Fine-grained emotions influence on implicit hate speech detection

Amir Reza Jafari, Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, Noel Crespi

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Fine-grained Emotions Influence on Implicit Hate Speech Detection

AMIR REZA JAFARI, GUANLIN LI, PRABODA RAJAPAKSHA, REZA FARAHBAKHSH, NOEL CRESPI

Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France

Corresponding author: Amir Reza Jafari (e-mail: amir-reza.jafari_tehrani@telecom-sudparis.eu).

ABSTRACT Recent years brought an exponential growth of social media which revolutionized freedom of speech but significantly increased the propagation of hate speech and hate-based activities. Therefore, constructive countermeasures are necessary to prevent escalating hateful content on online social media. Many recent works target explicit hate speech, but only a few studies have utilized multiple fused features such as sentiment, targets, and emotions as attributes to enhance the detection of hate speech. In general, sentiment features help to discern feelings such as positivity or negativity, and emotion features provide a deeper level of granularity, focusing on a more comprehensive understanding of sensitivities. The aim of this paper is to investigate the significance of incorporating fine-grained emotions as an essential feature in improving the classification of implicit hate speech. First, we analyzed emotion variations of hateful and non-hateful content and explored their major fine-grained emotion discrepancies targeting implicit hateful content. Next, we introduce a multi-task learning approach that integrates emotions and sentiment features to classify implicit expressions of hatred. To evaluate the effectiveness of our multi-task learning approach, we compared it with baseline models using single-task learning approaches. The experimental results show that our multi-task approach outperformed in classifying implicit hate speech compared to the baseline models and demonstrates that fine-grained emotional knowledge decreases the classification error across multiple implicit hate categories.

INDEX TERMS Hate Speech, Emotion Analysis, Social Media, Implicit Hate, Multi-task learning

I. INTRODUCTION

Warning: *this paper contains content that may be offensive or upsetting.*

Hate speech or hate communication that disparages a person or a group based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics, has always been one of the biggest concerns since the advent of the Internet and social media. Popular social media platforms such as Twitter and Facebook allow people to share their thoughts and ideas freely and anonymously, which brings a vibrant environment for discussion but sometimes incites hurtful and insulting speeches. Due to the vast amount of content produced daily on social network platforms, it is impossible to monitor all the content manually and filter the harmful ones. Consequently, efficient automatic hate speech detection methods are required to study and prevent their spread [1], [2], [3].

As shown in Figure 1, hate speech can either be implicit or explicit [4]. The detection of implicit hatred is more challenging due to the lack of explicit cues to understand biases

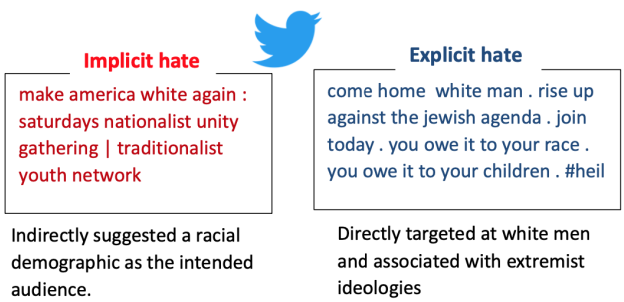


FIGURE 1: Implicit hate vs Explicit hate: The detection of explicit hate speech is relatively straightforward since it involves the use of direct and straightforward language to convey hatred. In contrast, detecting implicit hate speech poses greater challenges as it relies on understanding the overall meaning of a sentence and analyzing linguistic variations due to its subtle and indirect nature.

or ideologies [5]. Hence, implicit hate speech detection is a difficult task and requires identifying prominent textual features compared to explicit hate speech [6].

In the literature, various methods have been extensively utilized for explicit hate speech detection, including lexicon-based, rule-based, and linguistic feature-based and incorporated different embedding types with numerous Machine Learning (ML) techniques [7]. The aforementioned feature extraction techniques are not efficient to detect implicit hateful content due to sophisticated language patterns and expressions and a lack of hatred indicators. The recent advancements in hate speech detection usually focused on semantic features of textual content, but sentimental features were also utilized in hate speech detection [8] with a prevailing emphasis on the binary polarity (positive or negative). In addition, a few research works have utilized emotional features for the detection of hate speech, but their focus has been limited to specific emotional categories and a single hate category [9] [10], while implicit hate speech detection mainly revolves around data-driven approaches that used deep models and language models to capture semantic features [5]. These studies provide evidence that effectively addressing the nature of hate speech detection and techniques such as leveraging emotions and sentiment analysis can play an important role in uncovering hidden patterns and gaining insights into the underlying implications to better understand implicit hate speech.

Compared to binary polarity, emotion features are finer-grained and can express more subtle and complex sentiments in the text. Thus, exploring fine-grained emotions in hateful content for different hate categories is necessary to identify their impact on various hate speech types. In addition, this is also useful in exploiting emotion-agnostic hate speech detection models. This paper focuses on exploring the fine-grained emotions within implicit hateful content, highlighting their substantial variation in different emotion categories. After that, we conducted experiments to assess whether fine-grained emotions are useful features for classifying implicit hateful content.

The main contributions in this paper are two-folds, accompanied by several other sub-contributions:

1) *An extensive analysis of emotion distribution of hateful vs non-hateful content:*

- Examine the distribution of emotions in 3 levels (sentiment, Ekman, and fine-grained) in both non-hateful and hateful content.
- Explore the variations of fine-grained emotion categories within implicit hateful content.

2) *Implicit hate speech classification with single-task learning (STL) and multi-task learning (MTL) models:*

- Propose feature fusion STL embedding-based classifiers that integrate emotion and sentiment features with text features for the classification of implicit hate speech. Additionally, we conduct an ablation study to evaluate the effectiveness of the different features.
- Propose a multi-task approach based on a Transformer-based shared encoder, which incorporates three distinct heads: a hate speech head, a fine-grained emotions head, and a sentiment level head. We show the effectiveness of this model compared to the STL models when classifying implicit hate over several categories (incitement, inferiority, irony, stereotypical, threatening, and white grievance).
- To demonstrate the enhanced performances of the multi-task approach, we conduct an error analysis on the performance of both the single-task and multi-task setups.

The analysis results show that examining fine-grained emotions across multiple implicit hate categories provides different intensity levels within emotion categories. The experimental results validate that leveraging this variation improved the performances of the MTL-based implicit hate speech detection models.

The rest of the paper is structured as follows: Section II describes the background knowledge and related works. Datasets used for the analysis and evaluation are presented in Section III. In Section IV, the analysis of the emotional dimensions of hate speech is provided, while Section V presents the proposed architecture of our system with the experimental methodology and results along with error analysis on system performance. Lastly, Section VI concludes the work and presents future directions.

II. RELATED WORK

Hate speech is harmful and often targets an individual or group of people directly attacking hate towards them. Hate speech expressions can be explicit or implicit. Compared to explicit, implicit poses challenges such as fewer lexicons to learn the model, bias towards the dataset and its labelling etc. With the rise of hate speech in online social media, several automatic hate content detection models have been proposed in the literature which are based on natural language processing techniques. Most of the proposed models used supervised ML methods. However, the generalisability of those models is not noticeable when they are applied to other unseen datasets [11]. In addition, existing models are severely overestimated [12] and therefore, it still dares to apply these models in real-world scenarios.

A. HATE SPEECH DETECTION USING TEXT-BASED FEATURES

Many existing methods relied on the lexicons-based features, rule-based features, corpus-based approaches and probabilistic models [13], [14], [15]. These models solely depend upon domain-specific features and the co-occurrence of those features to decide whether the polarity of hate content is indirect. A number of corpus-based methods exhibit a high false positive rate since the classification entirely depends on a defined set of words which are used to recognize the polarity [16]. Another group of research works widely adapted linguistic features such as Bag-of-word (BoW) approaches to detect hate speech. These BoW-based hate speech detection methods rely on offensive keywords [17]. Apart from

that, n-grams are also used in hate speech classification and improved the performance of those models compared to the BoW. Davidson et al. [18] used n-grams to build their classifier and demonstrated and analyzed why hateful content is misclassified in the BoW model. The main reason for this is that n-gram models are capable of capturing consecutive words of varying size, but BoW losses this ability to recognize from the given sentence. In this study, we delve into the text-based characteristics of hate speech by employing diverse text-embedding techniques. Furthermore, we leverage these features, in conjunction with emotion and sentiment attributes, to focus on the implicit hate speech task.

B. HATE SPEECH DETECTION USING TEXT EMBEDDING-BASED METHODS

Text embedding methods are used to train hate speech classifiers and they are capable of capturing semantic meaning through word vectors. Djuric et al. [19] applied sentence embedding (paragraph2vec). Their results outperformed previous BoW representation-based methods. Sebastian et al. [20] explored the classification accuracy of multiple embedding types (BOW, 2-grams, 3-grams, linguistics, Word2Vec, Paragraph2vec, extended 2-grams, and extended 3-grams) by training a logistic regression classifier. Their results elaborated that Word2Vec and Extended 2-grams performed better than other embedding types. Sentence embedding-based hate speech detection methods have also been proposed in the literature [21]. Alorainy et al. [22] explored the performance of a wide range of classifiers through multiple word embedding techniques to detect hate speech (n-gram, comment embedding, Word2Vec, LSTM, paragraph2vec). By analyzing words and phrases using multiple features, they explored the context differences between hateful and non-hateful texts. Deep neural network models are also proposed for hate speech classification [23] [24] [25]. The performance of these models is much higher than lexicon-based methods and embedding-based approaches. In [24], authors used an ensemble model which integrates a series of features from the abusive text and user behaviours and in [26], an approach that utilized a blend of Glove and FastText word embeddings as input characteristics along with a BiGRU model, aiming to detect hate speech originating from social media platforms is introduced.

After the discovery of pre-trained models such as BERT [27], a significant number of hate-speech detection methodologies adopted them as embedding as well as some research used transformer model directly for the classification [28], [29]. Goran et al. [30] conducted a series of experiments by implementing monolingual and multilingual supported transformer-based models on hate speech detection. The performance of these models is substantially improved over the baseline approaches. Moreover, traditional methods such as TF-IDF and BoW are compared with ML models in [31]. In our experimental setup, we employed representation from different architectures to capture distinct aspects of word semantics and context. Firstly, we incorporated the TF-IDF approach, as a count-based vector space model. Next, for the non-context-based vector space model, we opted for GloVe [32], and lastly, to contrast traditional text-embedding methodologies with more advanced approaches, we selected BERT [27], as a context-based vector space model. Subsequently, we construct a multi-task framework using a shared BERT-based transformer to enhance the overall performance.

C. HATE SPEECH DETECTION USING SENTIMENT AND EMOTION FEATURES

In addition to developing a single classifier for hate speech detection, multi-model approaches are also proposed in the literature. A limited number of works used both emotions and sentiment to develop multi-model systems. For instance, a multi-modal approach has been proposed in [36], where authors tried to adapt topics (racism, xenophobia, sexism, misogyny) and hate speech targets. The emotions were encoded in hate lexicons and the experimental results showed that the emotion feature is useful when detecting hate speech. Niloofar et al. [10] introduced a single deep neural network approach introducing an emotion-aware attention model. Their emotion model is developed with emojis, which can identify how relevant an emoji is to a given text. Ricardo et al. [9] examined the possibility of integrating emotions into the available dataset and then, classifying hate speech using the newly created dataset. Their training dataset is prepared as a vector-based corpus, and each vector contains a few emotional scores, emotional intensities and sentiment polarity scores. The authors used a few ML classifiers and concluded that the emotion features help to increase the performance of the hate speech detection model. FADOHS is a framework implemented by Axel et al. [8] to detect hate speech using sentiment and emotions. Their experimental results have shown that FADOHS surpasses the performance of the previously proposed hate speech detection models. Also a novel framework that utilizes graph, sentiment, and emotion analysis techniques to automatically detect Facebook pages that promote hate speech in comment sections concerning sensitive topics, is introduced in [37]. Aneri et al. [38] introduced a new multimodel approach to filter hateful content using emotion as a feature. Their deep model used auditory features representing emotions and semantic features and proved that using emotion attributes significantly improves the performance of the hate detection models compared with text-based models. Moreover, Mohammad et al. [39] focused on sentiment analysis of tweets in the context of hate speech detection in the Urdu language, addressing challenges like skewed classes, high-dimensional feature vectors, and sparse data. Despite the utilization of emotions in hate speech detection, there remains a deficiency in adapting fine-grained emotions to address hate speech comprehensively. Furthermore, the majority of state-of-the-art emotion features are derived from hate speech detection methods that primarily concentrate on a single hate category. In this study, we address this gap by targeting multiple implicit hate speech categories, aiming to investigate the impact of fine-grained

TABLE 1: Class distribution of datasets used in STL and MTL setup. For the GoEmotion dataset, the classes are presented in Hierarchical Grouping (Neutral is excluded) and sub-categories of each class are described in Table 2

	Dataset	Task	Size	Classes		
				Hate		non-Hate
D1	Latent Hatred [5]	Hate Speech	20,391	7,100		13,291
D2	White Supremacy Hate [33]		10,703	1,196		9,507
D3	Offensive Language hate [18]		24,783	20,620		4,163
				Positive	Negative	Ambiguous
D4	GoEmotion [34]	Emotion Detection	58,009	21,733	12,996	6,668
				Positive	Negative	Neutral
D5	SemEval 2017 task 4-A [35]	Sentiment Analysis	50,356	19,913	7,850	22,593

emotions.

D. HATE SPEECH DETECTION USING MULTI-TASK LEARNING APPROACH

As hate speech detection can be a challenging task regards performance, MTL is a beneficial method by joint training on multiple tasks, so the model can learn more comprehensive representations and improve its ability to accurately detect hate speech. It offers several advantages for hate speech detection, including improved performance, data efficiency, robustness, and bias mitigation. For instance, Akhtar et al. [40], presented a context-level inter-modal attention framework using deep MTL for sentiment and emotion analysis. Their framework has multi-modal input, such as multi-modal text, and acoustic and visual frames of a video and offers an improvement over the single-task framework. In [41], authors used the same related tasks to detect hate speech in Spanish tweets. They proposed a multi-task model to integrate sentiment and emotion knowledge and evaluated their model on two hate speech datasets containing Spanish tweets. Firstly, they trained and evaluated the model on the corresponding datasets with the Transformer-based BETO model to obtain a baseline for STL and then used a multi-task setting to compare the model performance with the baseline. Also in [42], authors proposed multi-tasking models to evaluate the impact of emotions, sentiment and target classification. They used Ekman's emotion models and binary sentiments of content. The authors concluded that the use of emotions, sentiment and target classification via MTL improved the performance of hate speech detection. However, their analyses focused only on a single hate speech category. Md Rabiul et al. [43] proposed another MTL model called AngryBERT to classify hate speech features. AngryBERT is based on sentiment features and a target identification approach. They evaluated the performance of the model compared to several state-of-the-art hate speech classifier-based models including; DeepHate, CNN, LSTM, CNN-GRU etc. The AngryBERT model, which is based on sentiment features, outperformed the other baseline models. Moreover, a deep MTL framework is proposed by [44] using related tasks to hate speech such as offensive language identification, racism detection and sexism detection, while in [45], authors proposed the first joint model using emotion and abusive language detection in a multi-task setup.

To accomplish our objective of the impact of fine-grained emotion in detecting implicit hate speech, we employ a MTL paradigm as well as STL and incorporate relevant knowledge to enhance the performance of hate speech detection on the target implicit dataset.

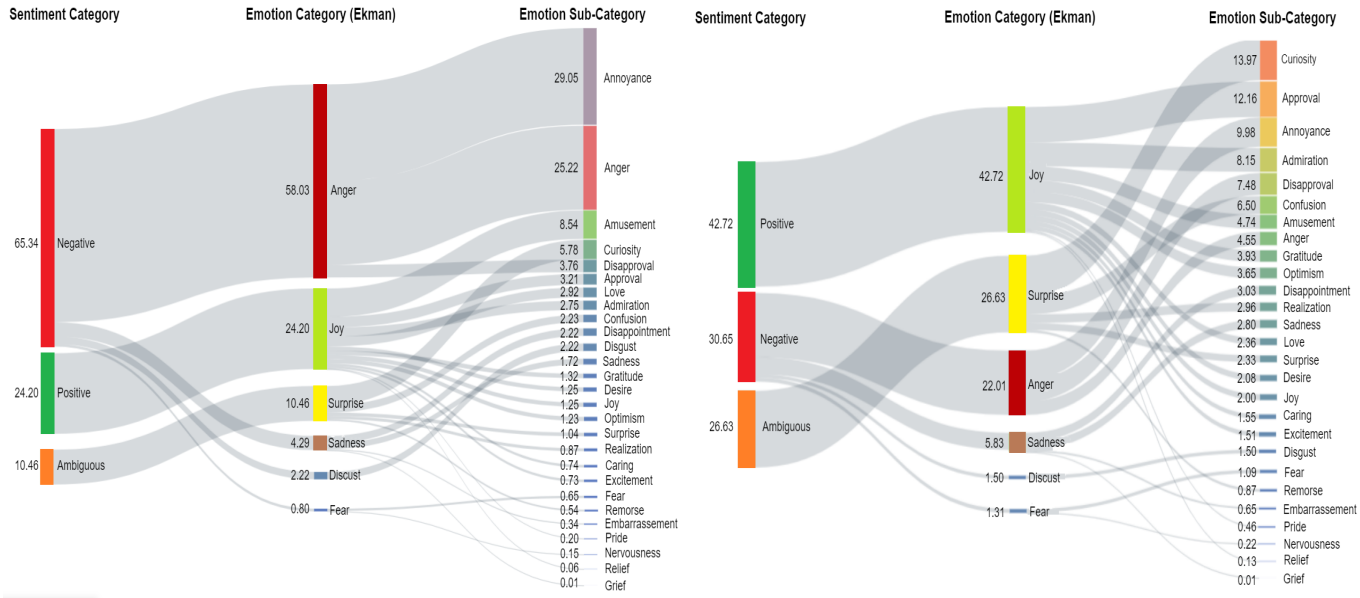
III. DATASETS

As shown in Table 1, our analysis and experiments rely on the utilization of five distinct labelled datasets, including three hate speech datasets (D1, D2, D3), one fine-grained emotion dataset (D4) and one sentiment dataset (D5). The experiments and the multi-task hate speech detection model evaluation are heavily based on the Latent Hatred dataset (D1) with the support of GoEmotion dataset (D4) and SemEval dataset (D5), while the other two hate speech datasets (D2 and D3) are used for the emotion analysis of hateful content to have a wider and more balanced range of data.

Latent Hatred dataset (D1): This dataset [5] contains about 20K English tweets with not-hate and implicit hate categories. Implicit hate tweets divide into six main categories with the following distributions: Grievance (24.2%), Incitement (20.0%), Inferiority (13.6%), Irony (12.6%), Stereotypes (17.9%), Threats (10.5%) plus Other (1.2%). The primary purpose of using this dataset is to analyze emotions in hate and non-hate content and perform a more detailed comparison of emotions used in each category to explore tricks used in tweets such as showing threats and abuse, which is more abstract compared to explicit hate.

White Supremacy Hate (D2): This dataset [33] contains 10K sentences extracted from Stormfront, which are manually labelled into four main labels: hate (11.29%), noHate (86.09%), relation (1.69%) and skip (0.93%). We used first two labels for our emotion analysis in hate and non-hate data.

Offensive Language hate (D3): This dataset [18] contains approximately 24K tweets with three main categories: hate speech (5.8%), offensive language (77.4%), and those with neither (16.8%). In their results, tweets related to racist and homophobic classified as hate speech, and sexist tweets are generally classified as offensive. However, as explained in Subsection IV-B, we consider both offensive language and hate speech as hateful content in our analysis.



(a) hateful content (b) non-hateful content

FIGURE 2: Emotion distribution (in percentage) in hateful and non-hateful contents.

GoEmotion dataset (D4): A human-annotated corpus consists of 58k Reddit English-language comments labelled with 27 classes including 12 positives, 11 negatives, and 4 ambiguous emotion categories described in Table 2. This dataset splits into a train set with a size of 43,410, and a test and validation set with a size of 5,427 and 5,426 respectively [34]. We used this dataset in our MTL experiment for the task of fine-grained emotion detection.

SemEval 2017 dataset (D5): This dataset [35] was released for Task 4 (subtask A) in SemEval 2017 which is a message polarity classification with positive, negative, or neutral sentiment. We used English examples of this dataset for the sentiment classification task in our MTL setup which contains around 50K messages including tweets on a range of topics, such as a mixture of entities, products and events.

IV. ANALYSIS OF FINE-GRAINED EMOTIONS IN HATE SPEECH

Analyzing fine-grained emotions used in both non-hateful and hateful content provides a comprehensive understanding of the emotional landscape, enabling us to differentiate between genuine expression and harmful hate speech. Therefore, it is necessary to understand deeply how the emotional difference between those two categories goes beyond binary polarity to refined emotions such as anger, fear, disgust, or sadness. In this section, we comprehensively examine the emotions employed in both non-hateful and hateful content in order to uncover the emotion distribution and patterns utilized in the dissemination of hatred within online content.

TABLE 2: Mapping of Original GoEmotions with Ekman and Hierarchical Grouping (Sentiment).

Sentiment	Ekman	Original GoEmotions
Negative	Anger	Annoyance, Anger, Disapproval
	Sadness	Disappointment, Sadness, Embarrassment, Grief
	Disgust	Disgust
	Fear	Fear, Nervousness
Positive	Joy	Approval, Admiration, Amusement, Gratitude, Optimism, Love, Desire, Joy, Caring, Excitement, Pride, Relief
Ambiguous	Surprise	Curiosity, Confusion, Realization, Surprise, Remorse

A. FINE-GRAINED EMOTION MODEL

To analyze the fine-grained emotions, we adopt the GoEmotions [34] taxonomy which is specifically developed for text data. It expands on Ekman's basic emotions and incorporates a broader range of emotional categories including a multitude of positive, negative, and ambiguous emotion categories. This taxonomy has been utilized in recent research as it provides a more comprehensive framework for understanding and analyzing emotions expressed in text. However, the majority of previous works used GoEmotions with Ekman's categories. Our objective is to leverage the detailed emotional categories introduced in the GoEmotion model to investigate relevant features from hatred content. By utilizing fine-grained emotional analysis, we seek to uncover valuable insights and patterns associated with expressions of hatred.

GoEmotions model incorporated with the following three taxonomies implemented with Huggingface Transformers[1].

[1]https://github.com/monologg/GoEmotions-pytorch

TABLE 3: Implicit hate categories definition proposed by [5] with examples and emotion model output

Category	Definition	Example tweet	Model output
Incitement	flaunting in-group unity to the implicit promotion of known hate groups and ideologies	this is an inherently white thing to do. keep up the good work	Admiration
Inferiority	using toxic language to show inferiority over a group	I really hate neocons like but not black people. it is not your fault that you are black	Anger
Irony	using sarcasm, humour, and satire to attack a group or individual	would you like fries with your mass amnesty and depressed wages?	Curiosity
Stereotypical	using negative attributes such as terrorism and metaphorical language to associate a group	You're Middle Eastern? Really? It's so good to know there are actually decent Middle Eastern people out there.	Admiration, Curiosity, Surprise
Threatening	attacking a group or individual with targeting pain, injury, damage, and violation	we need to stop the flow of immigration in our country! all must be vetted! just obey the laws! deport criminals!	Anger
White grievance	showing frustration over a minority group	not a good time to be an old white guy	Disapproval

- **Original GoEmotions** (27 categories + neutral):
  admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise + neutral.
- **Hierarchical Grouping (Sentiment)** (3 categories + neutral):
  positive, negative, ambiguous + neutral.
- **Ekman** (6 categories+ neutral):
  anger, disgust, fear, joy, sadness, surprise + neutral.

Table 2 shows how categories of each taxonomy are correlated. Our experiments are based on the Original GoEmotions categories excluding the neutral category.

B. EMOTIONAL CONTRASTS IN HATE VS. NON-HATE CONTENT

This section mainly focuses on analyzing how emotional values are varied from hateful content to non-hateful content. This further helps us to gain insights into the types of emotional tendencies associate with hate speech in general and that certainly leads to tracking and monitoring implicit hate speech with certain emotional values. As stated in Section III, D1, D2, and D3 hate speech datasets have their own hate categories. In our analysis of emotions, the term 'hateful content' refers to the texts that fall under six hate categories: Grievance, Incitement, Inferiority, Irony, Stereotypes, and Threats, as identified in D1. Additionally, it includes texts categorized under the Hate category in D2, as well as those classified under hate speech and offensive language categories in D3.

To evaluate and gain a visual understanding of the emotion value variations in both hateful vs non-hateful content, we generated Figure 2 utilizing datasets. For both content, Figure 2 maps each sentiment level to the correlated Ekman level, and from Ekman level to its fine-grained emotion sub-categories. As illustrated in Figure 2.b, non-hateful content includes more positive sentiments than negative sentiment, whereas hateful content (Figure 2.a) includes higher negative content. In the analysis of non-hateful content, there is a relatively small difference between the occurrence of negative sentiments and ambiguous sentiments, with percentages of 30.65% and 26.63% respectively. However, in hateful content, the majority of sentiment portion is accounted for the negative label with over 65% of the total distribution, which is nearly doubled compared to non-hateful. Overall, hateful content tends to have more negative sentiments, and non-hateful content exhibits a higher number of positive sentiments than other sentiment categories.

Ekman emotion categories exhibit that the *Anger* comprises the highest proportion of hateful contents; it is in the third place in non-hateful contents. In contrast, the *Joy* category consists of much non-hateful content followed by *Surprise*. The other categories were much smaller in both contents. The conclusion here is that based on the Ekman emotion model, hateful content tends to have the highest number of anger content under the negative sentiment class.

Lastly, the fine-grained emotions are depicted in both Figure 2.a and Figure 2.b via the emotion sub-category column. A deeper look at the emotion sub-categories shows that emotions are less distributed in hateful content. In hateful content, the greatest portions of emotions are allocated with *Annoyance* and *Anger*, in total about 54%, under the negative sentiment category. On the other hand, fine-grained emotions within non-hateful content are widely distributed across various emotional categories, but a significant portion of them belong to the ambiguous sentiment category (*Curiosity* - 14%, *Approval* - 12%.)

In summary, hateful content exhibits a strong emphasis on negative sentiment and Anger Ekman level. Additionally, it primarily encompasses two types of fine-grained emotions: *Annoyance* and *Anger*. These findings are important in the relevant feature selection for hate speech classification.

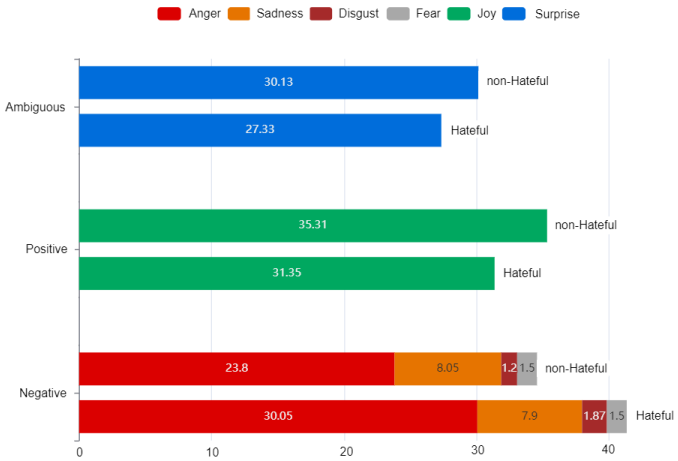


FIGURE 3: **Sentiment** and **Ekman** emotion distribution of hateful and non-hateful tweets of implicit dataset (in percentage)

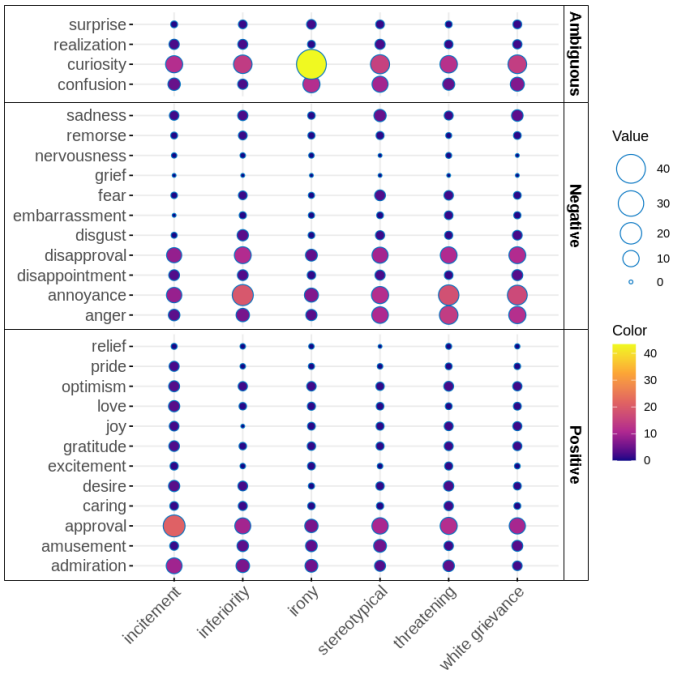


FIGURE 4: Fine-grained emotion distribution of six implicit hate categories. Circles and colours indicate the normalized emotional value in percentage.

Based on the analysis, the emphasis on negative sentiment can be beneficial for improving the model performances when classifying hateful content, and in addition *Annoyance* and *Anger* tend to have the most impact as additional features to enhance the model performances. Apart from that, the fusion of emotion and sentiment features can boost classification performances.

C. EMOTION ANALYSIS ON IMPLICIT HATE CATEGORIES

In Section IV-B, the analysis is mainly focused on hate speech in general, but this section focuses on the analysis of fine-grained emotions within implicit hatred content. We used the Latent Hatred dataset (D1) for this analysis targeting six hate categories: incitement, inferiority, irony, stereotypical, threatening, and white grievance. Table 3 shows six main hate categories with their definition proposed by [5] and samples tweets and their respective emotion classed by the GoEmotion model. We analyzed sentiment and related Ekman emotions on hateful and non-hateful content in the D1. After that, emotion frequencies in each category were analyzed. Although implicit hate detection is more challenging than explicit hate, we found different emotion patterns in each implicit hate category.

Figure 3 depicts the sentiment and Ekman emotion distribution of both hateful and non-hateful content in D1. Although the trend of emotion distribution in both groups follows a similar pattern as presented in the previous section, it is less distributed in terms of sentiment category in implicit hate. The main reason is the indirect language usage in implicit hate instead of the direct presentation of hatred in explicit hate.

Figure 4 provides a more detailed depiction of how fine-grained emotions and sentiment are utilized within implicit hatred across six different hate categories. As shown in Figure 4 in each implicit hate category for positive, negative and ambiguous sentiment labels, the most intense emotion categories are *Approval*, *Annoyance* and *Curiosity*, while the least frequent emotion categories are *Relief*, *Grief* and *Surprise*, respectively.

- In the *Incitement* hate category, the number of positive emotions is higher than the remaining sentiment categories in which *Approval* emotion label exhibits about 30% of the total emotions in this hate category. The main reason behind this occurrence can be explained based on the definition proposed for this category in Table 3 where incitements are used in ethnic or racial hatred, and users flaunt in-group unity and power, which leads to more positive emotions in their words. Generally, as reported by [5], *Incitement* category is even the most confusing category in both non-hate and hate among all implicit hate categories.
- In *Inferiority*, *Threatening* and *White grievance* hate categories, fine-grained emotion distribution is quite similar, where from negative emotions, the most intense emotion label is *Annoyance* which is approximately 20%, and *Anger* consists around 7% for *Inferiority* and 13% for other two categories in overall. These are the emotions used for spreading inferiority, frustration and threats in hateful content.
- In *Irony* hate category, the trend shows that using ambiguous emotions is more common. The *Curiosity* emotion label, with more than 40%, has the largest share of all emotions for ironic hate content, which refers to using sarcasm, humour and satire to spread hatred while attacking a group or an individual.
- The *Stereotypical* category exhibits more diverse emotion

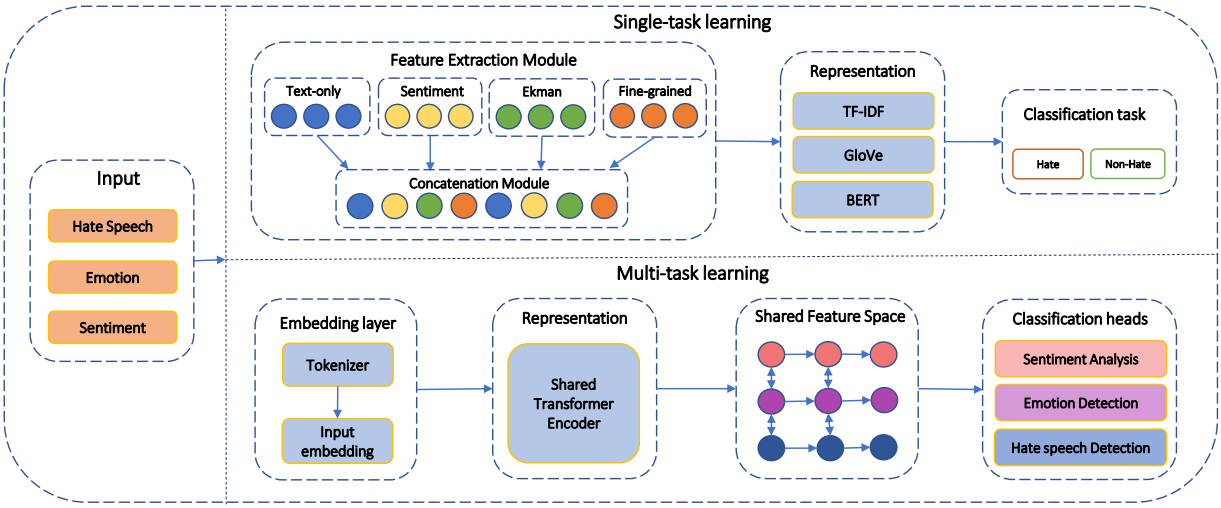


FIGURE 5: Experimental Setup for the implicit hate speech classification a) STL: a multi-modal model which uses different feature combinations, and b) MTL: a shared transformer encoded to train on related tasks

distribution than other implicit hate categories where frequent emotions used in this category are *Curiosity, Annoyance, Anger, Approval, Confusion, and Disapproval* with distribution range between 10-15% which is more balanced compared to other implicit hate categories. The main reason is that *Stereotypical* hate content uses metaphorical language that is intended to be taken as a metaphor to represent, suggest or compare something that leads to more emotion distribution based on the associated group of people.

To summarize, the fine-grained emotion analysis elucidates the most intense emotion labels for six hate categories. As seen in Figure 4, the most intense emotions are almost equal in all the hate categories but exhibit different intensity levels with the emotional score. Moreover, there is a considerable variation in the other categories of emotions in each hate category, which may provide helpful information for hate speech detection models and semantic features. Overall, leveraging the insights gained from the fine-grained emotion analysis, such as emotion intensity levels, and variations of emotions in different emotion categories, can enhance the effectiveness of implicit hate speech classification models.

In the next section, we aim to use these findings and analyze how fine-grained emotion and sentiment features can be used to classify implicit hateful content.

V. IMPLICIT HATE SPEECH CLASSIFICATION

This section highlights our methodology for classifying implicit hate speech and explores the influence of different levels of features in this task. We have developed a system, illustrated in Figure 5, which comprises two primary components.

The first component, STL, focuses on evaluating the impact of diverse feature levels. It incorporates emotion, sentiment, and the baseline text features [5] along with the concatenation of all features for classifying implicit hate speech to find the optimal performance. Our objective is to establish a new baseline performance for our STL system, which can be used for subsequent comparison with the MTL approach.

In the MTL component, we aim to demonstrate the effectiveness of employing related tasks within an MTL setup to enhance the performance and better generalization of implicit hate speech detection in comparison to the baseline achieved in the STL stage. By leveraging the advantages of MTL, each task extract related task features and also features that are relevant to all tasks based on the relationships between the tasks, and we seek to further improve the accuracy of implicit hate detection.

A. EXPERIMENTAL SETUP

The system depicted in Figure 5 contains several stages:

Input: The input for our system consists of three datasets (D1, D4 and D5) that are utilized for learning features related to hate speech detection, emotion detection, and sentiment analysis tasks.

Feature Extraction Module: We analyze the feature importance by doing the ablation study; we conduct an ablation study that compares the model's performance across five feature levels: i) text-only features [5]; ii) sentiment features; iii) Ekman-level emotion features; iv) fine-grained level emotion features; v) combination of sentiment and emotion features with text-only features.

We acquire the emotion features as follows: for each sample in the dataset, we run the emotion model mentioned in Section IV-A and generate a fine-grained emotion vector $v \in \mathbb{R}^n$ where n is the number of emotion categories. Here, we use the original GoEmotion taxonomy and take $n = 28$ for the emotion features in both binary and multiclass

classification. The emotion vector v contains fine-grained emotion information about the text.

As shown in Figure 5, after acquiring text features and emotion or sentiment features, we pass these features along with concatenated features (Concatenation Module) to models. The concatenation module performs simple splicing of the embeddings, for which several methods can be employed, including simple vector concatenation, gating, attention weighting etc. In our experiment, we choose the simple concatenation method with the best performance.

Single-task Setup: STL focuses on training a model to perform a specific task using a dedicated dataset to optimize performance on a single task. In this paper, we chose implicit hate speech detection using D1. Before the representation level, an appropriate tokenizer will be selected to transform textual data into a numerical format and prepare the input embedding. We further employ emotion features together with textual features for the detection of hate speech. The different levels of emotion features and the textual features are essentially different modalities, so we propose a feature fusion module to acquire a multi-modal representation of diverse features. For our baseline model, we perform a simple concatenation of the features to obtain the multi-modal features.

In order to establish a baseline for the evaluation of the system in single-task architecture, we adopt three commonly used text encoding models to acquire text features; sparse representation **(TF-IDF)**, a widely used technique that assigns weights to words in a document based on their frequency and importance within the document and across a collection of documents. TF-IDF represents each document as a vector in a high-dimensional space, where each dimension corresponds to a unique word in the document corpus. Next is pre-trained word embedding (**GloVe** [32]), an unsupervised learning algorithm that learns word embeddings, which are dense vector representations of words in a continuous space. It leverages the co-occurrence statistics of words across a large corpus to capture their semantic relationships. GloVe represents words as dense vectors of fixed dimensions, unlike TF-IDF, which represents each word as a sparse vector. Finally, a well-known **BERT** embedding [27], a model that utilizes a transformer architecture to capture the contextual information of words and sentences and is pre-trained on a large corpus containing a wide range of tasks. Unlike traditional models that process words in a left-to-right or right-to-left manner, BERT employs a bidirectional approach. The pre-training allows BERT to learn a general understanding of language and context. Fine-tuning BERT on specific downstream tasks enables it to provide highly accurate predictions, which in our work, it can obtain a better result for implicit hate detection. In our experiment, we use fully connected layers as the fusion module followed by another fully connected layer as the classification head. The layer number is set to one, with the input size equaling the concatenated feature number.

Multi-task Setup: MTL involves training a model to simultaneously learn multiple related tasks using a shared dataset to capture shared information and leverage it across multiple tasks, benefiting from the similarities and correlations between tasks [46], [47]. In our experiments, we use two auxiliary tasks for hate speech detection, namely, emotion detection task and sentiment analysis task. For the hate speech task, we only use the Latent Hatred dataset D1 in the learning setup. In multi-task setup, a crucial step is parameter sharing of hidden layers where we use hard-parameter sharing, which generally comprises a shared encoder where we used BERT as a transformer encoder that branches out into each task head.

B. PROPOSED MULTI-TASK LEARNING ALGORITHM

To use the MTL setup for our experiments, we propose an algorithm, to process input data, learn task-specific information, compute losses, and update its parameters and ultimately aiming to improve the performance of each task in the MTL setup.

Considering the given k tasks (which are sentiment analysis, emotion and hate detection), let $S \in \mathbb{R}^{d \times r}$ denote the shared transformer module used in our MTL setup and $P_i \in \mathbb{R}^r$ for task i, where r is the dimension of the shared module and d is data dimension. Moreover, for each task i, X_i represents covariates and y_i shows its label.

The main goal is to minimize the total loss over the Shared encoder module and each task.

$$f(P_1, P_2, \ldots, P_k; S) = \sum_{i=1}^{k} Lg(X_i S) P_i, y_i, \qquad (1)$$

Algorithm 1 Multi-task Learning with Transformer shared encoder. Coefficients in loss L are hyperparameters that determine the relative importance assigned to each task's loss.

Input: H - Hate speech annotations, E - Emotion annotations, S - Sentiment level annotations, embedding model, encoder model
  $h \leftarrow$ EMBEDDING(H)
  $h_e \leftarrow$ EMBEDDING(E)
  $h_s \leftarrow$ EMBEDDING(S)
  $h' \leftarrow$ ENCODER(h, h_e, h_s)
for *epoch* in range(*num_epochs*) **do**
    Compute task-specific scores
      $\hat{p}_e, \hat{p}_s, \hat{p} \leftarrow M(h'; \theta)$
    Compute task-specific losses
      $L_{sentiment} \leftarrow \mathrm{L}(\hat{p}_s, y_s)$
      $L_{emotion} \leftarrow \mathrm{L}(\hat{p}_e, y_e)$
      $L_{hate} \leftarrow \mathrm{L}(\hat{p}, y)$
    Compute multitask losses
      $L = \beta L_{sentiment} + \lambda L_{emotion} + \gamma L_{hate}$
    Update the multitask model
      $\theta \leftarrow \theta - \alpha \nabla L(\theta)$
end for

Algorithm 1 presents the pseudocode for our MTL approach. To provide the necessary information and resources

TABLE 4: Experimental results of the STL models for the binary classification of implicit hate speech. F1 scores are reported in the macro average.

Feature level	**TF-IDF**				**GloVe**				**BERT**			
	Precision	Recall	macro-F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc
Text-only (Latent Hatred) [5]	59.5	68.8	63.9	71.6	56.5	65.3	60.6	69.0	72.1	66.0	68.9	**78.3**
Sentiment	63.6	67.3	64.4	71.5	59.0	67.6	63.0	70.7	72.4	73.5	72.8	75.4
Ekman level	63.6	69.0	66.2	72.4	59.0	67.4	62.9	70.6	72.2	73.6	72.9	76.4
Fine-grained Emotion	**64.7**	67.0	65.8	71.4	**60.5**	67.1	63.6	70.9	72.7	**74.3**	**73.5**	77.2
All features	64.4	**69.1**	**66.7**	**72.6**	60.3	**67.9**	**63.9**	**71.8**	**72.9**	74.0	73.4	75.9

to perform subsequent computations and tasks, we consider hate speech annotations (H), emotion annotations (E), and sentiment level annotations (S), as well as the embedding model and encoder model in the input phase. After that, the algorithm applies an embedding model to convert the input annotations (H, E, and S) into numerical representations (h, h_e, and h_s) so that can be processed by the subsequent stages. To capture the joint information from multiple annotation types and improve the performance of subsequent tasks, an encoder model is used to combine the embeddings of hate speech (h), emotion (h_e), and sentiment level (h_s) into a single representation (h'). After the model learns and improves its performance, it computes task-specific scores to generate predictions or scores related to sentiment, emotion, and hate speech, which are the specific tasks the algorithm aims to address. Next, the model computes task-specific losses ($L_{sentiment}$, $L_{emotion}$, L_{hate}) by comparing the predicted scores with the ground truth labels (y_s, y_e, y) to quantify the discrepancy between the predicted and desired outputs, allowing the model to optimize and improve its performance for each task. In the last step, the model combines the individual task losses using weighted coefficients (β, λ, γ) to compute the multitask loss (L) and performs parameter updates to reach the main goal. This phase is significant as it balances the importance of each task and allows the model to jointly optimize across multiple objectives.

In the following sections, we describe the details of the experiments along with the results to explore how different levels of emotions and sentimental features influence performance on implicit hate speech detection.

C. EXPERIMENT TYPES

We designed two types of experiments to support our objective:

i) We conduct an experiment with several STL methods for implicit hate speech classification. In this experiment, we evaluate different text representations using sentiment, Ekman-level emotion and fine-grained emotion features along with text-only features. Due to the different modalities of text features and emotion features, we design a multi-modal model to concatenate text features with emotion and sentiment features and selected the best-performed STL method for further comparisons.

ii) We evaluate the Latent Hatred dataset D1 in an MTL approach with related tasks to check the performance in comparison with our STL baseline model. In order to reach the optimum approach, we consider three configurations. In all of them, the model trains and evaluates the corresponding dataset (D1) along with 1) SemEval 2017 dataset (D5) for the sentiment analysis task (MTL_sentiment) 2) GoEmotion dataset (D4) for the fine-grained emotion detection task (MTL_FG-emotions) 3) All mentioned datasets (D1, D4 and D5) combined tasks (MTL_all).

TABLE 5: Experimental result of the Multi-task learning models. F1 scores are reported in Macro-average.

Model	Precision	Recall	F1	Accuracy
Text-only (Latent Hatred) [5]	72.1	66.0	68.9	78.3
STL	72.7	74.3	73.5	77.2
MTL_sentiment	73.1	75.2	74.1	77.8
MTL_FG-Emotion	**76.9**	78.8	**77.8**	**80.2**
MTL_All	75.5	**79.3**	77.4	79.9

During training, we test the model and select hyperparameters using the validation split, and we report the final performance on a held-out test split. All the metrics are tested on an average of five runs. The model was trained for a total of 5 epochs using a batch size of 32. We initially fine-tuned the model with a learning rate set to 5e-6 for 3 epochs, which was later adjusted to 4e-5. The AdamW optimizer with a weight decay rate of 0.01 was employed to optimize the training process. Additionally, a maximum sequence length of 128 was enforced to handle input data.

For the experiments, we used PyTorch high-performance deep learning library [48] run on a Tesla-V100 32 GB GPU with 64 GB of RAM.

D. EXPERIMENT RESULTS

In this section, we show the result of our methodology evaluation. We used the macro-average of prevalent metrics such as Precision (P), Recall (R) and F1-score (F1). The results are described below:

1) Experiment 1

This experiment explores how fine-grained emotion features perform to distinguish between implicit hate speech and non-hateful speech and compare the performance with the broadly-adopted sentimental features. In order to compare and analyze the feature importance of binary sentimental features and the fine-grained emotion features, we perform ablation experiments with different text encoding models

TABLE 6: STL and MTL example on implicit hate dataset. The red color represents error in classification.

	Text	Actual(Category)	STL	MTL		
				Hate	Sentiment	Emotion
T1	nice to know that tsunami didn't dampen their energy to stay in the 7th century	hate (Irony)	non-hate	non-hate	positive	admiration, approval
T2	it's wonderful if these white women will wake up to anti-white racism. but some whites love to be punished for being white.	hate (White grievance)	non-hate	hate	neutral	admiration
T3	we obviously need a nation for white people. and the antifa white race traitors can live with the black people.	hate (Incitement)	non-hate	hate	neutral	approval
T4	sounds violent to me. white people advocating for a white homeland. the horror of it	non-hate	hate	hate	negative	fear
T5	are you trying to demoralize us? huwites have the power to take the white house alone.	non-hate	hate	non-hate	negative	curiosity
T6	white flight didn't work because being afraid of black people is stupid.	non-hate	hate	non-hate	negative	anger, annoyance

on the task of binary classification of implicit hateful and non-hateful content. We used the Pysentimentio toolkit [2] for the sentimental analysis, which is an open-source library supported for multilingual sentiment analysis. A 3-dimension sentiment vector containing binary sentiment information is generated for each sample in the implicit hate speech dataset.

As shown in the results in Table 4, the emotion features demonstrate good discrimination ability for implicit hate speech and improve the performance of the baseline text-only model. Using emotion and sentiment knowledge as additional features achieve improved classification performances. We observe that binary-polarity sentiment features also help the classification and improve the classifier's performance by a small margin (around 2% on average in macro F1 score) compared to text-only features as reported in [5]. In contrast, finer-grained emotion features consistently outperform sentiment features and bring around 4% improvement in the F1 score, showing a better ability to capture the latent sentimental information in the text, which helps to identify implicit hate speech, and combining sentiment features and emotion features brings slight improvement over emotion features alone. Moreover, in order to compare fine-grained and coarse-grained emotions on system performance, we used Ekman-level emotion which contains 6 main categories. The results show that fine-grained features perform slightly better in F1 in all three encoders.

Based on the results in Table 4, the best-performed STL model uses fine-grained emotion features and BERT embedding. Compared to the binary classification result using text-only features reported in Latent Hatred [5], we achieved better performance in precision, recall and F1-score. Therefore, we consider this model as our STL baseline to compare with the multi-task learning setting.

2) Experiment 2

We further experiment with a multi-task model incorporating all feature levels used in the STL models. The experiment results are presented in Table 5, showcasing the performance of each configuration on the implicit hate speech task. Notably, all multi-task models outperformed the STL baseline across all evaluation metrics. Particularly noteworthy was the improvement in accuracy by approximately 2 percent in our MTL setup, which was the only metric we fell behind in our STL method compared to text-only (Latent Hatred).

Moreover, despite expecting significant enhancements in the sentiment setup, we only observed marginal growth in the F1 score in MTL_Sentiment. This could be attributed to the fact that implicit hate speech often employs indirect language, such as sarcasm and humour, making it challenging for the system to classify the polarity correctly. As a result, the system may frequently misclassify the labels in such cases. These findings underscore the efficacy of incorporating fine-grained emotion features in the MTL setup, which significantly aids in improving the performance of implicit hate speech detection.

E. ERROR ANALYSIS

We perform an error analysis on the performance of our proposed STL and MTL models. Table 6 provides a few example cases to compare how single-task and multi-task models classified implicit hate. More specifically, it contains 3 false positives and 3 false negatives regarding the single task baseline. For those examples where both STL and MTL models failed to predict actual labels, it can be observed that even sentiment and emotion tasks failed to help the system, particularly in T1 and T4; we can derive by the predicted labels in sentiment and emotion task in MTL setting that the system would classify as non-hate and hate respectively while the actual label is opposite due to the nature of implicit hate. In the two false negative examples (T2 and T3), STL predicted non-hate while MTL classified correctly despite the system could not recognize any polarity for the text. On the contrary, regarding false positive examples, at a glance, the text might indicate hate speech; nonetheless, the actual label is non-hate and MTL could classify correctly. Presumably, in T7, although the detected sentiment is negative, the fine-grained emotion helps the hate speech detection task as it classified the text as "curiosity" which is a sub-category of

[2]https://github.com/pysentimiento/pysentimiento

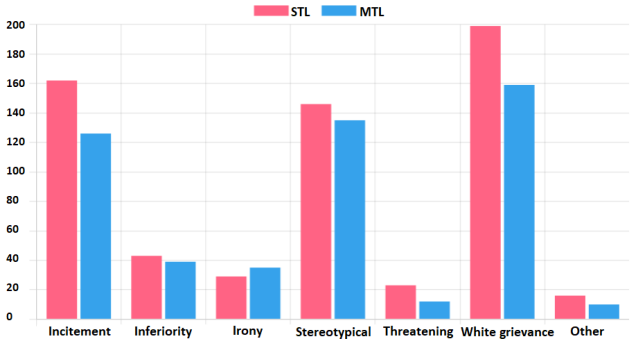
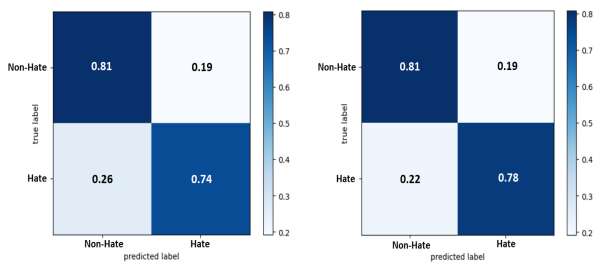


FIGURE 6: Comparing STL and MTL in number of false negatives for each implicit hate category test set

ambiguous class in the GoEmotion model.

To have a more comprehensive look at the errors of each model, Figure 6 compares each model on the number of wrong predictions for each implicit hate class. As can be seen, MTL has the best performance on "Incitement" and "White grievance" where the number of wrong predictions decreased sharply from 162 to 126 and 199 to 159 cases respectively. As a result, learning on related tasks has shown its greatest impact in reducing these two classes, especially "Incitement" which is the most confusing category to detect among all implicit hate categories. Similarly, in other categories, MTL reduced errors slightly as expected, except for "Irony" where MTL failed to outperform, presumably, since in this category, texts usually contain sarcasm, humour, etc, MTL setup could not be useful compared to STL.

In addition, confusion matrices of STL and MTL are demonstrated in Figure 7. A comparison of two matrices reveals that the MTL setup successfully improves implicit hate speech detection by increasing the number of true positives as well as decreasing false negatives. On the other hand, no considerable difference is observed regarding the non-hate class in both models.



(a) Single-task learning (b) Multi-task learning

FIGURE 7: Confusion matrix on implicit hate data

VI. CONCLUSION AND FUTURE WORK

Many hate speech detection models were proposed in the literature, but a few of them utilize additional features such as fine-grained emotions, especially in domain-specific hate speech. This work explores the influence of emotion as a feature to build a model to detect implicit hate speech by proposing a MTL model based on related tasks such as sentiment analysis and emotion detection. First, we analyzed fine-grained emotion distribution in both non-hateful and hateful content and explored that hateful content has more negative sentiment while non-hateful content includes more positive sentiments. We further analyzed the variations of fine-grained emotions in implicit hate to explore potential emotional patterns. The results show that *Anger* and *Annoyance* in the negative sentiment category tend to spread more hateful content. Moreover, we explore that implicit hate speech has different fine-grained emotion intensities. Then, we implemented a set of word embedding-based classifiers (STL models) to identify whether the emotion features are useful in implicit hate speech classification and compared their performance with sentiment feature-based classifiers and other baseline models. Finally, we experimented with multi-task models incorporating sentiment and emotion features to obtain the best performance. The results show that the multi-task classifier outperformed other baseline models considered in this paper. Hence, implicit hate speech detection models improved their performances by integrating emotions as a feature.

In our future work, we intend to leverage these findings to enhance the classification of various hate categories, particularly in the multilingual context. Furthermore, we see the potential to utilize fine-grained emotional analysis in related tasks such as fake news detection and user intent and behavior analysis. This integration will enable us to gain a deeper understanding of the underlying emotions and motivations behind hate speech instances.

To address the limitations of our current algorithm, we will explore data augmentation strategies and domain adaptation techniques to reduce dependence on limited annotations; allowing us to scale our system effectively to be capable of adapting to different online platforms and communities across diverse environments and improve its performance on various data sources.

Moreover, we plan to enhance hate speech detection performance through innovative techniques, including task weighting mechanisms, auxiliary loss functions, and refined hyperparameter tuning. By fine-tuning these aspects, we aim to achieve higher accuracy and reliability in identifying instances of hate speech. Moreover, we plan to utilize other large transformer-based models to enhance the contextual and semantic representations of implicit hate speech. By pursuing these future works, we envision a more robust and adaptable hate speech detection system, better equipped to combat hate speech across different languages, platforms, and communities, while mitigating the challenges associated with limited data and varying contexts.

REFERENCES

[1] G. del Valle-Cano, L. Quijano-Sánchez, F. Liberatore, and J. Gómez, "Socialhaterbert: A dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles," Expert Systems with Applications, vol. 216, p. 119446, 2023.

[2] A. Alrehili, "Automatic hate speech detection on social media: A brief survey," in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2019, pp. 1–6.

[3] G. Koushik, K. Rajeswari, and S. K. Muthusamy, "Automated hate speech detection on twitter," in 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA). IEEE, 2019, pp. 1–4.

[4] L. Gao, A. Kuppersmith, and R. Huang, "Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach," in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 774–782.

[5] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," in Proceedings of Conference on Empirical Methods in Natural Language Processing, 2021, pp. 345–363.

[6] M. Wiegand, J. Ruppenhofer, and E. Eder, "Implicitly abusive language–what does it actually look like and why are we not getting there?" in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 576–587.

[7] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain. Association for Computational Linguistics, 2019, pp. 1–10.

[8] A. Rodriguez, Y.-L. Chen, and C. Argueta, "Fadohs: Framework for detection and integration of unstructured data of hate speech on facebook using sentiment and emotion analysis," IEEE Access, vol. 10, pp. 22 400–22 419, 2022.

[9] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in 7th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2018.

[10] N. S. Samghabadi, A. Hatami, M. Shafaei, S. Kar, and T. Solorio, "Attending the emotions to detect online abusive language," in Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020, pp. 79–88.

[11] S. D. Swamy, A. Jamatia, and B. Gambäck, "Studying generalisability across abusive language detection datasets," in Proceedings of the 23rd conference on computational natural language learning (CoNLL), 2019.

[12] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, 2019, pp. 45–54.

[13] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 4, pp. 215–230, 2015.

[14] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," Policy & internet, vol. 7, no. 2, pp. 223–242, 2015.

[15] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[16] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in Tenth international AAAI conference on web and social media, 2016.

[17] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 27, no. 1, 2013, pp. 1621–1622.

[18] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1, 2017, pp. 512–515.

[19] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in Proceedings of the 24th international conference on world wide web, 2015, pp. 29–30.

[20] S. Köffer, D. M. Riehle, S. Höhenberger, and J. Becker, "Discussing the value of automatic hate speech detection in online debates," Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana, Germany, 2018.

[21] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, and V. Varma, "Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in twitter," in Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 70–74.

[22] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, ""the enemy among us" detecting cyber hate speech with threats-based othering language embeddings," ACM Transactions on the Web (TWEB), vol. 13, no. 3, pp. 1–26, 2019.

[23] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in European semantic web conference. Springer, 2018, pp. 745–760.

[24] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in twitter data using recurrent neural networks," Applied Intelligence, vol. 48, no. 12, pp. 4730–4742, 2018.

[25] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[26] N. Badri, F. Kboubi, and A. H. Chaibi, "Combining fasttext and glove word embedding for offensive and hate speech text detection," Procedia Computer Science, vol. 207, pp. 769–778, 2022.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[28] P. Liu, W. Li, and L. Zou, "Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers." in SemEval@ NAACL-HLT, 2019, pp. 87–91.

[29] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on bert model," PLOS ONE, vol. 15, no. 8, pp. 1–26, 08 2020.

[30] G. Glavaš, M. Karan, and I. Vulić, "Xhate-999: Analyzing and detecting abusive language across domains and languages." Association for Computational Linguistics, 2020.

[31] S. Akuma, T. Lubem, and I. T. Adom, "Comparing bag of words and tf-idf with different models for hate speech detection from live tweets," International Journal of Information Technology, pp. 1–7, 2022.

[32] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of EMNLP, 2014, pp. 1532–1543.

[33] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20. [Online]. Available: https://www.aclweb.org/anthology/W18-5102

[34] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4040–4054.

[35] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in Proceedings of the 11th International Workshop on Semantic Evaluation, ser. SemEval '17. Vancouver, Canada: Association for Computational Linguistics, August 2017.

[36] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, "Emotionally informed hate speech detection: a multi-target perspective," Cognitive Computation, vol. 14, no. 1, pp. 322–352, 2022.

[37] A. Rodriguez, C. Argueta, and Y.-L. Chen, "Automatic detection of hate speech on facebook using sentiment and emotion analysis," in 2019 international conference on artificial intelligence in information and communication (ICAIIC). IEEE, 2019, pp. 169–174.

[38] A. Rana and S. Jha, "Emotion based hate speech detection using multimodal learning," arXiv preprint arXiv:2202.06218, 2022.

[39] M. Z. Ali, S. Rauf, K. Javed, S. Hussain et al., "Improving hate speech detection of urdu tweets using sentiment analysis," IEEE Access, vol. 9, pp. 84 296–84 305, 2021.

[40] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," arXiv preprint arXiv:1905.05812, 2019.

[41] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," IEEE Access, vol. 9, pp. 112 478–112 489, 2021.

[42] F. M. Plaza-del Arco, S. Halat, S. Padó, and R. Klinger, "Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language," arXiv preprint arXiv:2109.10255, 2021.

[43] M. R. Awal, R. Cao, R. K.-W. Lee, and S. Mitrović, "Angrybert: Joint learning target and emotion for hate speech detection," in Pacific-Asia conference on knowledge discovery and data mining. Springer, 2021, pp. 701–713.

[44] P. Kapil and A. Ekbal, "A deep neural network based multi-task learning approach to hate speech detection," Knowledge-Based Systems, vol. 210, p. 106458, 2020.
[45] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," arXiv preprint arXiv:2005.14028, 2020.
[46] R. Caruana, "Multitask learning," Machine learning, vol. 28, pp. 41–75, 1997.
[47] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 7, pp. 3614–3633, 2021.
[48] A. Paszke, S. Gross, F. Massa, A. Lerer, and J. P. Bradbury, "An imperative style, high-performance deep learning library," Advances in Neural Information Processing Systems, vol. 32.

AMIR REZA JAFARI received the B.Sc. degree in computer engineering from the Isfahan University of Technology, Iran, in 2018, the M.Sc. degree in information technology - multimedia systems from the Tehran University, Iran, in 2021. He is currently pursuing a Ph.D. degree in Data Intelligence and Communication Engineering Laboratory (DICE) at Telecom SudParis, Institut Polytechnique de Paris, France. His research interests include social network analysis, data science, hate speech detection, and knowledge graphs.

GUANLIN LI Guanlin Li is currently a Ph.D. student in Data Intelligence and Communication Engineering Laboratory (DICE) at Samovar, Telecom SudParis, Institut Polytechnique de Paris, France. In 2022, he received his M.Sc. degree in Data & Artificial Intelligence from Telecom Paris, Institut Polytechnique de Paris, France. His primary research interests include natural language processing and social network analysis.

PRABODA RAJAPAKSHA Praboda Rajapaksha is a Research Fellow at the Institut Polytechnique de Paris, France. She holds a Ph.D. in Computer Science from the Institut Polytechnique de Paris, France (2021). She received her M.Eng. in Computer Science from the Asian Institute of Technology, Thailand; her M.Sc. in Communication Networks and Services from Institut Mines-Telecom, France and her B.Eng. in Computer Engineering from the University of Peradeniya, Sri Lanka. Her primary research interests include NLP, Language Modelling and Data Science.

REZA FARAHBAKHSH (Member, IEEE) received his Ph.D. degree from Paris VI (UPMC) jointly with the Institut-MinesTelecom, Telecom SudParis (CNRS Lab UMR5157), in 2015. He is currently an Invited Assistant Professor at the Institut-Mines Telecom, Telecom SudParis, and Lead Data Scientist at TotalEnergies SE. His research interests include AI and data science in scale, social networks analysis, language modeling, generative AI, and user behavior analysis.

NOEL CRESPI (Senior Member, IEEE) received the master's degree from the Universities of Orsay (Paris 11) and Kent (U.K.), the diplome d'ing nieur degree from the Telecom ParisTech, the Ph.D. and Habilitation degrees from UPMC (Paris-Sorbonne University). Since 1993, he has been working at CLIP, Bouygues Telecom, and then at Orange Labs in 1995. He took leading roles in the creation of new services with the successful conception and launch of Orange prepaid service, and in standardization (from rapporteurship of IN standard to coordination of all mobile standards activities for Orange). In 1999, he joined Nortel Networks as a Telephony Program Manager, architecting core network products for EMEA region. He joined Institut Mines-Telecom in 2002. He is currently a Professor and the Program Director, leading the Service Architecture Lab. He coordinates the standardization activities for Institute Mines-Telecom at ITU-T and ETSI. He is also an Adjunct Professor at KAIST, South Korea, an Affiliate Professor at Concordia University (Canada), and a Guest Researcher at the University of Goettingen, Germany. He is the Scientific Director of the French-Korean Laboratory ILLUMINE. His current research interests are in data analytics, the Internet of Things, and softwarization.