosine similarity and Euclidean distance. Euclidean distance measures how far apart two labels are in the feature space. Each tweet is represented as a point in this space based on its linguistic features, and the distance between two points reflects their dissimilarity. When comparing labels like normal and offensive, the Euclidean distance tells us how distinct the tweets in these categories are from each other. A smaller Euclidean distance means the tweets classified under those labels are more similar, with overlapping features, making it harder for a model to differentiate between them. Essentially, Euclidean distance helps reveal how much the language

used in different categories diverges. Smaller distances highlight subtle differences that models might struggle with, while larger distances suggest more defined distinctions between labels.

Table B: Inter-Class cosine similarity and Euclidean distance. Measures of How Similar or Distinct Tweets Are Between Different Labels in a Dataset

Labels	Cosine Similarity	Euclidean Distance
Label comparisons across Davidson		
normal vs offensive	0.9724	0.8033
offensive vs hateful	0.9947	0.3416
hateful vs normal	0.9736	0.8200
Label comparisons across Founta		
normal vs offensive	0.9647	0.9255
offensive vs hateful	0.9896	0.5018
hateful vs normal	0.9872	0.5572
Label comparisons across HateXplain		
normal vs offensive	0.9993	0.1450
offensive vs hateful	0.9988	0.1799
hateful vs normal	0.9977	0.2648

Davidson Dataset

In the Davidson dataset, the cosine similarities between each pair of labels are quite high, ranging from 0.9724 (normal vs offensive) to 0.9947 (offensive vs hateful). The Euclidean distances, however, show more distinction. The lower distance between offensive and hateful (0.3416) suggests that these two labels share more overlapping features, making them harder to distinguish. In contrast, the distances between normal and offensive (0.8033) and normal and hateful (0.8200) indicate greater separation, implying that there are clearer boundaries between these labels.

Founta Dataset

The Founta dataset shows a similar trend, though the distinctions are slightly less pronounced. The normal vs offensive similarity is slightly lower (0.9647), with a higher

Euclidean distance (0.9255), suggesting that Founta differentiates normal and offensive language more distinctly than Davidson. However, the offensive vs hateful labels remain closely related, with a high cosine similarity (0.9896) and a relatively low Euclidean distance (0.5018), reinforcing the notion that these categories overlap significantly.

HateXplain Dataset

In the HateXplain dataset, the labels are remarkably similar across the board, with cosine similarities exceeding 0.997 for all comparisons. The Euclidean distances are correspondingly low, particularly between normal and offensive (0.1450), indicating a substantial overlap in how these labels are represented. This suggests that HateXplain may not be drawing strong distinctions between these categories, potentially making it harder for models to accurately separate them.

Overall, the closer relationship between offensive and hateful language across datasets highlights the challenge in distinguishing these forms of harmful speech, as the boundaries between offensive and hateful content are often blurred. Meanwhile, the varying levels of similarity between normal and offensive language suggest that different datasets may have different thresholds for what constitutes "offensive" speech, reflecting subjective biases in annotation.

A.3.3 Cross-Dataset Comparison

By measuring Euclidean distance, we can quantify how different the representations of each class – normal, offensive, and hateful – are between datasets. A lower Euclidean distance means that the features representing a particular class are more similar across datasets, indicating greater alignment in how the datasets define and capture that type of language. Conversely, a higher Euclidean distance suggests that the datasets represent the same class in noticeably

different ways, reflecting potential differences in data collection, annotation, or interpretation. By examining these distances, we gain insight into how consistently each class is represented across datasets, which is crucial when evaluating model performance and generalizability. Table C shows these results.

Table C:Cross-Dataset cosine similarity and Euclidean distance. Measures of how similarly tweets are represented across different datasets within the same label.

Label	Cosine Similarity	Euclidean Distance
Comparison across Davidson and Founta		
normal	0.9931	0.4352
offensive	0.9931	0.3885
hateful	0.9861	0.6605
Comparison across Davidson and HateXplain		
normal	0.9897	0.5707
offensive	0.9738	0.8491
hateful	0.9809	0.7831
Comparison across Founta and HateXplain		
normal	0.9931	0.4284
offensive	0.9778	0.7630
hateful	0.9961	0.3132

Normal Class

The normal class shows consistently high cosine similarity across all dataset comparisons, with values around 0.993, indicating that the normal tweets in each dataset have highly similar representations. The Euclidean distances for the normal class are also relatively low, especially between Founta and HateXplain (0.4284), suggesting that the feature spaces for normal tweets are closely aligned. This consistency implies that the definition and characteristics of normal speech are relatively stable across datasets.

Offensive Class

The offensive class shows slightly lower cosine similarity, particularly between Davidson and HateXplain (0.9738), indicating more variation in how offensive language is represented

across datasets. The Euclidean distances for offensive tweets are higher compared to the normal class, especially between Davidson and HateXplain (0.8491). This could reflect differences in the types of offensive language captured by each dataset or slight variations in annotation criteria.

Hateful Class

The hateful class displays the most variation across datasets. The lowest cosine similarity is observed between Davidson and Founta (0.9861), suggesting that the datasets diverge the most when identifying hateful content – though this difference is marginal. The Euclidean distances for the hateful class vary more, with the largest distance found between Davidson and HateXplain (0.7831), reinforcing the idea that these datasets may not fully align on what constitutes hateful speech. Interestingly, the comparison between Founta and HateXplain yields a very high cosine similarity (0.9961) and the lowest Euclidean distance (0.3132), implying that these two datasets may have more agreement on the features representing hateful content.

Overall, these results suggest that the normal class is the most consistently represented across datasets, with lower Euclidean distances indicating a more stable definition of what constitutes "normal" speech. However, the offensive and hateful classes show greater variation, particularly when comparing Davidson and HateXplain. This variability points to potential discrepancies in how offensive and hateful language are perceived and labeled across datasets, which could impact model performance when applied to new data.

APPENDIX B

This appendix contains additional figures and tables generated during the experimentation process that were not central to the final analysis. While these results did not directly contribute to the conclusions drawn in the main body of the thesis, they are included here for transparency and reference. Readers interested in deeper exploration of the model behavior and dataset characteristics may find these materials informative.

B.1 Additional Confusion Matrices From Imbalanced Dataset Model Results

Figure A: Confusion Matrix for TF-IDF+SVM trained on imbalanced Davidson Dataset.

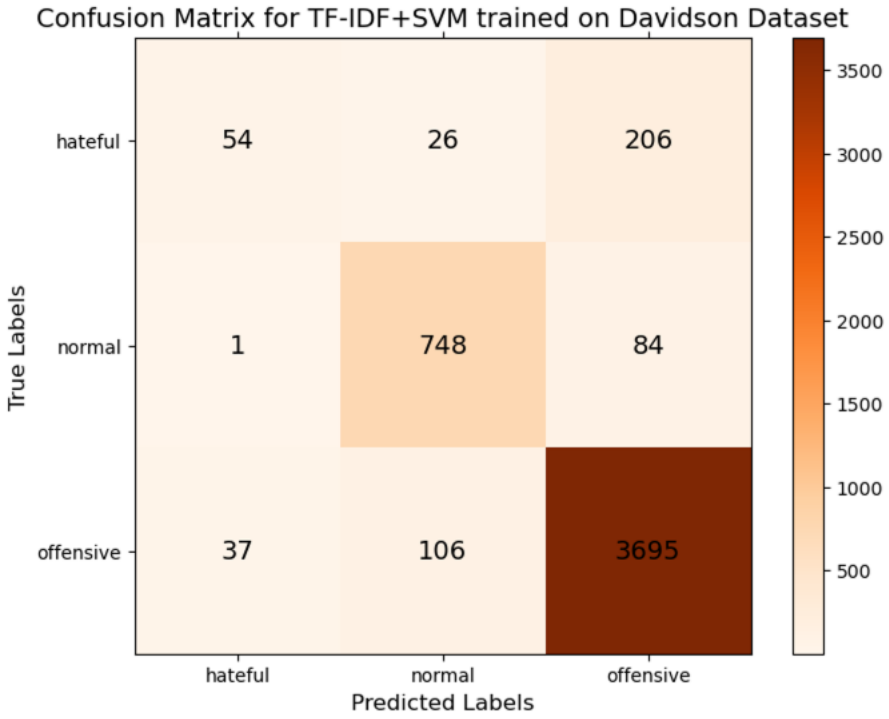


Figure A shows a confusion matrix for the TF-IDF + SVM model trained on the Davidson dataset which reveals that while the model effectively classifies offensive speech, it struggles with distinguishing between offensive and normal language. Specifically, the model correctly identifies 3,695 offensive instances but misclassifies 106 normal and 37 hateful tweets

as offensive. For normal speech, 748 instances are correctly classified, with only one being mistaken for hateful and 84 misclassified as offensive. However, the classification of hateful speech is notably poor, with only 54 instances correctly identified, while 206 are mislabeled as offensive and 26 as normal. This indicates a strong bias toward the offensive class, leading to a high false positive rate for offensive speech as hateful.

Figure B: ROC Curve for TF-IDF+SVM trained on imbalanced Davidson Dataset.

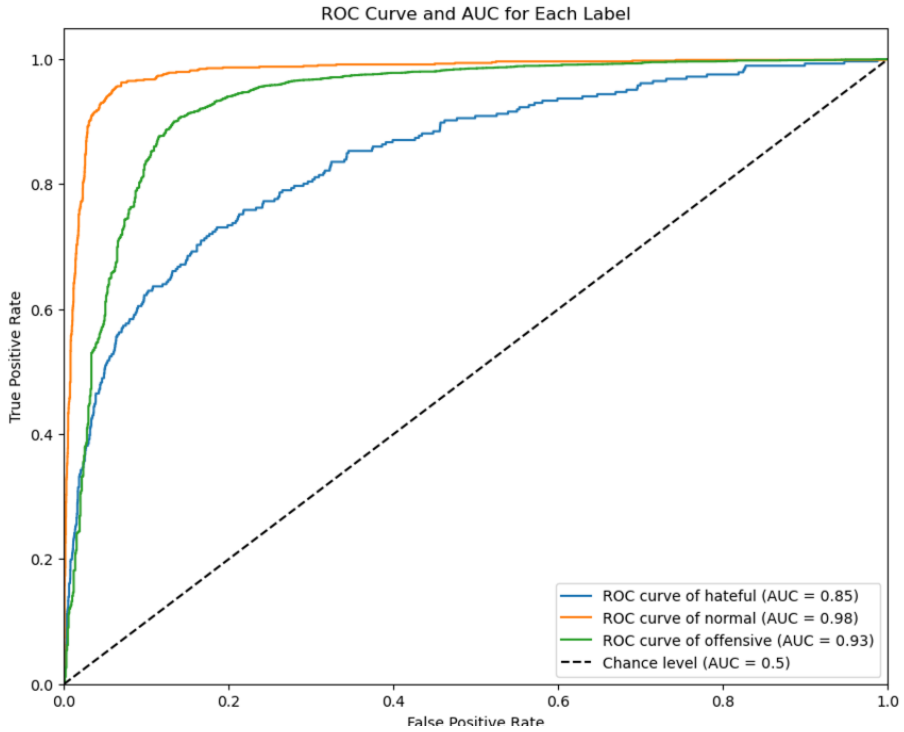


Figure B shows an ROC curve and AUC scores provide insight into the model's classification confidence and overall discriminatory power. The AUC for the normal class is the highest at 0.98, suggesting that when the model correctly classifies normal tweets, it does so with high confidence. Offensive speech follows closely with an AUC of 0.93, indicating strong classification ability for this category as well. However, the hateful class has the lowest AUC at

0.85, reflecting a weaker, albeit still strong, ability to differentiate hateful content from other categories.

Figure C: Confusion Matrix for TF-IDF+SVM trained on imbalanced Founta Dataset

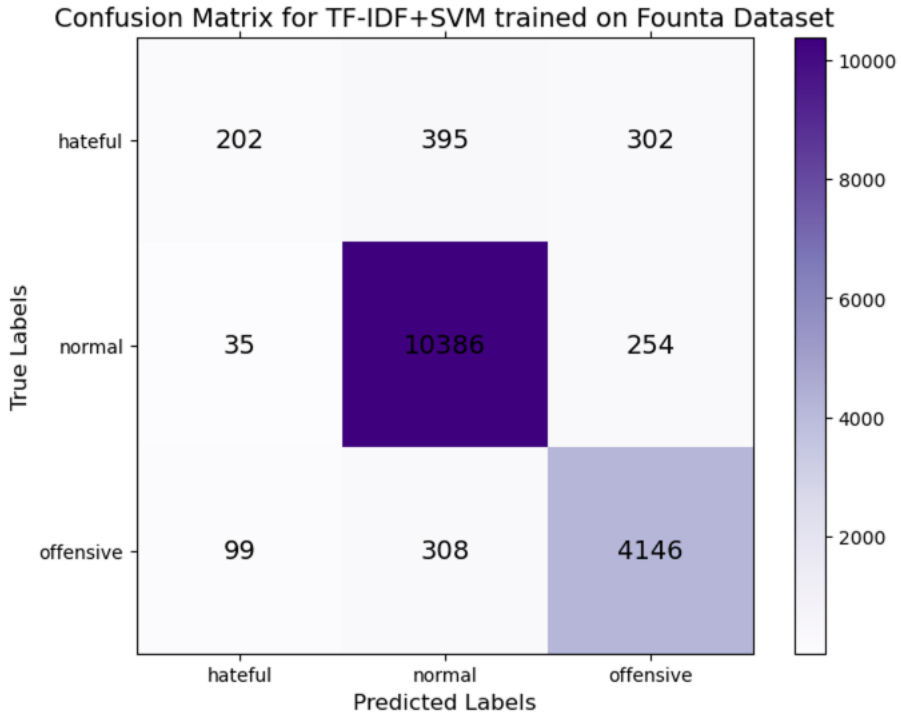


Figure C's confusion matrix represents results from the TF-IDF+SVM model trained on the Founta dataset. The model performs well in classifying normal tweets, correctly identifying 10,386 instances, with only 35 misclassified as hateful and 254 misclassified as offensive. This suggests that the model has a strong bias towards correctly identifying normal tweets, but may still struggle with some edge cases. However, when it comes to hateful tweets, the model performs much worse. Out of all hateful tweets, only 202 were correctly classified, while 395 were misclassified as normal and 302 as offensive. This indicates that the model has difficulty distinguishing hateful speech from both normal and offensive language, potentially due

to overlapping linguistic patterns. Similarly, offensive tweets are somewhat well classified, with 4,146 correct predictions, but 99 instances were misclassified as hateful and 308 as normal, suggesting some level of confusion between offensive and normal speech.

Figure D: ROC Curve TF-IDF+SVM trained on imbalanced Founta Dataset

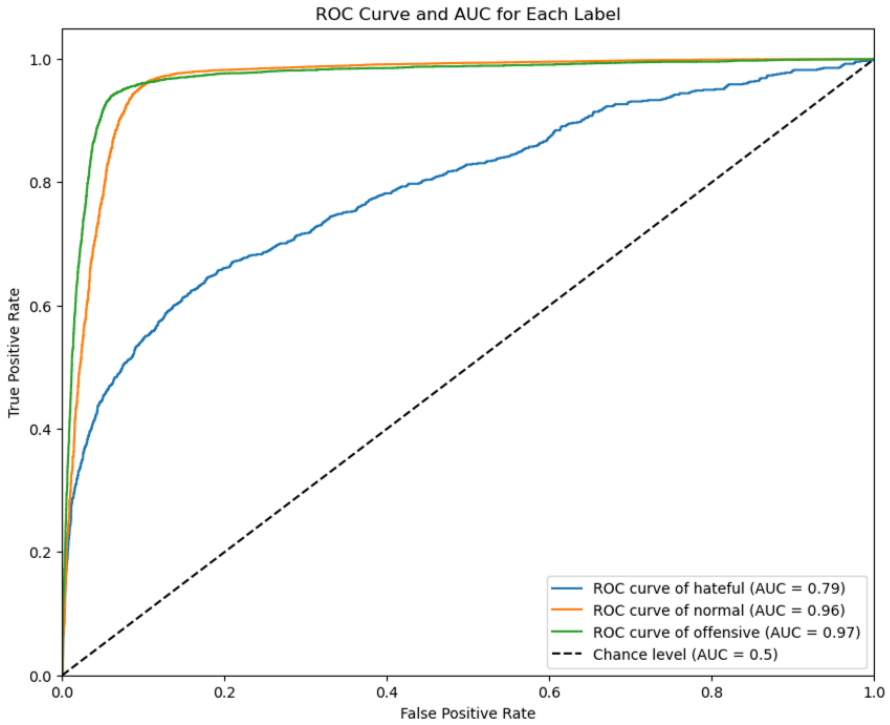


Figure D displays the ROC curve for the TF-IDF+SVM model trained on the Founta dataset. It further supports the conclusion that the model struggles the most with the hateful class. The AUC for hateful tweets is 0.79, significantly lower than the scores for normal (0.96) and offensive (0.97) tweets. This indicates that the model has difficulty distinguishing hateful tweets from the other two classes, leading to more false positives and false negatives. In contrast, the model performs well in classifying normal and offensive tweets, as evidenced by their near-perfect AUC scores. The high performance for these two categories suggests that the feature

extraction method (TF-IDF) and the SVM model are effective in detecting general patterns of normal and offensive speech but fall short when handling hate speech.

Figure E: Confusion Matrix for TF-IDF+SVM trained on balanced Davidson Dataset.

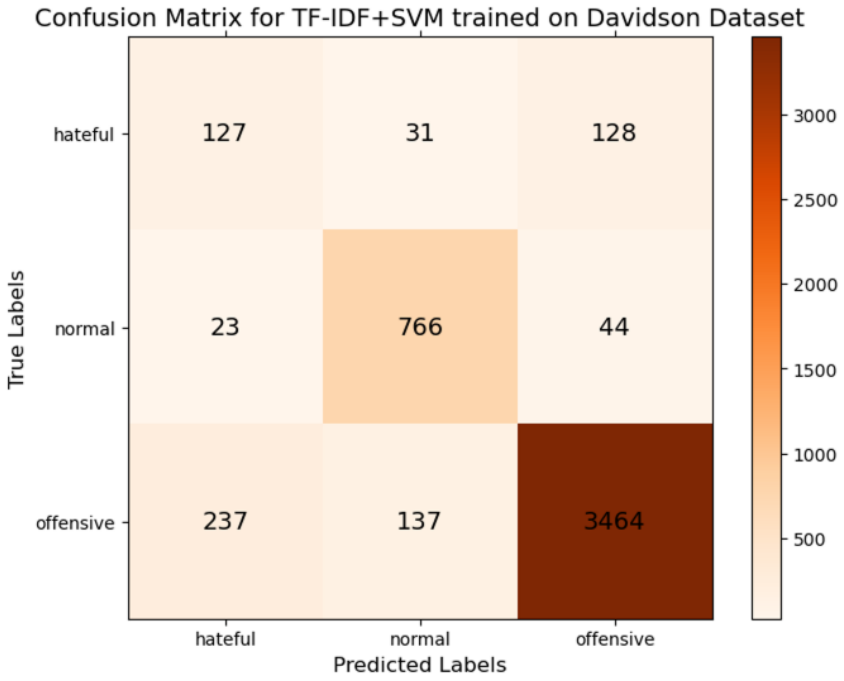


Figure E shows a confusion matrix for the TF-IDF + SVM model trained on the Davidson dataset with balanced class distributions. The model correctly identifies 3,464 offensive instances but misclassifies 137 normal and 237 hateful tweets as offensive. In this case, balancing class distributions led this model to have greater difficulty distinguishing between the hateful and offensive class, compared to the results of the model trained on an imbalanced class distribution, to the detriment of performance on the normal class. For normal speech, 766 instances are correctly classified, with 23 being mistaken for hateful and 44 misclassified as offensive. The classification of hateful speech is poor, but improved from figure A, with 127 instances correctly identified (compared with 54 previously), while 128 are mislabeled as offensive and 31 as normal.

Figure F: ROC Curve for TF-IDF+SVM trained on balanced Davidson Dataset.

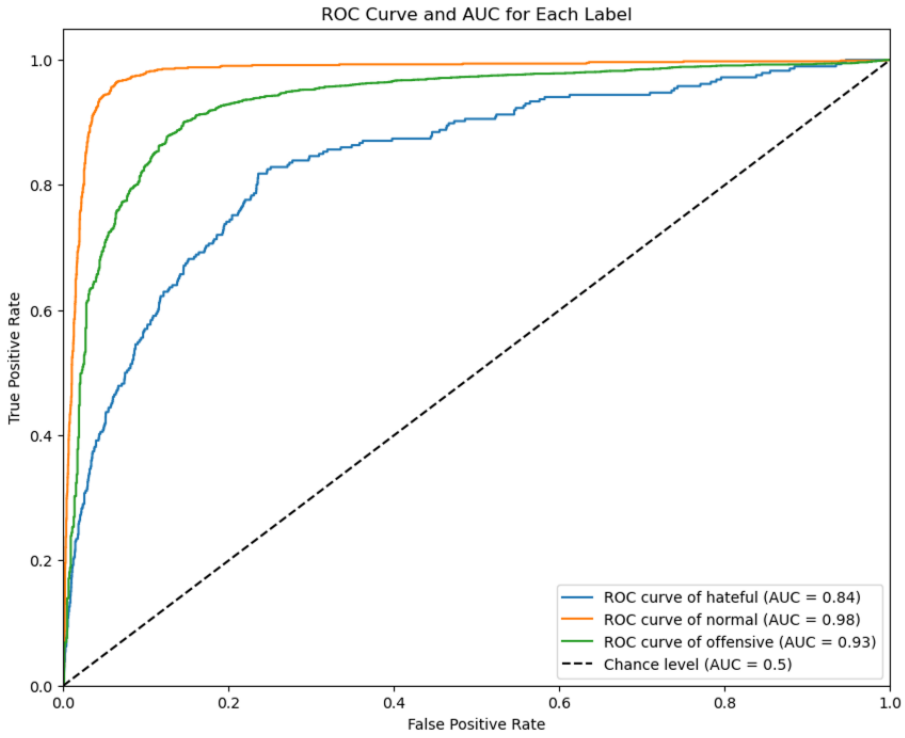


Figure F shows an ROC curve and AUC scores for the TF-IDF+SVM model trained on the Davidson Dataset with balanced class distributions. The AUC for the normal class is the highest at 0.98, suggesting that when the model correctly classifies normal tweets, it does so with high confidence. Offensive speech follows closely with an AUC of 0.93, indicating strong classification ability for this category as well. However, the hateful class has the lowest AUC at 0.84, reflecting a weaker ability to differentiate hateful content from other categories.

B.2 Model Performance on Balanced Dataset Class Distribution

In order to improve performance on the minority class, we balanced dataset class distributions. Table D reports findings from our experimentation with balanced class distributions in the datasets.

Table D: Balanced Dataset Results. Accuracy results from the initial Glove+LSTM model on three hate speech datasets. Although originally reported to achieve state-of-the-art performance, later findings suggest the model was overfit to its training data [42].

Model	Macro			Weighted Average			AUC Average	Overall Accuracy
	Precision	Recall	F1	Precision	Recall	F1		
Results on Davidson Dataset								
TF-IDF + SVM	0.70	0.76	0.72	0.89	0.88	0.89	0.92	0.88
Glove + LSTM	0.02	0.33	0.04	0.01	0.06	0.01	0.50	0.06
DistilBERT	0.73	0.78	0.76	0.91	0.90	0.90	0.86	0.90
Results on Founta Dataset								
TF-IDF + SVM	0.70	0.74	0.70	0.90	0.86	0.87	0.91	0.86
Glove + LSTM	0.09	0.33	0.15	0.08	0.28	0.12	0.50	0.28
DistilBERT	0.71	0.75	0.73	0.90	0.88	0.88	0.84	0.88
Results on HateXplain								
TF-IDF + SVM	0.61	0.62	0.61	0.62	0.62	0.62	0.79	0.62
Glove + LSTM	0.10	0.33	0.15	0.08	0.29	0.13	0.50	0.29
DistilBERT	0.65	0.65	0.64	0.66	0.64	0.65	0.73	0.64

With the balanced class distributions, we still see TF-IDF + SVM performing well, achieving the highest weighted average F1 scores across all datasets, with 0.89 on Davidson, 0.87 on Founta, and 0.62 on HateXplain. DistilBERT follows closely, with strong performance across datasets, particularly on Davidson (0.90 weighted F1) and Founta (0.88 weighted F1). Notably, it surpasses TF-IDF + SVM in macro F1 scores for all datasets, indicating stronger balance across classes. Glove + LSTM continues to struggle with performance. It struggles with both precision and recall, leading to very low F1 scores – especially on Davidson, where its macro F1 is just 0.04 and weighted F1 is 0.01. This trend is consistent across the Founta and HateXplain datasets, where its highest weighted F1 is only 0.29. The AUC scores further reinforce this pattern, as Glove + LSTM reaches only 0.50 on Davidson and Founta, barely above random chance.

The best model-dataset combination appears to be TF-IDF + SVM on the Davidson dataset, where it achieves the highest weighted average F1 score of 0.89 and an overall accuracy

of 0.88. Additionally, it has an AUC score of 0.92, indicating strong classification performance across all classes. A close second is DistilBERT on the Davidson dataset, which achieves a weighted average F1 score of 0.90 and a slightly lower AUC of 0.86. A further look into these results can be found in the following confusion matrices and ROC curves.

Figure G: Confusion Matrix for TF-IDF+SVM trained on balanced Davidson Dataset.

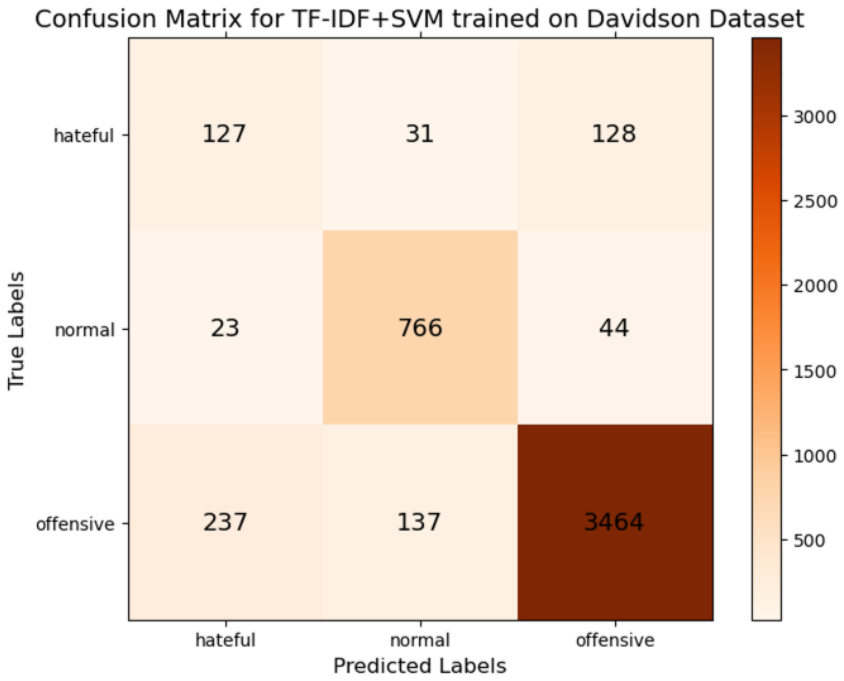


Figure G shows a confusion matrix for the TF-IDF + SVM model trained on the Davidson dataset with balanced class distributions. The model correctly identifies 3,464 offensive instances but misclassifies 137 normal and 237 hateful tweets as offensive. In this case, balancing class distributions led this model to have greater difficulty distinguishing between the hateful and offensive class, compared to the results of the model trained on an imbalanced class distribution, to the detriment of performance on the normal class. For normal speech, 766 instances are correctly classified, with 23 being mistaken for hateful and 44 misclassified as offensive. The classification of hateful speech is poor, but improved from figure A, with 127

instances correctly identified (compared with 54 previously), while 128 are mislabeled as offensive and 31 as normal.

Figure H: ROC Curve for TF-IDF+SVM trained on balanced Davidson Dataset.

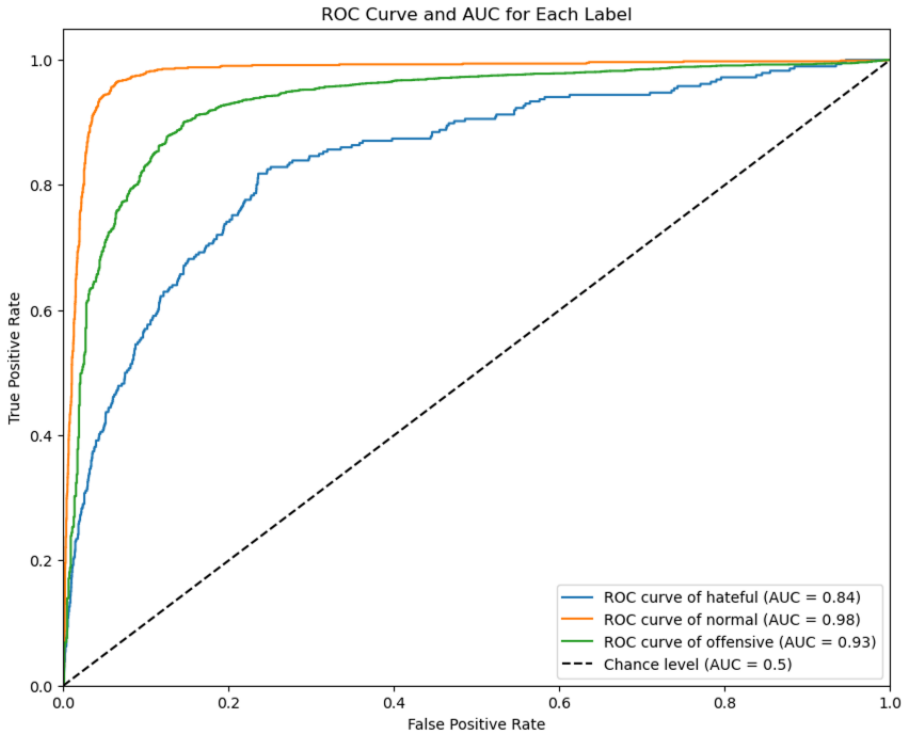


Figure H shows an ROC curve and AUC scores for the TF-IDF+SVM model trained on the Davidson Dataset with balanced class distributions. The AUC for the normal class is the highest at 0.98, suggesting that when the model correctly classifies normal tweets, it does so with high confidence. Offensive speech follows closely with an AUC of 0.93, indicating strong classification ability for this category as well. However, the hateful class has the lowest AUC at 0.84, reflecting a weaker ability to differentiate hateful content from other categories.

Figure I: Confusion Matrix for DistilBERT trained on balanced Davidson Dataset.

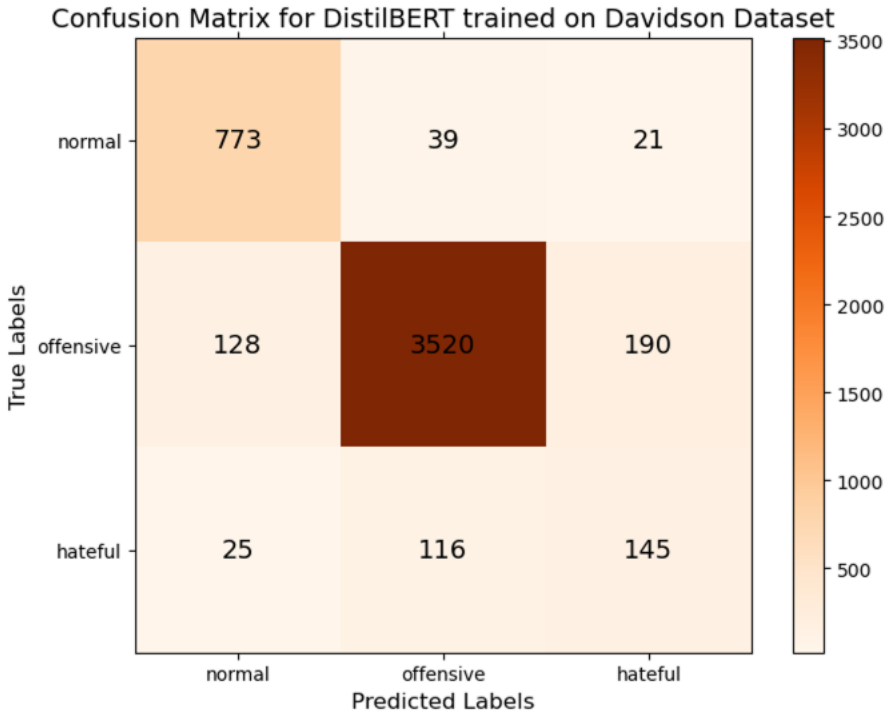


Figure I's confusion matrix represents results from the DistilBERT model trained on the Davidson dataset with balanced class distributions. The model performs well in classifying offensive tweets, correctly identifying 3,520 instances, with 128 misclassified as hateful and 190 misclassified as offensive. This suggests that the model has a strong bias towards correctly identifying offensive tweets. Similarly, normal tweets are well classified, with 773 correct predictions, but 39 instances were misclassified as offensive and 21 as hateful. However, performance dips when it comes to hateful tweets. Out of all hateful tweets, only 145 were correctly classified, while 25 were misclassified as normal and 116 as offensive.

Figure J: ROC Curve for DistilBERT trained on balanced Davidson Dataset.

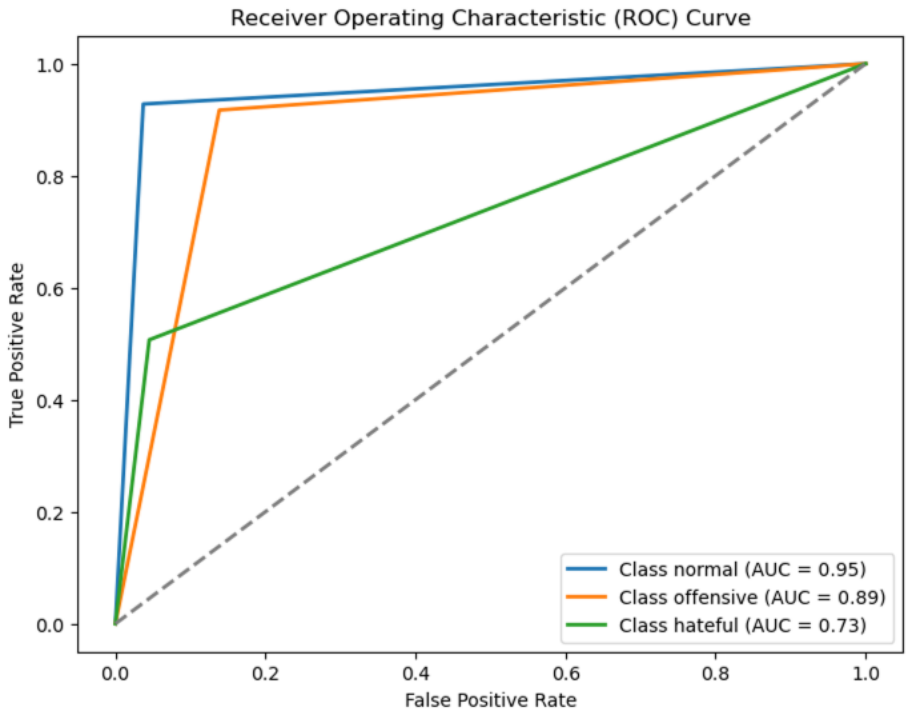


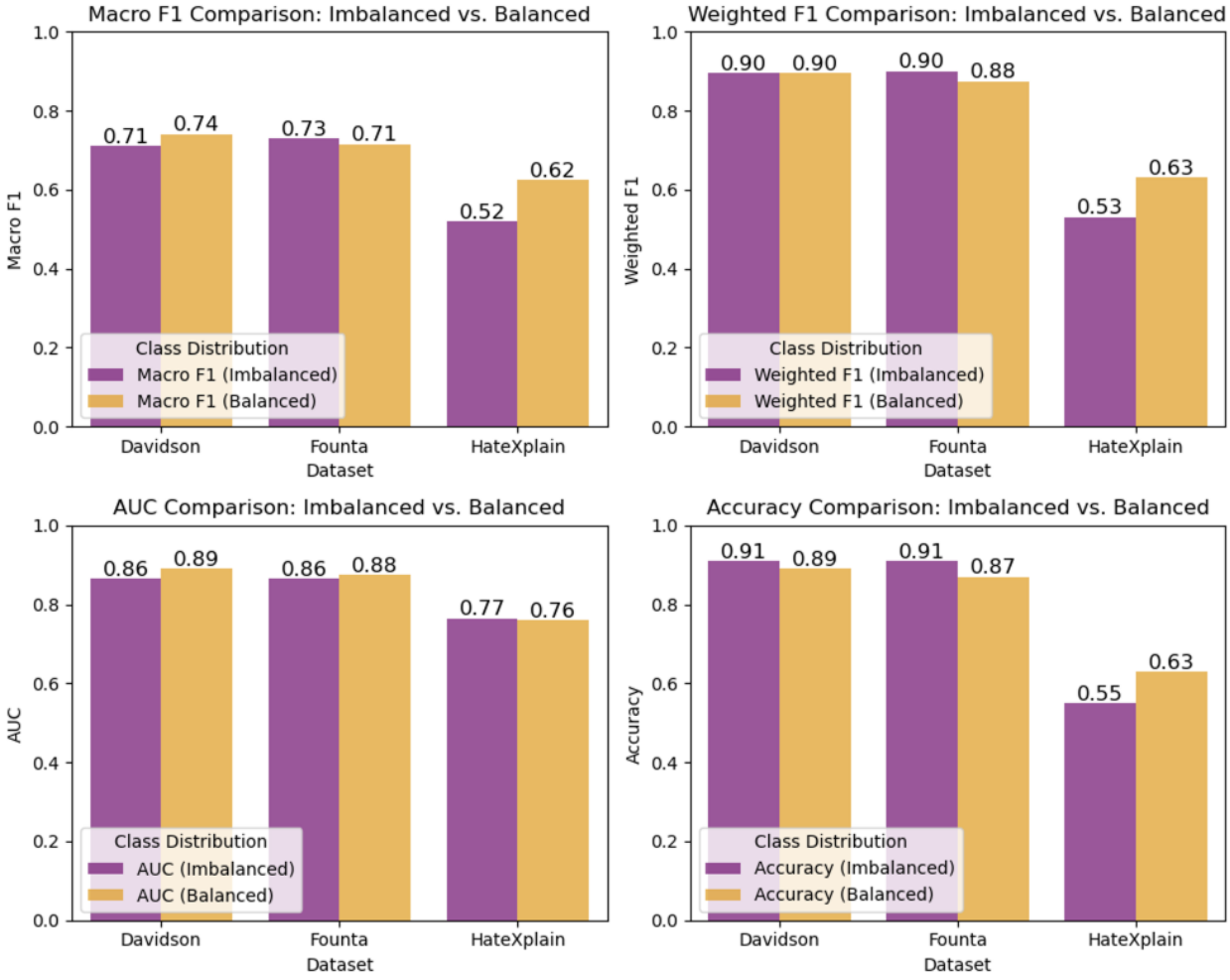
Figure J displays the ROC curve for the DistilBERT model trained on the Davidson dataset with balanced class distributions. It further supports the conclusion that the model struggles the most with the hateful class. The AUC for hateful tweets is 0.73, significantly lower than the scores for normal (0.95) and offensive (0.89) tweets. This indicates that the model has difficulty distinguishing hateful tweets from the other two classes, leading to more false positives and false negatives.

Analysis of Model Performance: Imbalanced vs. Balanced Class Distributions

When comparing models trained on imbalanced and balanced class distributions, we see that TF-IDF + SVM remains one of the most stable models, with only minor changes in overall performance. On the Davidson dataset, balancing led to a slight improvement in macro recall (0.76 vs. 0.68), indicating better classification of minority classes, but a slight drop in overall accuracy (0.88 vs. 0.92), which is expected when removing the bias toward majority classes. A

similar trend appears in the Founta dataset, where macro F1 slightly declines from 0.73 to 0.70, but AUC remains constant at 0.91, showing that the model's ranking ability is unaffected. However, on the HateXplain dataset, balancing significantly improves TF-IDF + SVM's performance, increasing macro and weighted F1 from 0.40 to 0.62, while overall accuracy jumps from 0.44 to 0.62. Suggesting that balancing class distributions had the greatest effect on this dataset.

Figure K: Comparing results of balanced class distribution on model performance



As shown in Figure K, HateXplain demonstrates significant improvements across all metrics, with the exception of its AUC score, which saw a minimal change (0.77 vs. 0.76). For the Davidson and Founta datasets, balancing the class distributions resulted in both slight

improvements and slight declines in performance. Most notable is the improvement in the Davidson Macro F1 (0.71 vs 0.74) and AUC score (0.86 vs 0.88). Founta also saw the most improvement in AUC score (0.86 vs 0.88).

B.3 Model Performance by per class distribution

Although macro and weighted average results presented so far align with previous work, a direct comparison of per-class metrics is challenging, as they are not often reported in the literature due to the common issue of class imbalances in datasets. In this section, we present the per-class performance metrics for the models we have already evaluated above for both the balanced and imbalanced class distributions.

B.3.1 Per-Class Metrics on Imbalanced Datasets (Original Class Distribution)

In this section, we examine the per-class performance metrics for our models when evaluated on the imbalanced datasets, using the original class distributions. Given the challenges posed by class imbalance in many real-world datasets, it is crucial to assess how well the models perform on each individual class. This detailed analysis helps to highlight the strengths and weaknesses of our models across different classes, which may not be apparent when only considering aggregate metrics.

Table E: Per-Class metrics with models trained on original class distributions

Label	Precision	Recall	F1-Score	Accuracy
Davidson trained on DistilBERT				
Normal	0.84	0.91	0.88	0.91
Offensive	0.94	0.95	0.95	0.95
Hateful	0.50	0.27	0.35	0.27
Davidson trained on TF-IDF				
Normal	0.85	0.90	0.87	0.90
Offensive	0.93	0.96	0.94	0.96
Hateful	0.59	0.19	0.29	0.19
Founta trained on DistilBERT				
Normal	0.94	0.97	0.95	0.97
Offensive	0.88	0.90	0.89	0.90
Hateful	0.60	0.26	0.36	0.26
Founta trained on TF-IDF				
Normal	0.94	0.97	0.95	0.97
Offensive	0.88	0.91	0.90	0.91
Hateful	0.60	0.22	0.33	0.22
HateXplain trained on DistilBERT				
Normal	0.73	0.68	0.70	0.68
Offensive	0.50	0.54	0.52	0.54
Hateful	0.73	0.75	0.74	0.75
HateXplain trained on TF-IDF				
Normal	0.45	0.73	0.56	0.73
Offensive	0.42	0.29	0.34	0.29
Hateful	0.41	0.19	0.26	0.19

Table E presents the per-class performance metrics for the TF-IDF and DistilBERT models trained on the datasets with original (imbalanced) class distributions. For the Davidson dataset, the DistilBERT model shows strong performance in identifying normal and offensive classes. The normal class achieves a precision of 0.84, recall of 0.91, and an F1-score of 0.88, indicating that DistilBERT is effective at recognizing normal content while maintaining a good balance of precision and recall. The offensive class performs even better, with precision of 0.94, recall of 0.95, and F1-score of 0.95, which demonstrates the model's ability to accurately detect offensive content. However, the hateful class performance is notably poor, with a precision of 0.50, recall of 0.27, and F1-score of 0.35, suggesting that DistilBERT struggles to identify

hateful content in the Davidson dataset. When trained with TF-IDF, the Davidson model shows similar trends. The normal class has slightly improved precision (0.85) and slightly lower recall (0.90), compared to DistilBERT, while the offensive class maintains strong performance with precision of 0.93 and recall of 0.96. However, the hateful class's performance with TF-IDF significantly drops, with precision of 0.59, recall of 0.19, and F1-score of 0.29, indicating that TF-IDF struggles substantially to identify hateful content in the Davidson dataset.

In the Founta dataset, DistilBERT performs well in identifying normal content, achieving a precision of 0.94, recall of 0.97, and F1-score of 0.95. The Offensive class also shows strong performance, with precision of 0.88, recall of 0.90, and F1-score of 0.89. However, similar to the Davidson dataset, the hateful class presents a challenge, with a precision of 0.60, recall of 0.26, and F1-score of 0.36. When the Founta dataset is trained with TF-IDF, the performance on the normal and offensive classes is almost identical to the DistilBERT model, with precision values of 0.94 and recall values of 0.97 for normal tweets, and precision of 0.88 and recall of 0.91 for the offensive class. However, the hateful class still shows poor results, with precision of 0.60, recall of 0.22, and F1-score of 0.33. These findings underscore the limitations of the TF-IDF model in detecting hateful content, as it fails to accurately identify this class across both the Davidson and Founta datasets.

When the HateXplain dataset is trained on DistilBERT, the model shows relatively balanced performance across the different labels. The hateful class achieves the highest scores across all metrics, with a precision of 0.73, recall of 0.75, and an F1-score of 0.74, which indicates that the model is quite effective at identifying hateful content. The normal class also performs reasonably well, with a precision of 0.73, but its recall drops to 0.68, reflecting some challenges in fully capturing all instances of normal content. The offensive class shows the

weakest performance, with precision and recall values of 0.50 and 0.54, respectively, resulting in an F1-score of 0.52, which suggests that the model struggles to consistently identify offensive content (which is the minority class in the HateXplain dataset). For HateXplain trained on TF-IDF, the performance is notably poorer, particularly for the offensive and hateful classes. The normal class, however, sees a high recall of 0.73, but the precision drops to 0.45, indicating that the model struggles with false positives – classifying many non-normal instances as normal. The offensive class' performance is poor, with both precision and recall low at 0.42 and 0.29, respectively, yielding a very low F1-score of 0.34. This suggests that the TF-IDF model has difficulty identifying offensive content accurately. The hateful class also suffers from low precision (0.41) and recall (0.19), leading to a poor F1-score of 0.26.

In summary, DistilBERT outperforms TF-IDF in all cases across all datasets. DistilBERT also achieves the best performance on the hateful class when trained on the HateXplain dataset. These results highlight the strengths of transformer-based models like DistilBERT in handling nuanced text classification tasks, but also indicate that further improvements are needed to better detect hateful speech.

B.3.2 Per-Class Metrics on Balanced Datasets

In this section, we focus on the per-class performance metrics for our models when trained on balanced datasets, where class distributions have been adjusted to mitigate the effects of imbalance. By examining these metrics, we can assess how well the models generalize across classes when given equal class representation. This analysis provides insight into the model's ability to detect the minority classes more effectively, offering a clearer picture of its performance in scenarios where class balance is intentionally managed.

Table F: Per-Class metrics with models trained on balanced class distributions

Label	Precision	Recall	F1-Score	Accuracy
Davidson trained on DistilBERT				
Normal	0.83	0.93	0.88	0.93
Offensive	0.96	0.91	0.94	0.92
Hateful	0.41	0.51	0.45	0.51
Davidson trained on TF-IDF				
Normal	0.82	0.92	0.87	0.92
Offensive	0.95	0.90	0.93	0.90
Hateful	0.33	0.44	0.38	0.44
Founta trained on DistilBERT				
Normal	0.93	0.93	0.94	0.93
Offensive	0.87	0.84	0.86	0.84
Hateful	0.32	0.49	0.38	0.49
Founta trained on TF-IDF				
Normal	0.95	0.91	0.93	0.91
Offensive	0.89	0.81	0.85	0.81
Hateful	0.24	0.50	0.33	0.50
HateXplain trained on DistilBERT				
Normal	0.74	0.59	0.66	0.59
Offensive	0.48	0.59	0.53	0.59
Hateful	0.73	0.75	0.74	0.75
HateXplain trained on TF-IDF				
Normal	0.67	0.66	0.66	0.66
Offensive	0.50	0.46	0.48	0.46
Hateful	0.67	0.73	0.70	0.73

Table F presents the per-class performance metrics for the three different datasets trained on the TF-IDF and DistilBERT models. When trained on the Davidson dataset, DistilBERT shows strong performance on the normal and offensive classes. Precision for normal is 0.83, with a high recall of 0.93, resulting in a solid F1-score of 0.88. The offensive class performs well, with precision of 0.96, recall of 0.91, and F1-score of 0.94, suggesting that DistilBERT effectively identifies offensive content. However, the hateful class shows notable weakness, with precision of 0.41, recall of 0.51, and F1-score of 0.45, which is considerably lower compared to the other two classes. This indicates that DistilBERT struggles with identifying hateful content in the Davidson dataset. TF-IDF on the Davidson dataset

exhibits similar trends, with normal having precision of 0.82 and recall of 0.92, leading to an F1-score of 0.87. The offensive class performs well, achieving precision of 0.95 and recall of 0.90, resulting in an F1-score of 0.93. However, like DistilBERT, TF-IDF also struggles with the hateful class, as its precision of 0.33 and recall of 0.44 result in a poor F1-score of 0.38.

When trained on the Founta dataset, DistilBERT shows good performance on the normal class, with precision of 0.93, recall of 0.93, and an F1-score of 0.94. The offensive class also performs well, with precision of 0.87 and recall of 0.84, resulting in an F1-score of 0.86. However, the hateful class remains challenging for DistilBERT, with a precision of 0.32, recall of 0.49, and an F1-score of 0.38. TF-IDF on the Founta dataset results in slightly higher performance for the normal class, with precision of 0.95 and recall of 0.91, resulting in an F1-score of 0.93. The offensive class sees a decline in performance, with precision of 0.89 and recall of 0.81, leading to a lower F1-score of 0.85. The hateful class' performance with TF-IDF is also poor, with precision of 0.24, recall of 0.50, and an F1-score of 0.33.

When trained on the HateXplain dataset, DistilBERT performs most poorly on the normal and offensive classes. The precision for normal is 0.74, with a low recall of 0.59, resulting in a poor F1-score of 0.66. Similarly, the offensive class shows precision of 0.48, recall of 0.59, and an F1-score of 0.53. On the other hand, hateful tweets are detected much better by DistilBERT, with a precision of 0.73, recall of 0.75, and an F1-score of 0.74, suggesting that DistilBERT excels at identifying hateful content in this dataset. TF-IDF on the HateXplain dataset shows even weaker performance across the board. The normal class has a precision of 0.67 and recall of 0.66, resulting in an F1-score of 0.66. The offensive class performs poorly with a precision of 0.50 and recall of 0.46, yielding an F1-score of 0.48. Hateful content, however, is better detected

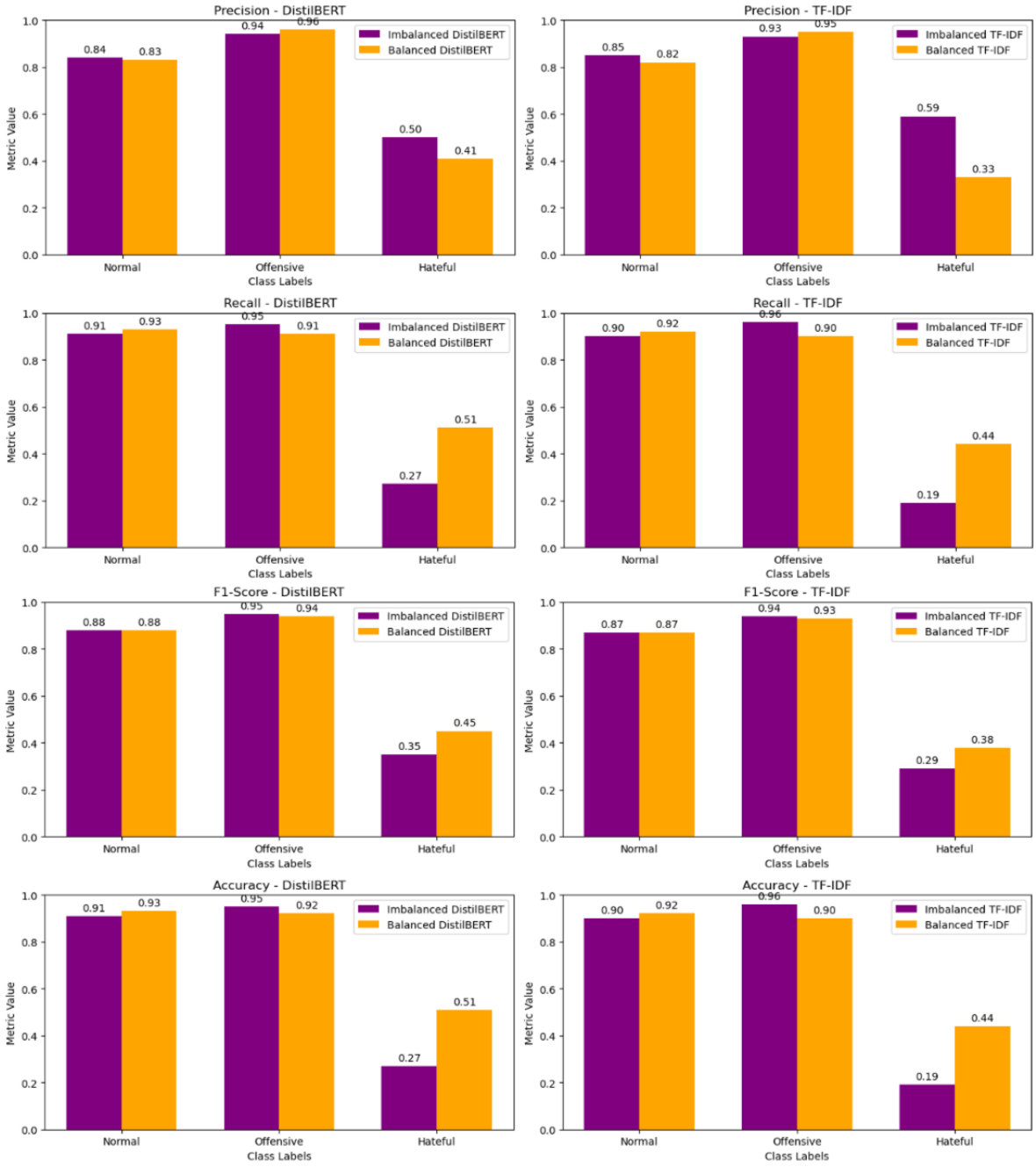
with TF-IDF compared to the other classes, with precision of 0.67, recall of 0.73, and an F1-score of 0.70.

In summary, DistilBERT consistently outperforms TF-IDF across all datasets for detecting offensive and normal classes, showing higher precision, recall, and F1-scores. However, both models show notable difficulties in identifying hateful content, with the hateful class performing significantly worse than the normal and offensive classes, except in the HateXplain dataset. DistilBERT generally performs better than TF-IDF for detecting hateful content but still faces challenges.

Analysis of per-class performance: Imbalanced vs. Balanced Class Distribution

When comparing the results of models trained on imbalanced versus balanced datasets, we observe differences in per-class performance across datasets. This comparison helps in understanding how class balancing affects model performance, particularly for minority classes like hateful.

Figure L: Per-Class metric comparison (imbalanced vs balanced) for Davidson dataset



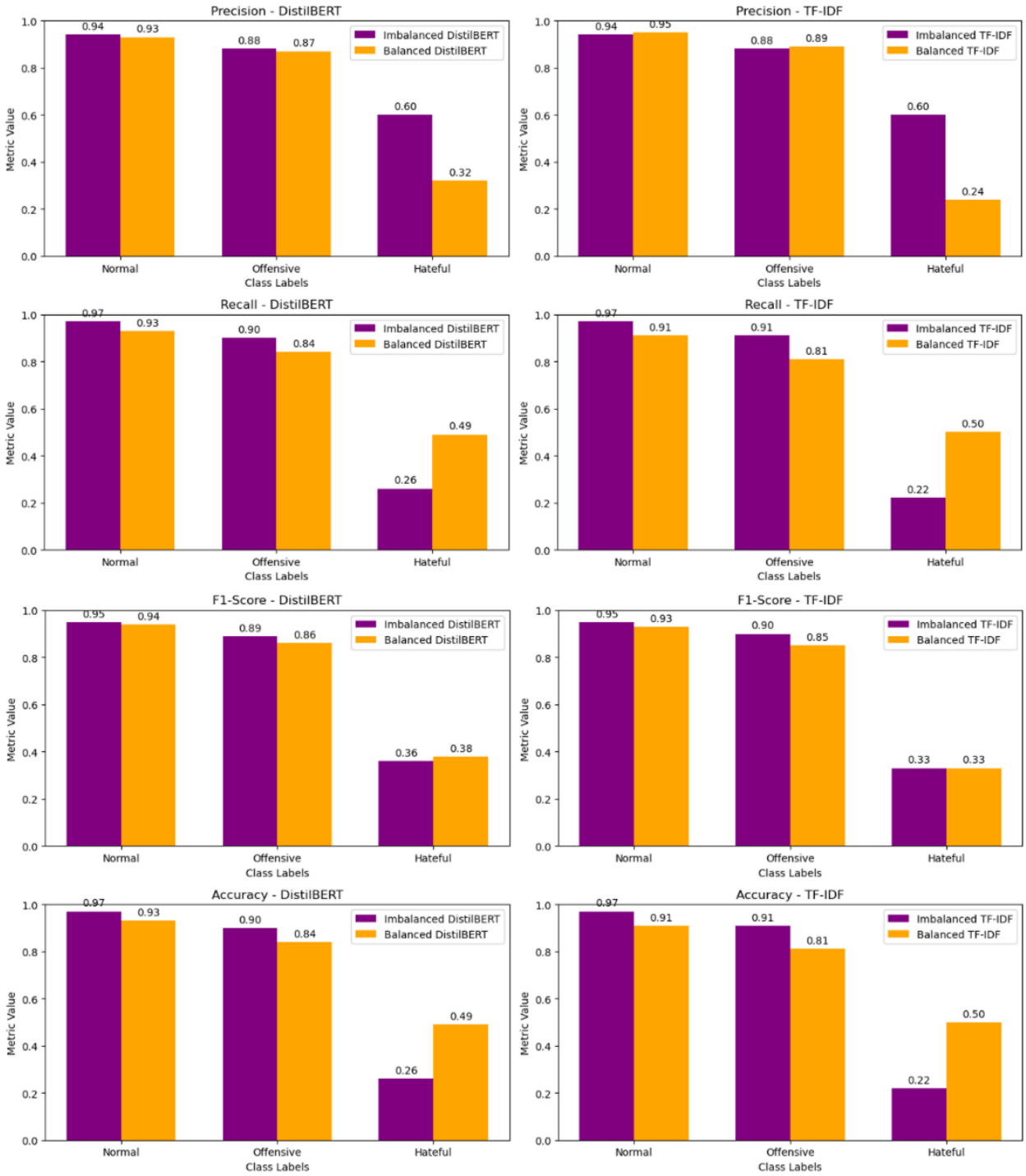
Davidson Dataset

When comparing the performance of DistilBERT on imbalanced vs. balanced datasets, a few key trends emerge, particularly in how the model handles the hateful class. For the normal

class, DistilBERT performs well on the imbalanced dataset, achieving a precision of 0.84 and recall of 0.91, which leads to an F1-score of 0.88. On the balanced dataset, precision slightly drops to 0.83, but recall improves to 0.93, keeping the F1-score constant at 0.88. In terms of the offensive class, DistilBERT demonstrates excellent performance in both cases. When the dataset is balanced, recall decreases slightly to 0.91, which leads to a small drop in F1-score from 0.95 to 0.94. The most notable difference appears in the hateful class. When the dataset is balanced, performance improves. Precision rises to 0.41, recall increases to 0.51, and the F1-score jumps to 0.45. This suggests that balancing the dataset helps improve the detection of hateful speech, although the performance is still much lower than for the other classes.

The performance of TF-IDF on the imbalanced vs. balanced datasets shows similar trends to those seen with DistilBERT, though there are some differences in the degree of improvement for each class. For the normal class, precision and recall remain very close across both datasets, indicating that balancing the dataset does not significantly affect the performance on the normal class. For the offensive class, performance is almost identical across both datasets. The hateful class once again presents a challenge for TF-IDF, particularly on the imbalanced dataset, where precision is 0.59 and recall is just 0.19, resulting in a poor F1-score of 0.29. When the dataset is balanced, there is some improvement. Precision rises slightly to 0.33, and recall increases to 0.44, which leads to a small boost in the F1-score to 0.38. Despite this improvement, performance on the hateful class remains far behind that of the other classes, highlighting that balancing the dataset does help, but challenges in detecting hateful speech persist.

Figure M: Per-Class metric comparison (imbalanced vs balanced) for Founta dataset



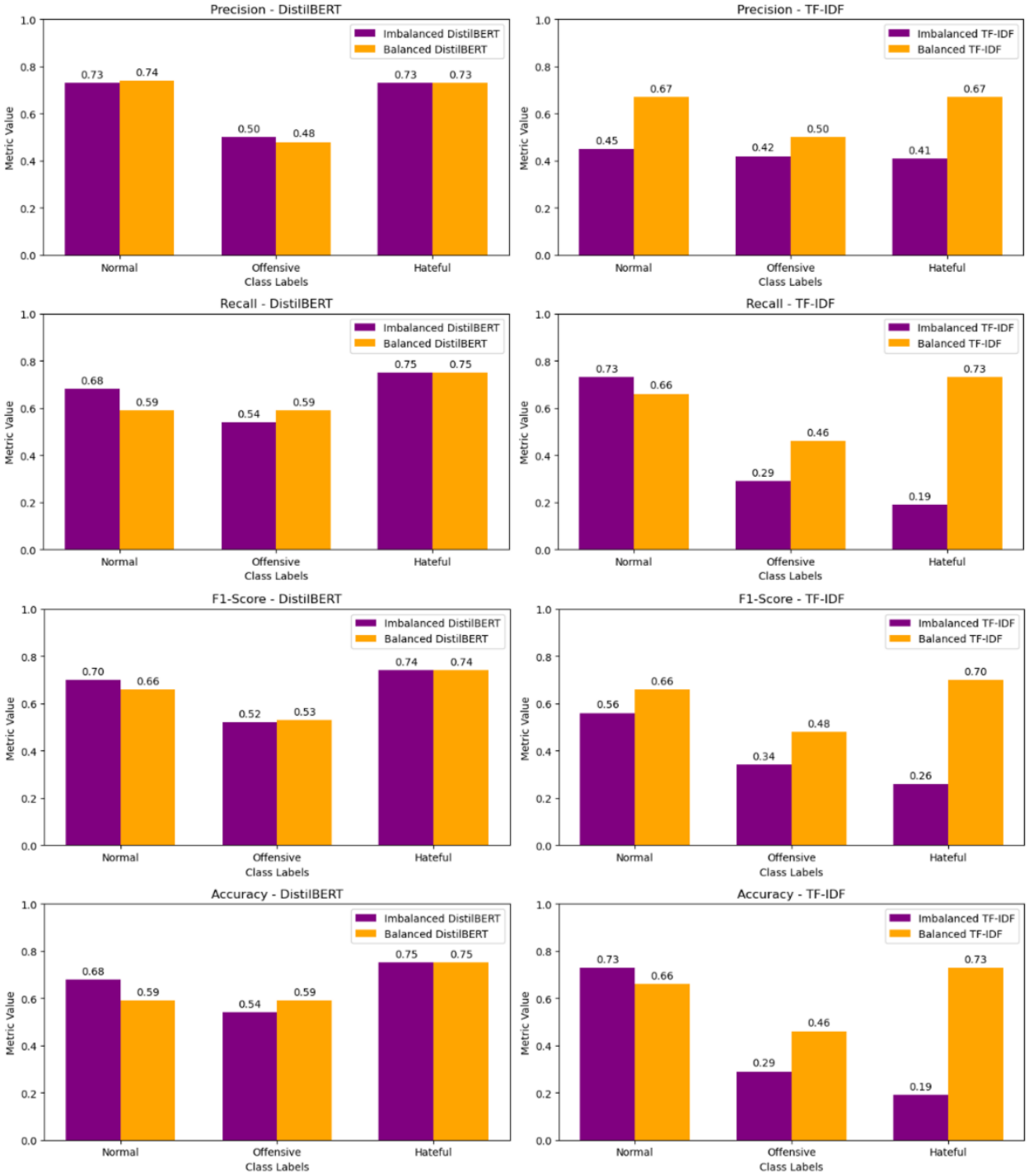
Founta Dataset

In comparing DistilBERT's performance when trained on the balanced vs imbalanced Founta dataset, we see consistently good performance on the normal class. However, the

offensive class experiences a noticeable decline in performance when the dataset is balanced. Precision drops to 0.87 and recall to 0.84, resulting in an F1-score of 0.86, suggesting that balancing the dataset may reduce the model's sensitivity to offensive content. The hateful class continues to struggle, with a drop in precision (0.32). However, recall, and overall accuracy improved.

The TF-IDF model shows similar results with relatively strong performance on the normal class across balanced and imbalanced datasets. We also see a similar drop in metrics for the offensive class. Again, precision dropped for the hateful class, but all other metrics saw improvement with class balancing.

Figure N: Per-Class metric comparison (imbalanced vs balanced) for HateXplain dataset



HateXplain Dataset

For the HateXplain dataset, the impact of balancing the dataset varies across different classes and models. When using DistilBERT, the normal class experiences a slight drop in

performance when the dataset is balanced, with its F1-score decreasing from 0.70 to 0.66 due to a noticeable decline in recall (from 0.68 to 0.59). Similarly, the offensive class sees a minor increase in recall (from 0.54 to 0.59) but a slight drop in precision (from 0.50 to 0.48), keeping the F1-score relatively stable at around 0.52–0.53. However, the hateful class remains unaffected by balancing, maintaining a performance with an F1-score of 0.74 across both imbalanced and balanced conditions. This suggests that DistilBERT is consistently able to detect hateful content, while balancing the dataset slightly reduces its ability to identify normal speech.

With TF-IDF, the effect of balancing the dataset is more pronounced. The normal class sees a notable improvement, with its F1-score increasing from 0.56 to 0.66 as precision rises significantly (from 0.45 to 0.67). Similarly, the hateful class benefits greatly from balancing, with its F1-score improving from 0.26 to 0.70, primarily due to a significant boost in recall (from 0.19 to 0.73). However, the offensive class does not see similar gains; although precision increases slightly (from 0.42 to 0.50), recall decreases (from 0.29 to 0.46), leading to only a modest improvement in the F1-score (from 0.34 to 0.48). These results indicate that while balancing helps TF-IDF better detect both normal and hateful speech, it does little to improve the model's ability to classify offensive content correctly.

Key Observations

When comparing the imbalanced and balanced datasets, we generally see that models trained on balanced datasets perform worse on the offensive and normal classes, especially for the Founta and HateXplain datasets. This could be due to the increased emphasis on the minority hateful class after balancing the dataset, which reduces overall performance on the other classes. The hateful class tends to perform better when models are trained on balanced datasets, especially in HateXplain for DistilBERT and TF-IDF. However, the improvement is not uniform

across all datasets, and some minority class performance still lags. DistilBERT tends to outperform TF-IDF on all datasets, with DistilBERT consistently yielding higher precision, recall, and F1-scores for both imbalanced and balanced scenarios.

B.4 Additional Confusion Matrices From Cross-Domain Model Results

Figure O: Founta source domain testing on Davidson (DistilBERT)

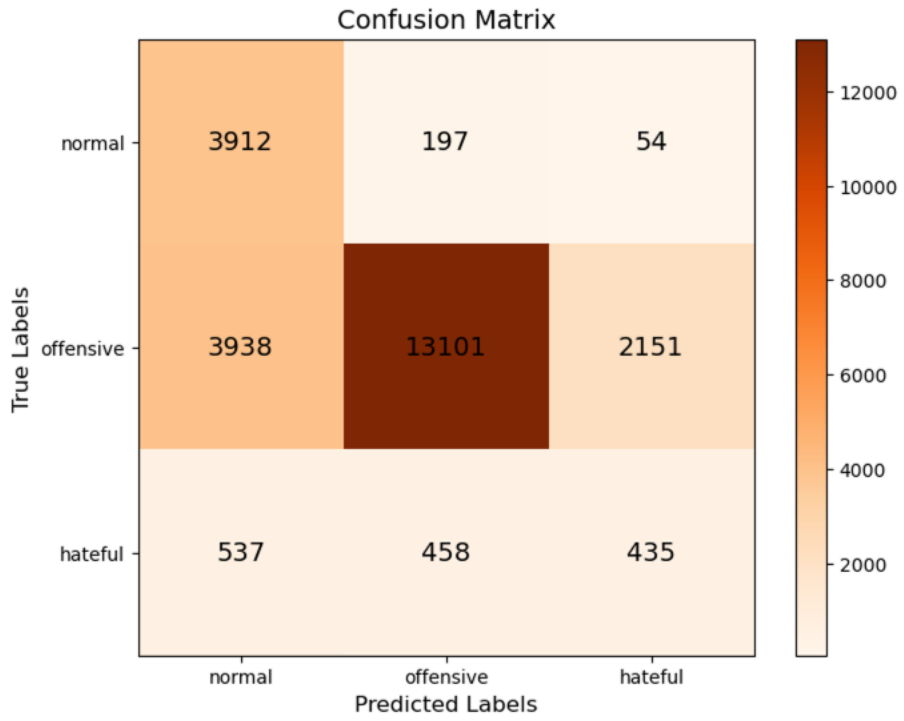


Figure O is a confusion matrix representing the results from testing the DistilBERT model (trained on the Founta dataset) on the Davidson validation set. It shows that the model does well in classifying offensive and normal tweets, however some offensive tweets are still classified as normal or hateful. The model does not classify hateful tweets well, and most commonly classifies them as normal.

Figure P: Founta source domain testing on HateXplain (DistilBERT)

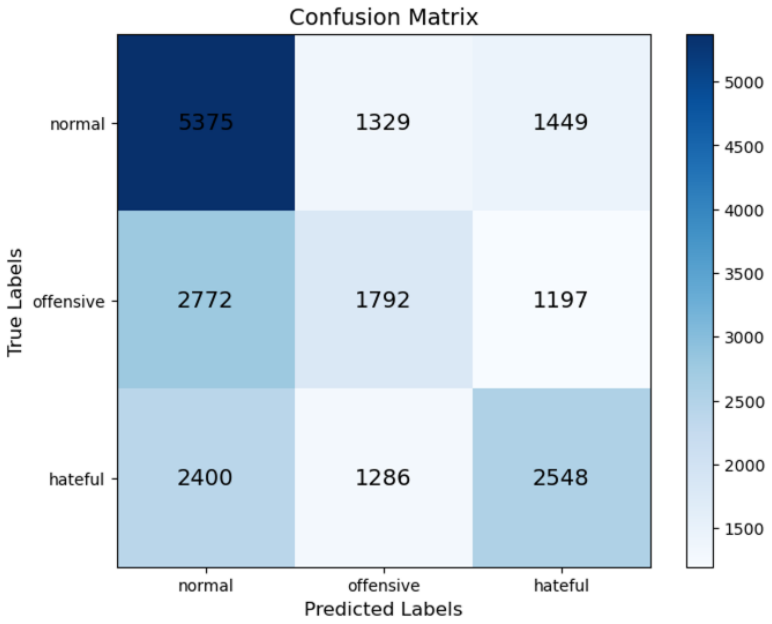


Figure P is a confusion matrix representing the results from testing the DistilBERT model (trained on the Founta dataset) on the HateXplain validation set. This model does best at calculating normal tweets. It also correctly classifies many hateful tweets correctly, but almost the same amount of hateful tweets are misclassified as normal, indicating that the model has a strong bias towards the normal class. This is also apparent in the offensive class, where the model largely misclassifies offensive tweets as normal.

Figure Q: HateXplain source domain testing on Founta (DistilBERT)

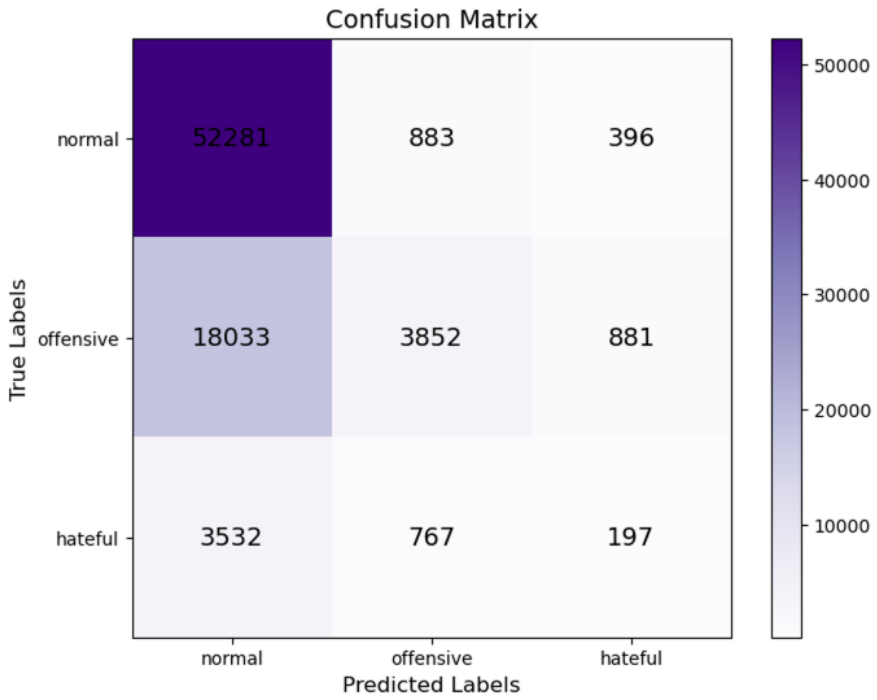


Figure Q is a confusion matrix representing the results from testing the DistilBERT model (trained on the HateXplain dataset) on the Founta validation set. It shows that the model does well in classifying normal tweets, but struggles in classifying hateful and offensive tweets. There appears to be a bias towards the normal class, as most hateful and offensive tweets are misclassified as normal.

Figure R: HateXplain source domain testing on Davidson

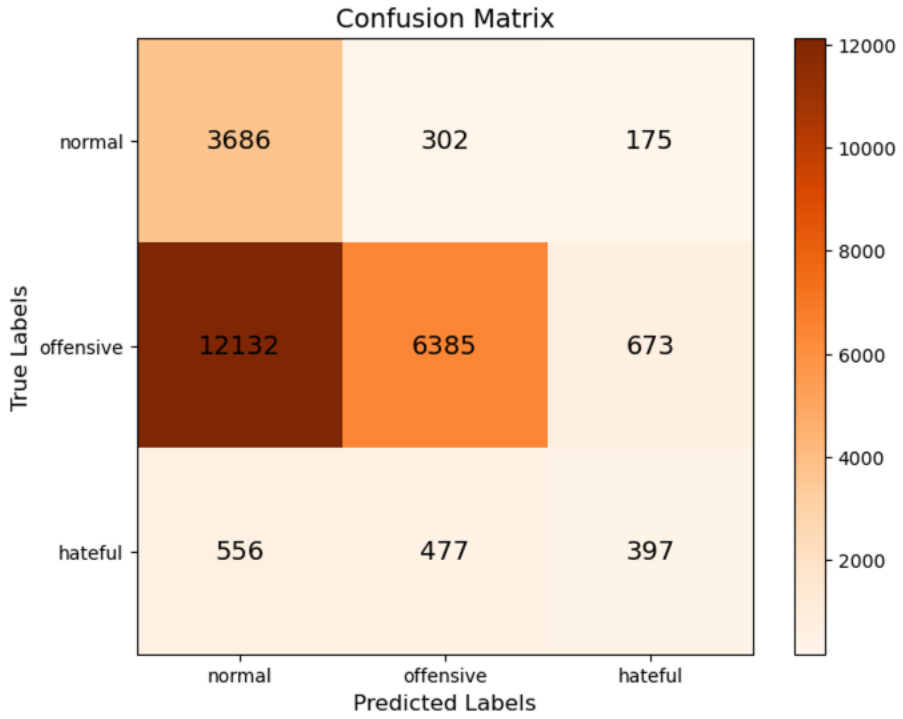


Figure R is a confusion matrix representing the results from testing the DistilBERT model (trained on the HateXplain dataset) on the Davidson validation set. It shows that the model does well in classifying normal tweets but struggles in classifying hateful and offensive tweets. There appears to be a bias towards the normal class, as most hateful and offensive tweets are misclassified as normal.