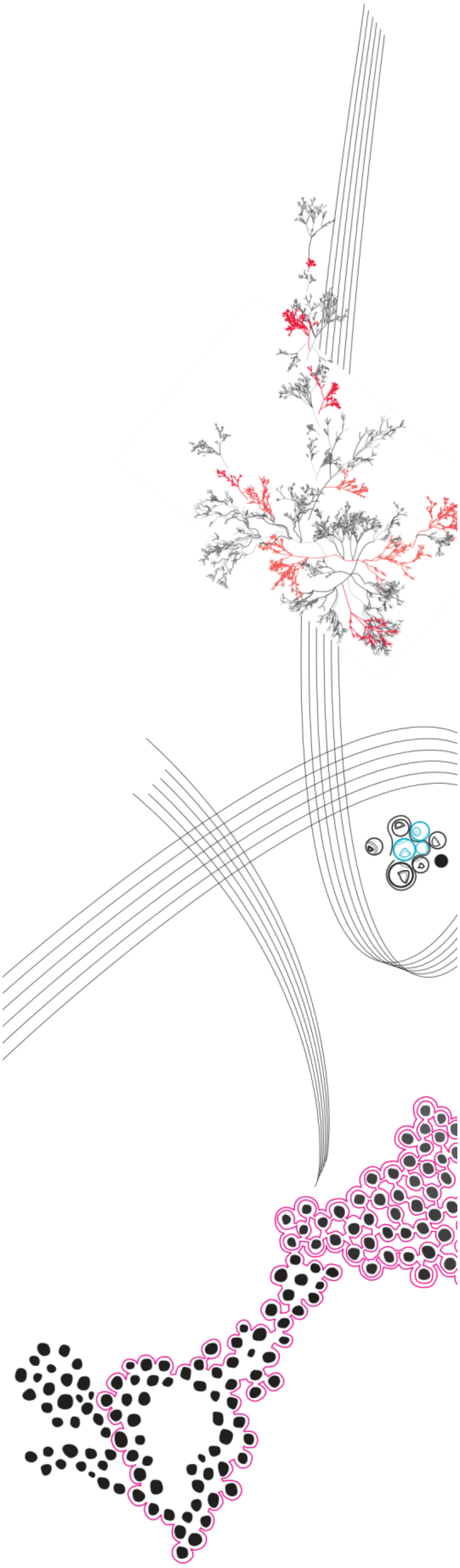


MSc Computer Science
Final Project

Leveraging Disagreement among Annotators for Text Classification

Jin Xu

Supervisor: Mariët Theune
Daniel Braun

May, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Abstract

Utilizing multiple annotators for data annotation is a common practice since it can ensure the quality of annotated data by facilitating error detection and reducing individual biases. However, this approach inevitably results in annotation disagreements, especially in subjective topics such as hate speech, emotion or sexism, as well as in some objective tasks like law and medical decision-making. While methods such as refining annotation guidelines, resorting to majority voting or dataset curator intervention have been employed to address annotation discrepancies, there is growing recognition of the limitations of traditional approaches that force a single "ground truth" label. It can sacrifice the valuable nuances and diverse perspectives inherent in annotators' assessments, thereby compromising the authenticity and richness of annotated dataset. Thus, there is a shifting paradigm towards preserving annotation disagreements to maintain the diversity of opinions and objectiveness of labeled dataset.

In this study, we introduce approaches that incorporate annotation disagreement into the model training process. We mainly focus on hate speech detection and abusive conversation detection, tasks inherently entailing a high degree of subjectivity. Our approaches construct models using three different strategies: probability-based multi-label method, ensemble system and instruction tuning. The probability-based multi-label method treats the detection tasks as a multi-label text classification problem and gives probability distribution across different labels. The ensemble system imitates the annotation process that involves multiple annotators. It consists of multiple sub-models that are trained individually, thereby incorporating diverse perspectives within the annotations. The predictions from all sub-models are combined and transformed into the final decisions. Both the multi-label method and the ensemble system use BERT as their foundation models. Instruction tuning shares the same principle with the ensemble system but employs LLaMa 2 as the foundation model and fine-tunes it through the use of natural language instructions. Cross entropy is utilized as a metric to compare the performance of these three approaches. Moreover, to evaluate the effectiveness of embracing annotation disagreements for model training, we conduct an online survey that compares the performance of the multi-label model against the baseline model. The baseline model shares the identical structure with the multi-label model but is trained with the majority label. In the survey, participants are asked to give their preferences for the outputs generated by these two models.

Our experimental results show that, in hate speech detection, the multi-label method outperforms the ensemble system and instruction tuning, even though the latter two approaches have much more complicated model structures and larger parameter sizes. For abusive conversation detection, instruction tuning achieves the best performance because unlike the other models, it relies less on using an extensive training dataset. Through significance testing, we find that the outputs from the multi-label model are considered more reasonable than those from the baseline model to characterize samples from the online survey. This proves the effectiveness of leveraging annotation disagreements for model training.

Acknowledgements

First and foremost, I would like to thank Dr. Daniel Braun for giving me this opportunity to work on this project on leveraging annotation disagreements among annotators. There were some minor obstacles on the road to obtaining approval for this undertaking. Therefore, I would also like to express my deep thanks to Dr. Mariët Theune for taking the responsibility of chair of the thesis committee. Otherwise, it would not be possible for me to work on this project. Next, I would like to extend my heartfelt gratitude to them for providing patient and detailed guidance throughout the process. The weekly meetings over the last six months have been very beneficial thanks to their academic insight and expertise, which ensured that this research work progressed in the right direction. Over this period, they consistently gave valuable feedback and raised thought-provoking questions. Furthermore, I appreciate their meticulous review of this thesis, which has significantly enhanced its quality and clarity.

I am extremely grateful for the friends who generously dedicated their precious time to participate in the online survey, which was an important component of this research. Without their involvement, the completion of this project would not have been so smooth, and I am deeply thankful for their support.

This thesis marks the end of my master's studies, and I would like to express my sincere gratitude to those who have supported me during this journey. Their guidance, encouragement, and unwavering belief in my abilities have been indispensable over the past two years.

Contents

1 Introduction

Employing multiple annotators for data annotation is a widely adopted practice since it can mitigate the individual biases and allows for error detection and correction, thereby ensuring the quality and reliability of annotated data. Nevertheless, this practice will inevitably lead to annotation disagreement. Annotation disagreement refers to instances where annotators, often individuals tasked with labeling or marking data, have not reached a consensus on the appropriate annotation in the process of annotating data. This lack of agreement can arise due to various factors, including differences in interpretation, subjective judgment, or ambiguity in the guidelines provided to annotators. Implementing inter-annotator agreement metrics can help quantify and manage annotation discrepancies. However, the importance of addressing annotation disagreement extends beyond this, especially in tasks related to natural language processing (NLP), computer vision, and machine learning, where accurately labeled datasets are foundational for training models (Pang & Lee, 2004; Snow et al., 2008).

Strategies to deal with annotation disagreement may include refining annotation guidelines, providing clearer instructions to annotators, and conducting regular training sessions to ensure a shared understanding of the labeling criteria. Additionally, the annotations provided by individuals can be combined through methods such as majority voting (Sabou et al., 2014) or other alternative procedures (e.g. the involvement of domain experts). In these ways, the consensus is achieved and subsequently utilized in training supervised machine learning algorithms. However, for some subjective annotation tasks involving hate speech, emotion or sexism, as well as for some objective tasks such as law and medical decision making (Dumitrache et al., 2018), there may be not a definitive right answer or true label (Alm, 2011; Cabitza et al., 2023). Under these circumstances, preserving annotation disagreements is essential for maintaining the quality and reliability of labeled datasets used in training and evaluating machine learning models. Instead, forcing a single "ground truth" label can sacrifice the valuable nuances inherent in annotators' assessments of the stimuli and their disagreement, thereby diminishing the authenticity and representativeness of annotated data (Cheplygina & Pluim, 2018). Therefore, there is a paradigm shift in the academic community that moves away from the conventional approach of constructing monolithic, majority-aggregated gold standards. In contrast, harmonization, which typically involves aligning annotations through methods like majority voting, is not universally favored as the primary means of generating gold standards.

In this study, we propose approaches that incorporate annotation disagreement into model training process. We choose two text classification tasks in the field of NLP which inherently entail a high degree of subjectivity: hate speech detection and abuse detection in conversation AI. In our chosen datasets, these two tasks exhibit different complexity and difficulty in terms of the label space. Hate speech detection is to determine whether one given text is hate speech or not, while abusive conversation detection requires not only identifying abusive text but also classifying the severity of the abuse. For these two detection tasks, the baseline model only considers the majority labels derived from the multiple annotations for training. Conversely, our three proposed approaches formulate the incorporation of multiple annotations during model training with different strategies: the probability-based multi-label method, the ensemble system and instruction tuning. Firstly, given that the final annotation for each instance is the probability distribution over different classes after the integration of the annotations from all the annotators, we tackle this task as a probability-based multi-label text classification problem. Instead of predicting specific label(s) to one instance, the model gives a probability distribution. Each value in this distribution represents the level of likelihood regarding the instance's association with each label. Secondly, we imitate the process of annotation from multiple annotators and approach this task by proposing an ensemble system. The ensemble system consists of many sub-models. Each sub-model is trained on its respective labels to capture the diverse viewpoints embedded in the annotations. Thirdly, since instruction tuning facilitates injecting explicit guidance into the training process and allows for explicit customization of model's behavior, it is applied to both detection tasks. Specifically, we use the pre-trained generative model and fine-tune it typically for the datasets. The response of the model is one annotation. In the baseline, multi-label method and ensemble system, BERT is adopted as their foundation model, while LLaMa 2 is utilized for instruction tuning. The performance of the proposed models on two datasets is compared using cross entropy. Besides, to evaluate the effectiveness of incorporating multiple labels during model training, we conduct an online survey. This survey aims to investigate individuals' preferences between the outputs generated by the multi-label model and those by the baseline model. These two models share an identical structure, and their only difference is the labels they are trained on. Therefore, although the baseline model is trained with single labels, it is tasked with generating probability distributions in the inference phase, which facilitates the comparison. For each selected sample, participants are presented with probability distributions generated by both the multi-label model and the baseline model during the online survey. They are then asked to indicate which of these two probabilistic annotations they find more reasonable to describe the sample. By employing significant testing on the collected data, we aim to reveal whether individuals exhibit a preference for one of the probabilistic annotations to characterize the samples they encounter in the online survey.

With these background and introduction, our research questions are as follows:

Firstly, we aim to design models that can incorporate individual annotations from multiple annotators during the model training phase.

- **RQ 1:** How can models be designed to incorporate individual annotations from multiple annotators during the training, instead of only considering the majority label derived from these annotations?

Then, with these proposed models, we intend to compare their performances across the two selected tasks.

- **RQ 2:** How do the proposed models perform in hate speech detection and abuse detection in conversational AI?

Lastly, we aim to explore whether incorporating multiple annotations can improve model performance compared with solely using the majority label. Since these two approaches generate different output formats, we need to design a method that allows for the comparison between them.

- **RQ 3:** How can we design a method to evaluate the effectiveness of incorporating multiple labels for model training against the model that only considers the majority label?

The remainder of this thesis is structured as follows: Chapter 2 presents an overview of related work, including sources of annotation disagreement, approaches to tackling such disagreements, and key techniques relevant to this research, such as BERT, large language models, parameter-efficient fine-tuning and instruction tuning. Chapter 3 discusses our two experimental datasets which correspond to two text classification tasks: hate speech detection and abuse detection in conversation AI. The methodology to be employed is illustrated in Chapter 4. It contains the baseline model, the three primary models that we propose for incorporating annotation disagreement during model training, and the evaluation metrics. Subsequently, Chapter 5 gives a description of the conducted experiments along with the experimental results, and Chapter 6 delves into a comprehensive analysis of these outcomes. Finally, the conclusion and limitations are summarized in Chapter 7.

2 Background

Recently, a large number of studies have been conducted to tackle annotation disagreement. This chapter primarily concentrates on reviewing related work in the sources of annotation disagreement, tackling annotation disagreement and key techniques relevant to our research, such as BERT, large language models, parameter efficient fine-tuning and instruction tuning, etc.

2.1 Related Work

2.1.1 Sources of annotation disagreement

The disagreements in annotation can come from different sources, such as the inherent ambiguity of text subjectivities and variations in value systems of annotators.

On one hand, natural language, especially texts, can be inherently complex and can be interpreted in multiple ways within a given context (Poesio, 2020). There are many subjective elements existing in the texts which may add an additional layer of intricacy to the understanding of texts, such as sentiments, opinions or nuanced expressions. Therefore, it is common that there are divergent interpretations among annotators. Furthermore, some sentences or even labels may contain vague or ambiguous statements (Russell et al., 2008), making it challenging for annotators to understand or reach an agreement.

On the other hand, some characteristics of annotators can have a significant impact on the annotation results, such as cultural differences, individual value systems or personal discrepancies, etc (Davani et al., 2022). For example, through post-annotation interviews, Patton et al. revealed that annotators who come from communities discussed in gang-related tweets are more likely to rely on their lived experiences in the process of annotating when compared to graduate student researchers. This divergence results in distinct label judgments (Patton et al., 2019). Additionally, Luo et al. found that the political affiliation of annotators can significantly shape how they assess and annotate the neutrality of political stances (Luo et al., 2020).

The above-mentioned studies reveal the multifaceted nature of annotation disagreement, underscoring both text-related complexities and annotator-specific influences.

2.1.2 Tackling annotation disagreement

Majority voting involves aggregating annotations by selecting the label that the majority of annotators agree upon. It is obvious that majority voting is intuitive, easy to understand and implement (Uma et al., 2021). Furthermore, it tends to perform well when the annotators share unanimous perspectives. However, the employment of a majority voting method in annotation processes can unintentionally obscure nuanced viewpoints, especially for groups that are underrepresented in annotator pools (Prabhakaran et al., 2021). For instance, this is particularly evident when considering older adults, who may hold distinctive views on aging that differ from those of crowd workers, the majority of whom are typically younger. The reliance on a majority vote mechanism may lead to the overshadowing of unique insights and experiences. To address this concern, it is important to ensure a diverse representation among annotators to foster a more comprehensive understanding of various perspectives, particularly those from underrepresented demographics (Wan et al., 2023). Recognizing this necessity is crucial for fostering inclusivity and preventing the marginalization of specific viewpoints in annotation tasks.

Therefore, there are some studies that have introduced alternative methods to majority voting when aggregating multiple annotations. In 2012, De Marneffe et al. trained a classifier that can predict event veridicality distributions (whether events described in a text are viewed as happening or not). Three types of features (lexical features, structural features and world knowledge) were used and selected through 10-fold cross validation (de Marneffe et al., 2012). All of these features aim to capture different factors that can influence the veridicality of the text. From the perspective of information theory, Waterhouse measured a contributor's judgment based on how much the judgement helps to reduce the entropy of our finding the "true" labels. These labels are estimated by a collective judgment resolution algorithm that takes into account the measurement of annotation workers' contributions. And this quantity is expressed by the pointwise mutual information (PMI) between the annotated labels and the "true" ones (Waterhouse, 2013). Furthermore, they also used conditional PMI to measure the intersections between annotators. With these measurements, the "true" labels can be correctly estimated. In 2013, Hovy et al. proposed an item-response model, which was trained in an unsupervised way. By treating the "correct" labels as latent variables, the model gains the ability to predict whom to trust and when, rather than relying solely on applying majority voting on all samples (Hovy et al., 2013). Experimental results showed remarkable improvements over baselines for predicting label and estimating trustworthiness. In 2021, Gordon et al. introduced a disagreement transformation that leveraged multiple annotators' judgments and disentangled stable opinions from noise by estimating intra-annotator consistency (Gordon et al., 2021). Using this algorithm, the final annotation for each instance was sampled from the primary labels, with random noise added, while non-primary labels were excluded.

Some studies have developed methods for incorporating annotation disagreement in the process of model training. In 2019, Chou et al. incorporated the characteristics of each annotator in the inner layers of the neural network. In their experiment, they also introduced a joint learning methodology that simultaneously modelled the label uncertainty and annotator idiosyncrasy by using both hard label (majority voting) and soft label (the distribution of annotations) (Chou & Lee, 2019). The results showed that the added features contain useful information that significantly boosts the model performance. In 2021, Fornaciari et al. proposed a multi-task neural network that was trained on soft label distribution over annotator labels (Fornaciari et al., 2021). By integrating a divergence measurement between soft label and "true" label vector into the loss functions, they effectively mitigated overfitting and therefore improved performance across different tasks. In 2022, Davani et al. introduced multi-annotator models where each annotator's judgements were regarded as independent subtask with a shared common representation of the annotation task (Davani et al., 2022). This approach, on one hand, enables to preserve and model the internal consistency in each annotator's label. On the other hand, it also incorporates the systematic disagreements with other annotators. Similarly, the network architecture introduced by Guan et al. incorporates the concept of "crowd layers" to individually model annotation experts (Guan et al., 2018). In this approach, each expert's model weight is calculated independently, and these individual weights are then averaged to facilitate ensemble recognition. In order to include the knowledge from all the annotators, Fayek et al. employed neural networks to build an ensemble system that consists of many models, with each model representing one annotator. Then the final results are obtained by combining the individual model outputs. For the purpose of comparison, they also trained a model that is fed with the soft labels from all annotators (Fayek et al., 2016). The results showed that these two approaches exhibit similar performance, which means that the performance improvement from the ensemble could be achieved by using soft labels from annotators.

Although the approaches outlined above have improved the performance by leveraging annotation disagreements into model training, they remained limited to identifying the majority label. The outputs, in the form of "soft labels" (probability distribution over labels), were still aggregated to single labels as final predictions. Accordingly, there is also a lack of research focusing on evaluating the effectiveness of embracing multiple labels during model training. To address the first gap, this study proposes three approaches for constructing text classification models that use "soft labels" as targets. The relevant techniques and knowledge will be introduced in the following subsections. The handling of the second gap will be discussed in Section 4.3.2.

2.2 Relevant Techniques

2.2.1 BERT

BERT, short for Bidirectional Encoder Representations from Transformers, is a self-supervised learning model which can represent words as low-dimensional embeddings. The embedding provides rich contextual information and this is achieved through the utilization of self-attention mechanism. The BERT model consists of a series of encoders which are stacked on top of each other (see Figure 1). These encoders are based on the Transformer architecture, which is a type of neural network architecture specifically designed for processing sequential data, such as natural language text. There are two sub-layers within each encoder: a multi-head self-attention mechanism and a feed forward neural network (Devlin et al., 2018). The self-attention mechanism allows the model to selectively attend to the most relevant parts of the input sequence, while the feedforward neural network captures nonlinear dependencies between the sequence elements. Through the stacking of multiple encoders, the BERT model is able to learn increasingly complex representations of the input text, thereby capturing contextual information from different scales.

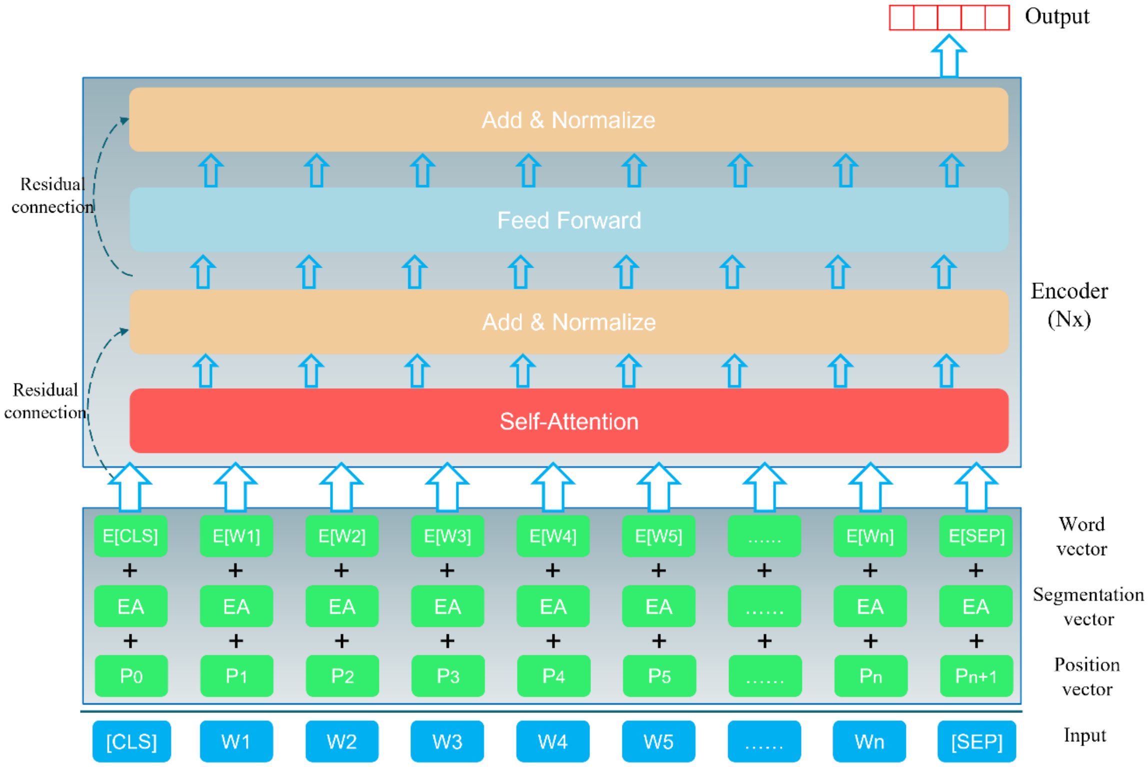


Figure 1 The architecture of BERT (Devlin et al., 2018; Vaswani et al., 2017)

The input to the BERT model consists of a sequence of words and other special tokens such as “CLS” and “SEP”. First of all, the positional embedding, segmentation embedding and word embedding are applied to each token, and these embeddings will be summed to form a new vector. After that, these embeddings are passed through multiple transformer encoder layers in stack. During pre-training, BERT employs a masked language modelling (MLM). In this approach, a random subset of the input tokens are replaced with a special mask token (“[MASK]” in Figure 1) and the model is trained to predict the masked tokens based on the surrounding context provided by other tokens in the sequence. In this way, BERT learns how words relate to each other, which forces it to develop strong contextual representations. Additionally, BERT employs a next sentence prediction (NSP) objective. In this objective, the model is provided two sentences and trained to predict whether they are consecutive in the original text or not (depicted as “CLS” as the output). The final output consists of a sequence of contextualized embeddings for each token in the input sequence. These embeddings encapsulate not only the meaning of the individual words but also their relationships with others in the sequence. Importantly, these embeddings are commonly used as input for downstream tasks.

The self-attention layer within each encoder block enables the model to selectively focus on specific parts of the input (Vaswani et al., 2017). The architecture of the attention mechanism is illustrated in Figure 2. There are three matrices: W_Q, W_K and W_V, which are derived via network training. As shown in this figure, by using these matrices, input vectors $X_1, X_2, \ldots \ldots, X_n$ are transformed into new vectors q_i, k_i and v_i $(1 \leq i \leq n)$. Subsequently, each element of the input sequence is compared with every other element in the same sequence. This is achieved by computing a score between each pair of elements (q_i and k_i), which indicates their relatedness. The score is calculated using a dot product operation between the embeddings of the two elements. Following this, scores are normalized using a SoftMax function to obtain a set of weights that signify the relative importance of each element in the sequence with respect to others. Utilizing these weights, a weighted sum of the embeddings (v_i) is computed, which is a contextual and low-dimensional vector that encapsulates the most relevant information within the sequence. The attention calculation in a self-attention layer is given in formula (1), where $Q = (q_1, q_2, \ldots \ldots, q_n)$, $K = (k_1, k_2, \ldots \ldots, k_n)$ and $V = (v_1, v_2, \ldots \ldots, v_n)$. In the denominator, d_k represents the dimension of q_i and k_i. It serves to scale the values before applying the SoftMax function. This scaling makes sure that the softmax outputs are not overly influenced by the magnitude of the input vectors. With these three matrices, this layer calculates the relation between different tokens with dot product and form a new vector for each token.

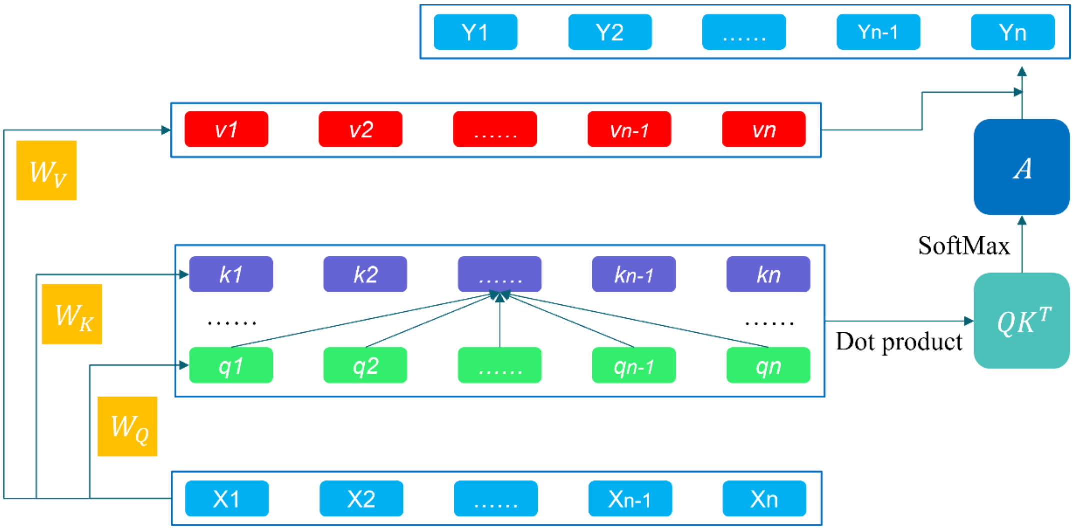


Figure 2 The architecture of the attention layer

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

The input of the self-attention layer will be added to its output (residual connection in Figure 1) and then layer-normalized ("Add & Normalize" in Figure 1). Next, the output of this "Add & Normalize" layer will be fed into a feed forward layer and another "Add & Normalize" layer. Finally, the last hidden state vector, which serves as the representation of the text, will be used as the output of this model. Generally, this vector will be put into a full-connected network to fit the downstream tasks.

2.2.2 Large language models

In recent years, there has been a surge in the development and utilization of large language models (LLM). As a result, "pretrain-then-finetune" has become a new paradigm for addressing various NLP tasks. Large pre-trained models have shown remarkable abilities in understanding and generating texts. For example, autoregressive models (decoders only) like GPTs (Generative Pre-trained Transformer) (Brown et al., 2020), LlaMa (Large Language Model Meta AI) (Touvron et al., 2023) exhibit strong ability in generating human-like text, while auto-encoding models (encoders only) like BERT, RoBERTa (Liu et al., 2019) and Electra (Clark et al., 2020) are good at natural language understanding. Different from these two types, T5 (Text-To-Text Transfer Transformer) is an encoder-decoder architecture designed specifically for conditional generation tasks (Raffel et al., 2020), where text is generated based on specific conditions or context. One of the most impressive aspects is that T5 serves as a unified text-to-text framework for different kinds of NLP tasks. Whether for text classification or generation, these challenges can be effectively tackled by T5 in a consistent manner.

These models are typically trained on massive datasets to learn the patterns and structures of language. Once pre-trained, they can be fine-tuned on specific tasks or domains to achieve better performance. Fine-tuning involves training the model on a smaller, task-specific dataset to adapt it to the nuances and requirements of a particular application, such as sentiment classification, summarization, translation, question-answering, etc.

The availability of pre-trained models and their fine-tuning capabilities has democratized access to advanced language processing for a wide range of applications. Researchers and developers can leverage these models to build applications with improved natural language understanding and generation, saving time and resources compared to training models from scratch.

2.2.3 Parameter efficient fine-tuning

Despite the widespread popularity and impressive performance across various research domains, the fine-tuning of LLMs can pose challenges since it is resource-intensive and time-consuming, especially for models with massive parameter sizes like LLaMA, Vicuna (Chiang et al., 2023). In some cases, it is even problematic to load such models onto a single GPU since they can lead to memory constraints and slower processing speeds. This can hinder the smooth fine-tuning of the models for downstream tasks. Consequently, various strategies have been proposed by researchers to facilitate the efficient deployment of these large models into memory and optimize the fine-tuning process to achieve acceptable efficiency.

For example, LoRA was introduced in 2021 as an innovative solution for parameter efficient fine-tuning (PEFT). It involves freezing the basic model and adding an adapter layer that consists of two trainable low-rank decomposition matrices. By freezing the base model and only optimizing these two matrices during fine-tuning process, LoRA significantly reduces the number of model's trainable parameter and avoids the computational overhead (Hu et al., 2021). Impressively, their experimental results demonstrated that LoRA can lead to a remarkable threefold reduction in GPU memory requirements. In 2022, Liu et al. introduced prompt-tuning, a technique involving the freezing of the basic model and updating the discrete embeddings of manual prompts during training. This approach substantially reduced per-task storage and memory usage (Liu et al., 2022). However, it is important to note that a slight change in prompts, despite targeting the same task, could influence model performance. Furthermore, the embeddings are discrete since they are derived from manual prompts. To address the instability associated with manual discrete prompt embeddings, Liu et al. proposed P-tuning, a technique that adds randomly generated continuous embeddings before the discrete embeddings of manual prompts. This method not only stabilized training by minimizing gaps between different discrete prompts, but also led to remarkable performance improvements across a wide range of tasks (Liu et al., 2023). In order to further optimize the memory usage of parameters, Dettmers et al. proposed NF4 (4-bit NormalFloat), a quantization method which store pre-trained neural network weights with integer rather than floating-point values (Dettmers et al., 2023). By utilizing a zero-centered normal distribution to divide the range of floating-point values, NF4 aligns well with the distribution of weight values within the model, thereby reducing the quantization errors. In 2021, with the similar idea with P-tuning, Li et al. proposed prefix-tuning technique. Different from P-tuning, which inserts continuous vectors solely in the input layer, prefix tuning inserts continuous vectors for each layer of the model, resulting in a larger parameter size. The authors discovered that prefix-tuning demonstrated impressive results. It can achieve comparable performance in full data settings and outperform fine-tuning in low data settings by learning only 0.1% of the parameters (Li & Liang, 2021).

In order to deal with the challenges that manually designing prompts is troublesome and automatically generated prompts is difficult and time-consuming in multi-class text classification tasks, Han et al. proposed a framework that creates a sub-prompt based on rules (Han et al., 2022). They injected these rules with the prior knowledge of the text classification through encoding. Their experiments were conducted on three multi-class classification tasks, including relation classification, entity typing, and intent classification. The results showed that the prompt tuning method with rules outperforms prompt-based methods in terms of the efficiency in constructing prompts.

2.2.4 Instruction tuning

With the wide-range application of transformer-based generative language models, the paradigm of using natural language to induce model's behaviors has become a popular research topic. First of all, in-context learning is a type of machine learning where the model learns from a specific context and generates predictions based on it (Radford et al., 2019). Different from traditional machine learning algorithms that train models on a fixed dataset, in-context learning allows models to adapt and learn from new information or data points encountered during inference phase (Brown et al., 2020). It is important to note that the in-context learning does not train and update the weights of the model. There are several types of in-context learning methods, such as few-shot, one shot and zero shot. In few-shot learning, the model is given a few examples of the task during inference as conditioning. In particular, one example usually consists of a context and a desired completion, and for few-shot learning, there are K examples given where K is often set in the range of 10 to 100. By contrast, in one-shot learning, K equals to 1. For zero-shot, instead of providing any examples, we only offer a natural language description of the task. One of the primary advantages of in-context learning is that it does not require a large amount of task-specific data. Even in few-shot learning, usually less than a hundred examples will be sufficient to control the model's generation (Zheng et al., 2021). However, due to the lack of training and fine-tuning process, the performance from this approach is significantly worse than many fine-tuned models (Sanh et al., 2021). Conducting in-context learning relies exclusively on the prior knowledge stored by a model during pre-training. In addition, in-context learning imposes significant computational, memory, and storage costs as it necessitates the processing of all training examples each time a prediction is generated (Liu et al., 2022). Also, the exact formatting of the prompt (including the wording and the arrangement of examples) can exert a substantial and unforeseeable influence on the model's performance, extending well beyond the variations observed during fine-tuning across different runs (Webson & Pavlick, 2021; Zhao et al., 2021).

Prompt tuning is also another way to control model's output with natural language. However, unlike in-context learning, prompt tuning has the fine-tuning process, but no weights of the model are updated (Shin et al., 2020). Instead, only the embeddings of the prompts are updated during the fine-tuning process. As a result, we do not have to save different models with different parameters for different downstream tasks. Conversely, only the fine-tuned embeddings of prompts corresponding to specific tasks need to be stored, and these are dramatically smaller than the whole pre-trained model. Nevertheless, prompt tuning can lead to model's overfitting to specific prompts. This process typically involves fine-tuning a pre-trained model on a specific dataset and task using a fixed set of prompts. Such an approach restricts the transferability of the model to other tasks or domains, as it may demand extensive fine-tuning with new prompts for each new task (Su et al., 2021). For example, Liu et al.

showed that even though prompt tuning shows superiority on some of the natural language understanding benchmarks, it performs poorly on typical sequence tagging tasks compared to fine-tuning (Liu et al., 2022). Furthermore, the model's performance is not always consistent across different model scales. Lester et al. proved that for medium-sized models (ranging from 100 million to 1 billion parameters) that are widely used, prompt tuning yields significantly inferior results compared with fine-tuning, while its performance can become comparable to fine-tuning when the model scales beyond 10 billion parameters (Lester et al., 2021). Thus, instruction-tuning was proposed and well-studied.

As a new tool of model tuning, instruction tuning has become increasingly popular and widely used to induce model through the usage of natural language instructions. In recent year, there has been a lot of work focusing on applying instruction tuning on a wide range of NLP tasks (Mehri & Eric, 2021; Wang et al., 2022; Wei et al., 2022), such as intent identification, sentiment analysis. The advantages of instruction tuning lie not only in its ability to explicitly guide the model's outputs to align with the desired response characteristics or domain knowledge, but also in its remarkable transferability across various tasks, showcasing superior generalization to previously unseen tasks. This process involves the fine-tuning of the pre-trained model with a specific dataset. In 2022, Gupta applied instruction tuning on dialogue to enhance model's performance on zero-shot and few-shot settings (Gupta et al., 2022). In their experiments, they adopted a unified text-to-text format based on 59 openly available dialogue datasets. Two pre-trained large language models were selected: T0-3B and BART0. In particular, T0-3B (Sanh et al., 2021) is a fine-tuned version of T5 (Lester et al., 2021) with three billion parameters. It is fine-tuned on a multitask mixture of general non-dialogue tasks, such as question answering, paraphrase recognition and sentiment analysis. BART0 has a parameter size of 406 million (Lin et al., 2022). It is fine-tuned on the same task mixture as T0-3B on the basis of Bart-large (Lewis et al., 2020). The experimental results indicated that their proposed model could lead to good zero-shot performance on unseen data, and in many NLP tasks, it can even outperform few-shot learning. Given that LLMs does not always follow the users' intent as they are becoming increasingly bigger, such as making up facts, generating biased or harmful text, or just simply obviating instructions from users (Bender et al., 2021; Weidinger et al., 2021), Ouyang et al. proposed InstructGPT which is fined-tuned via reinforcement learning to incorporate human preferences (Ouyang et al., 2022). In this way, they managed to make sure that the LLM can act in accordance with human's intentions. Specifically, they selected GPT-3 as the pre-trained model and fine-tuned it by writing some instructions which demonstrated the desired output behavior from the model. The primary fine-tuning strategy is reinforcement learning from human feedback (RLHF) since in the second step of their experiments, human preference was utilized as a reward signal to train a reward function. The reinforcement learning algorithm employed was Proximal Policy Optimization (PPO) (Schulman et al., 2017). From their experimental results, an improvement in the truthfulness and reductions in toxic output generation can be observed, even though InstructGPT still makes some small mistakes.

2.3 Chapter Summary

In this chapter, we mainly introduced the relevant background knowledge that is utilized in this research. Firstly, we discussed the sources of annotation disagreement and the way to tackle it. Then, we introduced some newly released techniques that are relevant to our experiments, including large language model, BERT, parameter efficient fine-tuning and instruction tuning. Due to the capacity of LLM in solving many NLP problems, this research will mainly unfold around LLM, instead of the traditional machine learning algorithms. To reduce the training time and the requirement of computational power, we used PEFT, which achieves this by reducing the size of the parameters that need to be upgraded in the process of fine-tuning. We also applied instruction tuning to one of our proposed models given its ability to directly control the model's output with natural language.

3 Datasets

In this study, we want to discover discrepancies in model performance across different datasets. To achieve this, two datasets are utilized in our experiments. They show differences in some aspects, such as data size, the number of annotators involved in each sample, classification difficulty, etc.

3.1 “Large-Scale Hate Speech” Dataset

The first dataset utilized in this study is the “Large-Scale Hate Speech Dataset”[1] (hereinafter referred to as “hate speech dataset”) (Toraman et al., 2022). This dataset consists of 200,000 tweets, evenly split between Turkish and English languages. For our experiments, we specifically focus on 100,000 English tweets, with 7,000 tweets for training, 1,500 for validation, and another 1,500 for testing. Five distinct domains are involved in this dataset: Religion, Gender, Race, Politics, Sports. For both languages, there are 2,000 tweets for each domain. There are a total of 20 annotators in the annotation panel. And each tweet is annotated by randomly selected five anonymous annotators. The label space is {0,1,2}, where 0 corresponds to “Normal”, 1 to “Offensive” and 2 to “Hate” (Toraman et al., 2022). According to the annotation guidelines utilized by Sharma et al. and Toraman et al., tweets are categorized as “Hate” if they target, incite violence against, threaten, or advocate for physical harm towards an in dividual or a group of people based on identifiable trait or characteristic. And if tweets humiliate, taunt, discriminate against, or insult an individual or a group of people, they are annotated as “Offensive”. In the absence of these criteria, the tweets are labeled as “Normal” (Sharma et al., 2018; Toraman et al., 2022). An example of a data sample is demonstrated in Table 1.

Table 1 An example from the hate speech dataset

Column Name	Description	Example
TweetID	Twitter ID of the tweet	1344215464352821248
Text	Tweet’s text contents	I don’t care what this man’s beliefs were, no one deserves to die this way. Hopefully this event will help raise awareness......
LangID	Language of the tweet 0-Turkish, 1-English	1
TopicID	Domain of the topic 0-Religion, 1-Gender, 2-Race, 3-Politics, 4-Sports	0
Label_1	Annotation of the first annotator 0-Normal, 1-Offensive, 2-Hate	1
Label_2	Annotation of the second annotator 0-Normal, 1-Offensive, 2-Hate	0
Label_3	Annotation of the third annotator 0-Normal, 1-Offensive, 2-Hate	0
Label_4	Annotation of the fourth annotator 0-Normal, 1-Offensive, 2-Hate	1
Label_5	Annotation of the fifth annotator 0-Normal, 1-Offensive, 2-Hate	0
HateLabel	Final hate label decision 0-Normal, 1-Offensive, 2-Hate	0

[1] https://github.com/avaapm/hatespeech/tree/master/dataset_v1

Here, we also present the statistical characteristics of tweets in the dataset. As illustrated in Table 2, a serious class-imbalance issue exists in this dataset. The predominant class, “Normal” constitutes approximately 66% of all samples across various domains. By contrast, only 27% and 7% of the tweets are from “Offensive” and “Hate” categories respectively. In our experimental results, we will present the model performance on each class.

Table 2 The statistical characteristics of tweets in the hate speech dataset (English part)

Domain	**Hate**	**Offensive**	**Normal**	**Total**
Religion	1,427	5,221	13,352	20,000
Gender	1,313	6,431	12,256	20,000
Race	1,541	3,846	14,613	20,000
Politics	1,610	6,018	12,372	20,000
Sport	1,434	5,624	12,942	20,000
Total	7,325 (7%)	27,140 (27%)	65,535 (66%)	100,000

3.2 “Abuse in Conversational AI” Dataset

The second dataset is “Abuse in Conversational AI” dataset[2] (hereinafter referred to as “abusive conversation dataset”) (Curry et al., 2021). The data was collected from conversations between users and three different conversational AI systems (Alana v2, CarbonBot and ELIZA), which have different goals and properties. Specifically, two of these systems are classed as chatbots, serving as social, open-domain platforms, while the third one operates as a transactional, goal-oriented system.

Alana v2 is one of the chatbots developed in Alexa Challenge 2018. This is a competition in which university teams were required to develop engaging social chatbots that can have conversations with users. The chatbot seamlessly integrated social chit-chat with the provision of information through entity linking. Users were informed about the competition at the beginning of the conversation. The dataset comprises automatically transcribed user utterances, inclusive of recognition noise, and was collected during the period from April 2017 to November 2018.

CarbonBot is an assistant developed by Rasa[3] and hosted on Facebook Messenger[4]. This bot’s primary goal is to persuade the user to consider buying carbon offsets for their flights. The data for CarbonBot was collected over a period spanning from 1st October 2019 to 7th December 2020. Additionally, it also informed the user that their conversations will be recorded for research purposes.

ELIZA is an implementation of a rule-based conversational agent designed to emulate the role of a psychotherapist (Weizenbaum, 1966). This agent serves academic purposes and is hosted at the Jozef Stefan Institute[5]. Its primary motivation is to engage the users through the presentation of open-ended questions such as “Tell me more about……”. The data collection for ELIZA took place from December 19, 2002 to November 26, 2007.

Regarding the annotation of this dataset, the author adopted the unbalanced rating scale proposed by Poletto et al. (Poletto et al., 2019), in which inputs are labelled on a scale from +1 (Not abusive) to −3 (Very strongly abusive). This annotation scheme offers insights into not only the presence of abusive content, but also the severity of the abuse. Specifically, -3 denotes content that is strongly negative with overt incitement to hatred, violence, or discrimination, and an attitude geared towards attacking or demeaning the target. A label of -2 indicates content that is negative, insulting, or abusive, with an aggressive tone. -1 represents content that is negative and impolite, mildly offensive but still within the

[2] https://github.com/amandacurry/convabuse/tree/main
[3] https://rasa.com/
[4] https://m.me/CarbonBot.from.Rasa
[5] http://www-ai.ijs.si/eliza

bounds of conversational language. A label of 0 suggests ambiguity or unclear sentiment, while a +1 indicates the non-abusive content. In the annotation process, eight annotators have been recruited, and each example is annotated by a minimum of three annotators to ensure annotation quality.

Table 3 gives the dataset size, the number of annotations and the percentage of abuse annotations from three different datasets. From the table, the ratio of annotation numbers to dataset is about 20,710/6,837 ≈ 3.0291, which means that, in average there are approximately three annotators designated for one sample in the dataset. Similarly, this is also a class-imbalanced dataset. The overall percentage of abuse samples is about 20.4%, while the figure for CarbonBot is only 6.7%. Due to privacy reasons, the authors were not able to provide the dataset from Alana that was used in their paper (Curry et al., 2021). Only data from CarbonBot and ELIZA has been released. Specifically, we have 2501 samples as training data, 831 as validation data and 853 as testing data.

Table 3 Dataset size, labelled examples and the percentage of abuse across different datasets

Dataset	Size	Annotation	Abuse Percentage
Alana v2	2,652	7,942	27.2%
CarbonBot	1,515	4,650	6.7%
ELIZA	2,670	8,118	21.2%
Total	6,837	20,710	20.4%

An example of a data sample is given in Table 4.

Table 4 An example from the abusive conversation dataset

Column Name	Description	Example
example_no	Index of the record	278
annotator_id	Annotator ID	2
conv_id	Conversation ID	164070
prev_agent	Agent’s previous utterance	How long have you been on the verge of suicide?
prev_user	User’s previous utterance	since I was 10 years old
agent	Agent utterance	Why do you tell me you were 10 years old just now?
user	User (target) utterance	were you not ten years old?
bot	Agent name (CarbonBot/Eliza)	Eliza
Annotator1_is_abuse.1	Not abusive	1
Annotator1_is_abuse.0	Ambiguous	0
Annotator1_is_abuse.-1	Mildly abusive	0
Annotator1_is_abuse.-2	Strongly abusive	0
Annotator1_is_abuse.-3	Very strongly abusive	0
Annotator2_is_abuse.1	Not abusive	0
Annotator2_is_abuse.0	Ambiguous	1
……	……	……
Annotator8_is_abuse.-3	Very strongly abusive	/

As we can see from this table, one conversation consists of two rounds of dialogue between the user and agent. The annotations provided by one annotator are depicted across five columns, where they assign their chosen class as 1 and the others as 0. For each sample, there are annotations from at least three annotators, and for annotators not assigned to a particular sample, their entries remain blank.

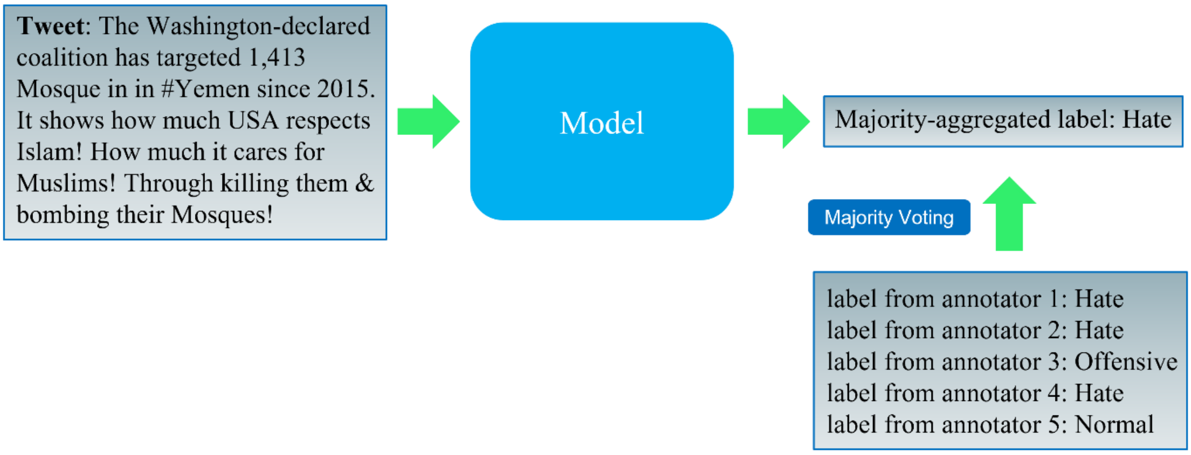
3.3 Chapter Summary

In this chapter, we mainly discussed our experimental datasets we intend to utilize. In this research, we have two different datasets, one is the hate speech dataset, and the other one is abusive conversation dataset. They were collected from different platforms and have different classification granularities: one is three-class classification, and the other is five-class classification. Furthermore, they have a huge discrepancy in terms of data size, which would be beneficial to testing our models' applications in different size of datasets.

4 Methodology

4.1 Baseline Model

In this study, our baseline model is trained on the "ground truth" label that is aggregated via majority voting. That is, the label assigned to a particular data point will be determined by the majority vote among annotators (see Figure 3). In cases of a tie situation, a dataset curator outside the initial annotator set determines the label as the final decision.



Tweet: The Washington-declared coalition has targeted 1,413 Mosque in in #Yemen since 2015. It shows how much USA respects Islam! How much it cares for Muslims! Through killing them & bombing their Mosques!

Majority-aggregated label: Hate

label from annotator 1: Hate
label from annotator 2: Hate
label from annotator 3: Offensive
label from annotator 4: Hate
label from annotator 5: Normal

Figure 3 The framework of the baseline model

Given BERT's notable performance in contextual understanding, we choose it as the pre-trained model. The anticipated output of this model should be a single label. Therefore, we augment its architecture by adding a fully connected layer to its last hidden state, thereby adapting the model structure to this specific prediction task. Within the model architecture, BERT functions as a feature extractor, extracting contextual information from the input text, while the fully connected layer appended to the last hidden state of BERT enables the model to learn task-specific patterns, relationships and knowledge.

4.2 Proposed Models

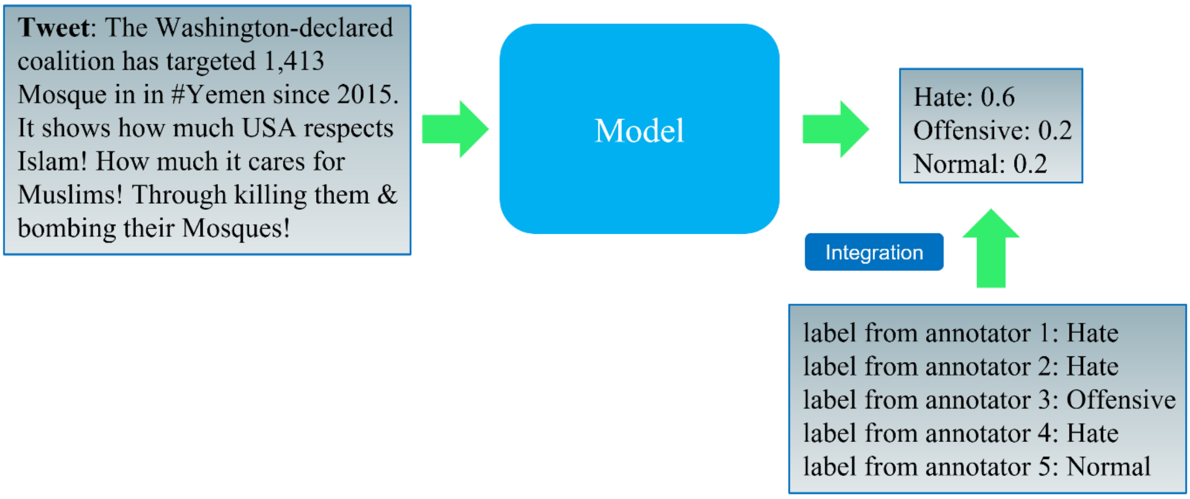
Although the baseline approach facilitates model training by providing a highly "precise" dataset with a clear and consistent target for supervised machine learning models to learn from, it can lose its meaning to a certain extent in some subjective tasks, such as sentiment analysis, hate speech identification. On one hand, majority vote-based harmonization tends to overlook minority perspectives, potentially leading to a loss of valuable nuances in the data. On the other hand, even if a "ground truth" label could be derived through harmonization, it might not serve as a robust basis for training a system aiming to reproduce annotators' judgements within a specific understanding task (Klenner et al., 2020). Instead, this approach can lead to brittleness or excessive generality, posing challenges in transferring annotated data across domains or limiting the practical applicability of the obtained results (Aroyo & Welty, 2013).

As a result, alternative methods that incorporate annotation disagreement into the model training process are proposed to improve model's robustness and capacity to generate diverse and inclusive predictions. The model becomes more informative by providing diverse predictions, rather than just adhering to one single assumed "ground truth" label. In our proposed methods, instead of training model on the majority-based label, annotations from all annotators are taken into account as valuable perspectives for model training.

Fornaciari et al. (2021) added multi-label training as an auxiliary task of the majority label training in a multi-task neural network. Inspired by this, we remove the single label training and propose probability-based multi-label model. Moreover, based on a similar idea employed by Fayek et al. (2016), we propose an ensemble system that consists of multiple sub-models, with each contributing to the final prediction. Instruction tuning shares the same idea with the ensemble system but employs a different foundation model and fine-tunes it through the use of natural language instructions. Notably, these existing studies incorporated multiple labels only to improve the performance of identifying the majority label. In this study, we use the probability distribution over labels as the model output, instead of aggregating it into a single label.

4.2.1 Probability-based multi-label method

The task of identifying hate speech or abusive conversation can be regarded as a multi-label text classification problem, where a given piece of text can be associated with one or multiple labels simultaneously. This paradigm is especially relevant in scenarios where content demonstrates diverse characteristics, and can be associated with various topics, themes, or attributes. In the probability-based multi-label method, the input is the text and the output is the probability distribution over the predefined label space. Unlike the traditional approaches that assign one or several exclusive labels to the input text (Jiang & Nachum, 2020), our model predicts the likelihood or probability of each label being associated with the given text. A framework of this approach is provided in Figure 4. As we can see from this figure, the model is trained on the probability distribution across different labels which is derived from individuals' annotations.



Tweet: The Washington-declared coalition has targeted 1,413 Mosque in in #Yemen since 2015. It shows how much USA respects Islam! How much it cares for Muslims! Through killing them & bombing their Mosques!

label from annotator 1: Hate
label from annotator 2: Hate
label from annotator 3: Offensive
label from annotator 4: Hate
label from annotator 5: Normal

Figure 4 The framework of model training within the probability-based multi-label method

In this approach, we mainly follow "pretrain then finetune" paradigm. Accordingly, since BERT is adept at generating contextualized word representations by considering the left and right context, we select it as the pre-trained model for training on the probability-based multi-label text classification task. During the fine-tuning process, an extra layer is added to the output layer of the pre-trained model. Specifically, the pre-trained model generates a vector as the presentation of the text in this phase (see Figure 5). This vector encapsulates the essential features and semantic information extracted from the text. Subsequently, this vector is fed into a fully-connected layer and mapped into a multidimensional vector, with each dimension corresponding to a distinct category associated with the text. Given that this study involves three or five types of label space, the output vector has three or five dimensions. Finally, in order to make sure values from all dimensions sum up to 1, this vector is normalized by the SoftMax function, thereby providing a probability distribution that reflects the model's confidence or certainty regarding each category.

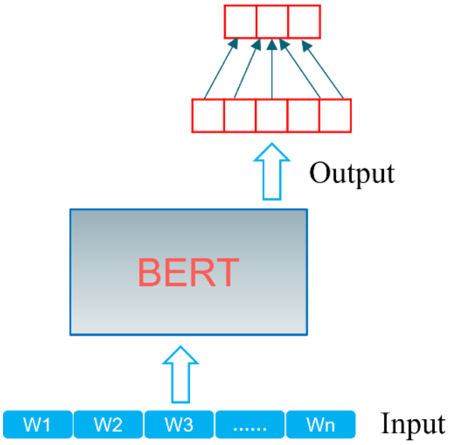


Figure 5 Fine-tuning BERT in the probability-based multi-label method

4.2.2 Ensemble system

In the process of label annotation, diverse labels are assigned by different annotators. And based on that, a probability distribution over different categories can be obtained by integrating all the annotations. Inspired by this idea, we propose the development of an ensemble system. The concept behind proposing an ensemble system for this task is to simulate the process of annotation by leveraging the diversity in labels provided by different annotators. Rather than relying on a singular annotator's perspective, this approach integrates the annotations from multiple annotators, thereby capturing the collective insights of all annotators.

The ensemble system consists of several sub-models. Each sub-model is trained independently on its respective set of labels. During the inference process, the final results concerning the probability distribution across different labels are obtained by combining outputs from all sub-models. Within the ensemble system, each sub-model contributes a unique perspective to the final predictions, collectively representing the diversity and nuances present in the annotations. In this way, it aims to achieve a more comprehensive understanding of the data by incorporating the varied viewpoints of multiple annotators.

In the abusive conversation dataset, the annotators assigned for each sample are clearly specified, as detailed in Table 4. Therefore, within the ensemble system, each sub-model represents one specific individual annotator and is trained on that annotator's provided labels. In the testing phase, these sub-models make their own predictions, which contribute to the final outcomes of the ensemble system. This can ensure a more comprehensive result by embracing the insights from each individual annotator, as represented by the sub-models.

In the hate speech dataset, each column of labels contains annotations from several anonymous annotators. Despite the anonymity, training model with such labels can potentially increase the robustness of sub-models since it helps to reduce the biases or inconsistencies introduced by individual annotator (Frenay & Verleysen, 2014). Furthermore, the resulting labels are likely to reflect a diverse range of perspectives and interpretations of the data. Training sub-models on these diverse annotations can capture the variability in annotator judgments and enhance the model's ability to generalize across different viewpoints (Audhkhasi & Narayanan, 2013). This collective wisdom pooled from multiple annotators has the potential to improve overall performance.

The foundation model used to train on individual annotations is the pre-trained BERT model. As we can see from Figure 6, each BERT model is trained individually on its corresponding labels. For each sub-model, the input is the text from one instance and the output is a multidimensional vector where each dimension corresponds to one category. After that, this vector is transformed by the SoftMax function and the dimension with highest probability is identified as the final output. Finally, the predictions from all sub-models are combined and converted into a probability distribution of three- or five-dimensional vector.

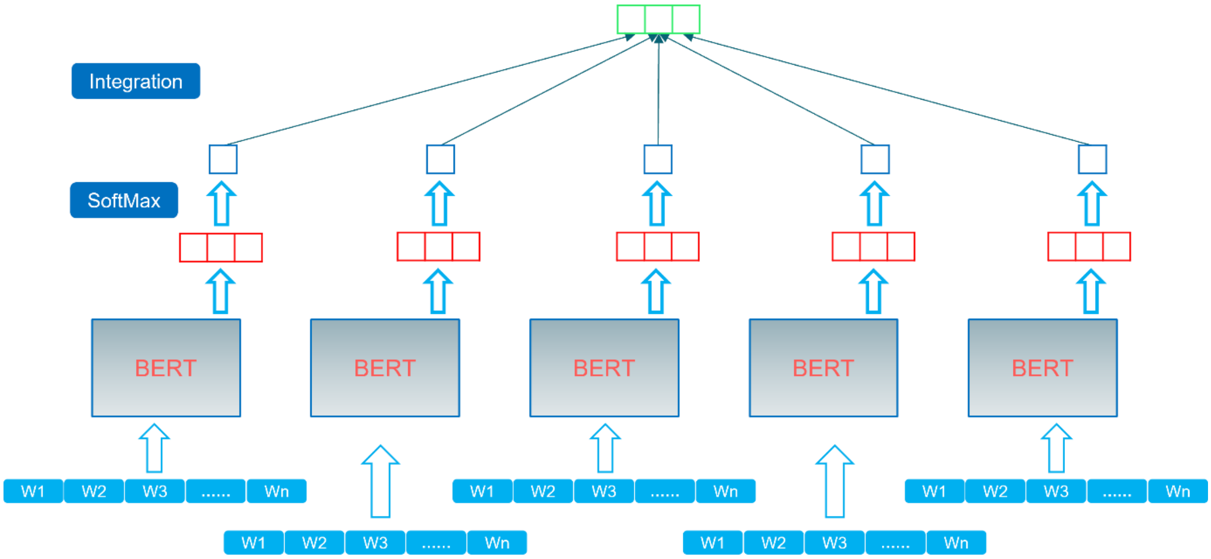


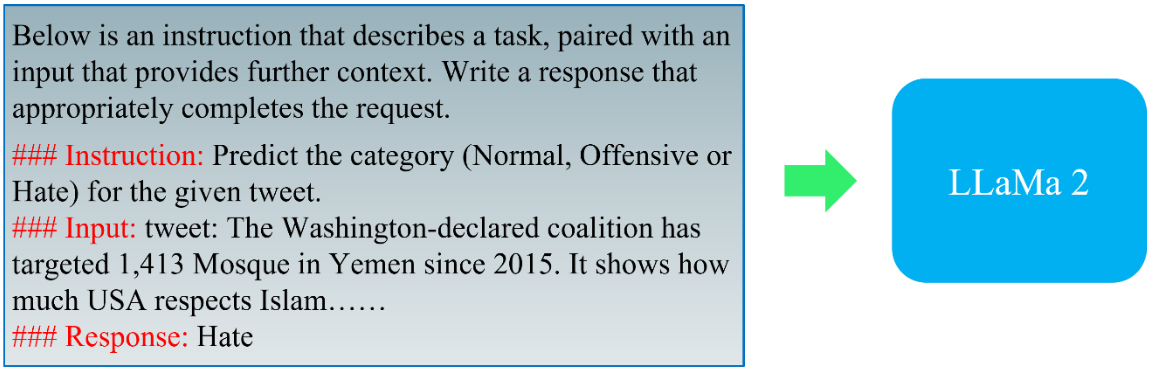
Figure 6 Fine-tuning BERT individually as sub-models within the ensemble system

4.2.3 Instruction tuning

Instruction tuning involves the process of further training LLMs on a dataset consisting of (instruction, output) pairs in a supervised fashion. Here, an "instruction" is constructed to guide the learning process. The key idea is to provide the model with explicit instructions, often in the form of paired input-output examples, to enhance its performance and align it with specific objectives. Unlike traditional training approaches where models learn from data alone, instruction tuning injects explicit guidance into the training process. This approach allows for explicit customization of the model's behavior. Researchers can provide specific guidance to shape the model's output, thereby adhering to desired properties or behaviors. In other words, this guidance comes in the form of instructions, which specify desired properties, behaviors, or characteristics that the model should exhibit in its outputs. In this study, we ask the model to predict the class of hate speech or abusive conversation based on the input we construct. The input contains task description, instruction, original text, and response, which is the annotation from a specific annotator. In this context, we need to design a template. This template can transform (text, label) pairs from existing annotated datasets to (instruction, output) pairs. In many cases, the construction of template typically involves two steps: firstly, manually composition of instruction and target templates; secondly, filling templates with data instances from the dataset (Zhang et al., 2023).

The advantages of instruction tuning are as follows. On one hand, instruction tuning enables the explicit incorporation of contextual information and constraints, ensuring that the model generates outputs that align with specific contextual requirements. In contrast, traditional supervised learning may struggle to implicitly capture or handle complex contextual dependencies. On the other hand, instruction tuning allows for few-shot learning. In other words, it enables the model to learn from a few examples or even just one, which makes it suitable for many few-shot learning scenarios. This facilitates training models on tasks where real-world labelled data is limited or expensive to obtain. By contrast, in traditional supervised learning, adverse effects may arise due to the lack of sufficient training data since it typically requires a large amount of dataset from which the model can learn the knowledge or pattern as precisely as possible (Zhou et al., 2017).

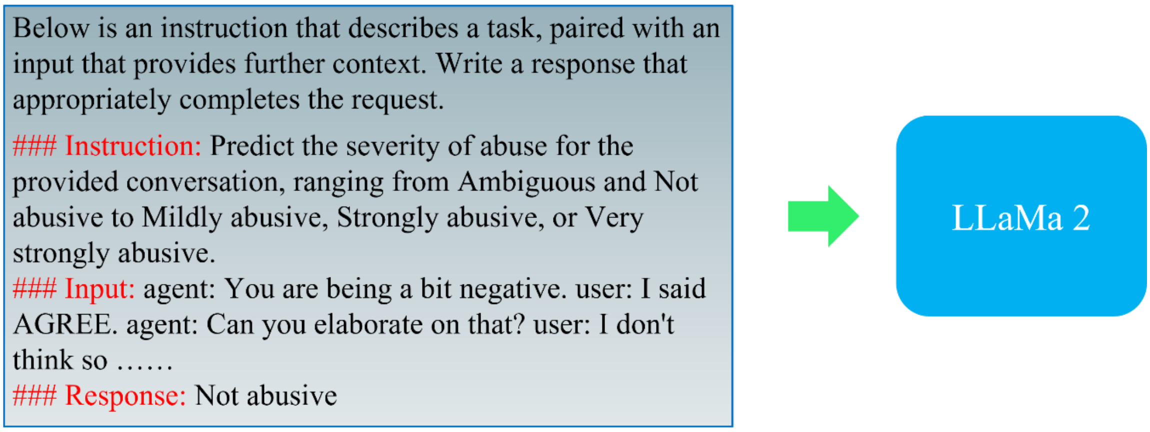
In the development of instruction tuning, numerous pre-trained LLMs, such as LLaMa 2, T5 and Vicuna can be employed. LLaMA 2 was pre-trained on an extensive dataset that encompasses a wide range of text and code, including books, articles, websites, and programming code. The dataset was compiled from various sources, including the Common Crawl corpus, the English Wikipedia, and GitHub. In the process of pre-training, the model was trained on a variety of NLP tasks, such as masked language modelling, text classification, and code generation. This multifaceted training equipped LLaMA 2 with the ability to learn a wide range of linguistic patterns and relationships. Operating as an auto-regressive language model, LLaMA 2 employs an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) (*Introducing ChatGPT*, n.d.; Ouyang et al., 2022) to align to human preferences, specifically targeting criteria related to helpfulness and safety[6]. Available through the Hugging Face platform under this repository, LLaMA 2 is configurable with parameters ranging from 7 billion, 13 billion, to 70 billion, and it is offered in both pre-trained and fine-tuned variations. Although Vicuna is an upgraded version of LLaMa, it has been particularly fine-tuned with a ChatGPT dialogue corpus, which enhances its performance in dialogue-related tasks and chatbot interaction (Xu et al., 2023). These capabilities are not directly aligned with the objectives of this study. Therefore, we will utilize LLaMa 2 (7 billion version) as the foundation model. LLaMa 2 is an auto-regressive language model, which means it generates text one word at a time, predicting the next word based on the words that came before it. As a result, different from the earlier models we have proposed, in which this task is treated as text classification problem, this approach addresses it as a question answering task. In this context, the model is trained to generate a response when given a designed input.



Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Predict the category (Normal, Offensive or Hate) for the given tweet.
Input: tweet: The Washington-declared coalition has targeted 1,413 Mosque in Yemen since 2015. It shows how much USA respects Islam……
Response: Hate

Figure 7 Fine-tuning LLaMa 2 as a sub-model with instruction tuning in the hate speech dataset



Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Predict the severity of abuse for the provided conversation, ranging from Ambiguous and Not abusive to Mildly abusive, Strongly abusive, or Very strongly abusive.
Input: agent: You are being a bit negative. user: I said AGREE. agent: Can you elaborate on that? user: I don't think so ……
Response: Not abusive

Figure 8 Fine-tuning LLaMa 2 as a sub-model with instruction tuning in the abusive conversation dataset

[6] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

The frameworks of fine-tuning LLaMa 2 via instruction tuning in two datasets are presented in Figure 7 and Figure 8. On the left sides of the figures are the inputs fed into the pre-trained model. From the figures, the instruction tuning process involves providing the pre-trained model with an input, which comprises the following four components: scenario description, instruction, text input and response. The instruction indicates the desired task, the text input comprises the tweet or conversation and the response is the annotation from a specific annotator. With this input format, the model is fine-tuned to become adept in the downstream task. After several attempts and checking the quality of model outputs, different instructions are allocated for hate speech detection and abusive conversation detection respectively. Following the fine-tuning process, the model gains the ability to predict the corresponding label assigned by a particular annotator. In the inference phase, the value under the "Response" key will be removed, and the fine-tuned model is tasked with generating its prediction.

However, due to the large parameter size of the pre-trained model, it would be very time-consuming to fine-tune the entire model throughout the whole training process. As a result, PEFT strategies will be applied to mitigate training time and memory requirements. The available strategies include LoRA, QLoRA, P-Tuning, Prefix-Tuning, and others, which have proven prominent in fine-tuning. Notably, LoRA and QLoRA are the most widely utilized techniques. Although QLoRA proves more memory-efficient and powerful in making the training process faster, it potentially leads to lower performance compared with LoRA, as it uses lower-precision weights to compress the model size and conserve memory. Considering that the selected version of LLaMa 2 (seven billion) lies within the acceptable limits of our memory and computational power, we employ LoRA as the PEFT method for instruction tuning. This allows for fine-tuning model by only optimizing a small number of parameters. The details of applying LoRA for PEFT is given in Figure 9. In the figure, the pre-trained weights are frozen and an adaptor is added alongside it. The adaptor is decomposed into two smaller matrices where r is much smaller than d and k. During fine-tuning, only the parameters in the adaptor are updated. As a result, the computations are reduced dramatically. In the process of inference, the matrices A and B will be multiplied and merged with pre-trained model, which ensures that the parameter size of newly fine-tuned model will stay unchanged.

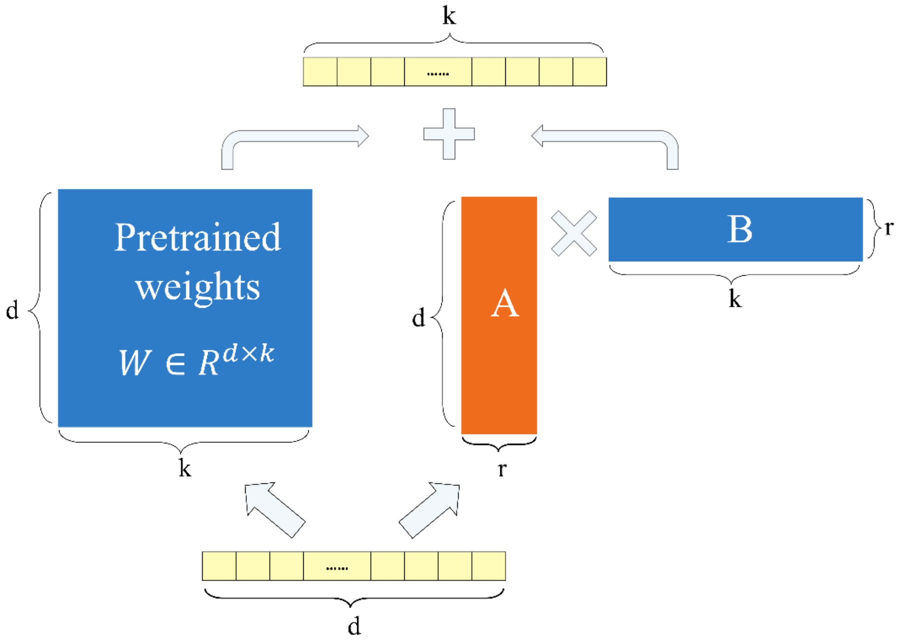


Figure 9 Details of using LoRA for model's parameter efficient fine-tuning

4.3 Evaluation Metrics

4.3.1 Regular metrics

The baseline model and sub-models within the ensemble system and instruction tuning method are trained using single labels. Therefore, we utilize precision, recall, accuracy and F1-score (Salton & Lesk, 1968) to evaluate the performance of these models. The ways of calculating these metrics are present in the formulas below. Given that there are several different classes and the dataset is class-imbalanced, we specifically present confusion matrix (Townsend, 1971) to elucidate the performance on each class.

$$Precision = \frac{True\ Positive}{True\ Postive + False\ Positive} \tag{2}$$

$$Recall = \frac{True\ Positive}{True\ Postive + False\ Negative} \tag{3}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Postive + True\ Negative + False\ Positive + False\ Negative} \tag{4}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

On the other hand, the final output of our proposed models is a probability distribution across different labels, and in this scenario, a single “ground truth” label is no longer applicable for model evaluation. Instead, we need to compare the distribution of annotations with model output. Here, cross entropy is used to measure the difference between probability distributions. Cross entropy is one kind of statistical distance which can be used to measure how a probability distribution is different from a reference probability distribution. By definition, it is the average number of bits (a format of quantified information) required to encode data coming from a target distribution A when using the approximation of the target distribution B (Murphy, 2012). For two discrete probability distributions A and B which have the same class space χ, the way of calculating the cross entropy between A and B is shown in formula (6) (Shannon, 1948). The higher the cross entropy is, the more different two distributions are. This metric is commonly used as the loss function for training neural networks. However, in previous research on NLP tasks, it has also been used to quantify how well the model’s predicted distribution matches the annotation distribution over possible categories from multiple annotators (Pavlick & Kwiatkowski, 2019).

$$H(A, B) = -\sum_{x \in \chi} a(x) \log b(x) \tag{6}$$

4.3.2 Alternative metric

Furthermore, in RQ3, we aim to evaluate the effectiveness of training model with multiple labels against the model that only relies on the majority label. Due to the format disparity between the outputs generated by these two models, it is impractical to use the abovementioned metrics for evaluation. To bridge this gap, we conduct an online survey where participants specify their preference between annotations generated from the probability-based multi-label model and the baseline model. For each dataset, we select 10 samples, each featuring two groups of annotations. In the first group, there are two annotations. One is the multiple labels from all the annotators, and the other one is the “ground truth” label derived from these multiple labels via majority voting. There are also two annotations in the second group. Both annotations are in the form of probability distributions across different labels. One is generated from the baseline model trained with majority label, which is, however, used to generate probability distribution in the phase of inference. The other one is from the probability-based multi-label model. This model has the same structure as the baseline model and their only difference is the labels in the data they were trained on. For each sample, participants are required to indicate which annotation they find is more reasonable to characterize the tweet or the abusive conversation in each of these two groups of annotations. The first group of annotations is to explore individuals' preferences regarding the labelling method for hate speech or abusive conversation. Specifically, we aim to understand whether individuals favor using a majority label or multiple labels. The second group aims to investigate whether training the model with the multiple labels improves performance.

In terms of the sampling strategy for the online survey, we incorporated samples with diverse annotation patterns. They, on one hand, have both diverse and concentrated annotations. For example, in hate speech part, three samples are relatively concentrated, with all the five annotators assigning the same label, while three others contain disperse annotations. On the other hand, we ensure the diversity of the majority labels across the samples.

On the visualization aspect, we use a different way for each group. The first group is presented via word cloud since we aim to display the multiple labels and majority label. The size of each word in the word cloud indicates the likelihood that the tweet belongs to a specific label. As for the second group, we visualize two distributions via stacked bars where the length of each section in the stacked bar indicates the probability value. For the specific details of the questionnaire, please refer to Appendix B.

In terms of analyzing the collected data, we have two goals. Firstly, we need to explore in general if individuals prefer multiple labels over majority label, and if they prefer the probability distribution generated from the probability-based multi-label model compared to the baseline model. Secondly, we seek to investigate the correlations between these two preferences and various demographic factors, such as gender, degree, ethnicity and familiarity with hate speech. To achieve the first objective, we employ binomial test and multinomial test to discover if people exhibit preference for certain option. The binomial test is a statistical hypothesis test used to assess if the proportion of observations in a sample deviate from a hypothesized value (Fisher, 1992). It is suitable for investigating labelling method preference since there are two options available. Moreover, this test is particularly effective for small to moderate sample sizes. However, when dealing with very large samples, even slight changes from the hypothesized proportion can yield statistically significant outcomes, which might not be very meaningful when put into practice. The multinomial test builds on the basis of binomial test since it handles scenarios with more than two possible outcomes (Read & Cressie, 2012). They are both valuable in analyzing categorical data and have wide applications such as market research, where they help to understand customer preferences across various product features or categories. As for the second objective, we adopt chi-square test for assessing the relationship between these two preferences and individuals' demographic factors. The chi-square statistic examines the independence between variables by quantifying the disparity between the observed and expected frequencies of outcomes within a specific variable set (William G. Cochran, 1952). It is commonly applied to categorical data, especially when dealing with nominal variables, such as marital status or gender, where the order of categories does not matter (Yeager, n.d.).

4.4 Chapter Summary

In this chapter, we mainly introduced the methodology that is used in this research. To begin with, we build the text classification models that embrace the annotation disagreement in different ways: the probability-based multi-label method, the ensemble system and instruction tuning method. Regarding the evaluation metrics, the common ones, such as precision, recall and F1-score will be applied on the baseline model and sub-models within the ensemble system and instruction tuning, as their targets are single labels. For the three proposed models which output probability distributions, cross entropy loss will be utilized. Additionally, in order to evaluate the performance of incorporating the multiple annotations for model training, we conduct an online survey. This survey contains two parts: the first part is to investigate individuals' preferences regarding the selection of either a majority label or multiple labels, as well as their preferences between two probability distributions, for labelling the hate speech or abusive conversation. One of these distributions is generated by the multi-label model, while the other one by the baseline model. The second part is used to investigate the correlations between these two preferences with demographic factors, such as degree, gender, etc.

5 Experiments and Results

The experiments in this study involve training models that can predict targets in different scenarios, including those with multiple labels or single labels. Our primary methodology follows the "pretrain then finetune" paradigm, which selects a model pre-trained on a large dataset, and subsequently, fine-tunes this model to tailor its knowledge for our specific tasks.

5.1 Baseline Model

The details of training configurations are given in Table 5.

Table 5 The training configurations of the baseline model

Hyperparameter	Value
Batch size	64
Learning rate	3e-5
Epoch	50
Dropout	0.3

5.1.1 Hate speech detection

The results of the baseline experiment for the hate speech dataset are shown in Table 6 and Table 7. From these tables, the baseline model achieves an accuracy of 0.7456 on the testing dataset. There exists significant discrepancy in performance across the three classes. Specifically, class "Normal" demonstrates the highest performance, attaining an F1-score of 0.8452. By contrast, the results for the other two classes are less satisfactory, especially for class "Hate", where the F1-score is only 0.3002. As shown in Table 2, both the "Offensive" and "Hate" classes are minority classes within the dataset, and this class-imbalance issue is the main reason why the model shows suboptimal performance on these classes. In both traditional machine learning and deep learning algorithms, the model might become biased towards the majority class if it is trained on a class-imbalanced dataset (Japkowicz & Stephen, 2002; Mazurowski et al., 2008). To minimize training loss and optimize fitting, the model may prioritize learning patterns from the majority class, potentially neglecting the minority classes.

Table 6 The loss and accuracy of the baseline model for hate speech detection

Training		Validation		Testing	
Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
0.6804	0.7243	0.6232	0.7482	0.6230	0.7456

Table 7 The baseline model's performance across classes on the testing data for hate speech detection

	Precision	Recall	F1-score
Normal	0.7906	0.9081	0.8452
Offensive	0.6364	0.5560	0.5935
Hate	0.5595	0.2051	0.3002
Macro avg	0.6621	0.5564	0.5796
Weighted avg	0.7266	0.7456	0.7251

5.1.2 Abuse detection in conversational AI

For the abusive conversation dataset, the results are shown in Table 8 and Table 9. According to Table 8, the baseline model demonstrates an accuracy of 0.8499, which is significantly higher than its accuracy on the hate speech dataset. Similarly, substantial performance disparities are evident across various classes. As a majority class, “Not abusive” achieves a remarkably high F1-score of 0.9447. By comparison, the model struggles to make correct predictions for “Ambiguous” class.

Table 8 The loss and accuracy of the baseline model for abusive conversation detection

Training		**Validation**		**Testing**	
Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
0.1459	0.9604	0.6451	0.8700	0.6997	0.8499

Table 9 The baseline model’s performance across classes on the testing data for abusive conversation detection

	Precision	**Recall**	**F1-score**
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9413	0.9481	0.9447
Mildly abusive	0.3556	0.2424	0.2883
Strongly abusive	0.5455	0.7619	0.6358
Very strongly abusive	0.2727	0.1875	0.2222
Macro avg	0.4230	0.4280	0.4182
Weighted avg	0.8388	0.8499	0.8421

5.2 Probability-based Multi-label Model

In the development of the probability-based multi-label model, we also leverage BERT as the pre-trained model. Further details on the training configurations can be found in Table 10.

Table 10 The training configurations of the probability-based multi-label model

Hyperparameter	**Value**
Batch size	32
Learning rate	2e-5
Epoch	50
Dropout	0.3
Optimizer	AdamW

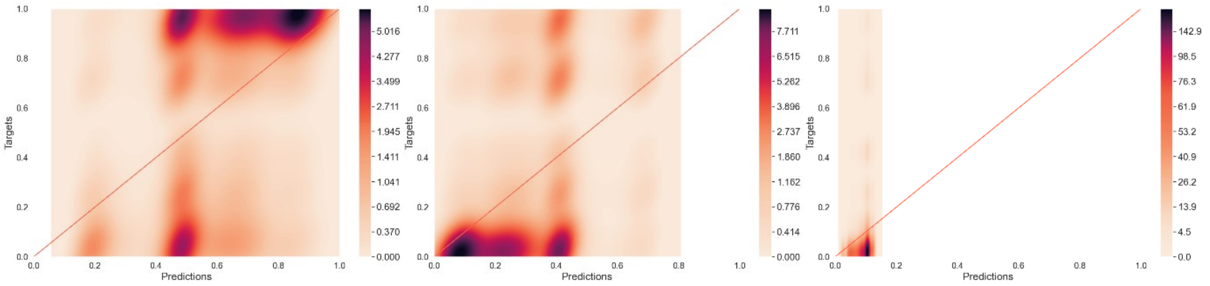
5.2.1 Hate speech detection

Table 11 The average cross entropy for hate speech detection from the probability-based multi-label model

	Training	**Validation**	**Testing**
Hate speech	0.7613	0.7569	0.7638

The model’s performance is evaluated by computing the cross entropy between the probability distributions generated by the multi-label model and annotators. The results of model training, validation and testing on this dataset are summarized in Table 11. Specifically, the model achieves a cross entropy loss of 0.7638 in the testing phase.

However, cross entropy, being a general metric, provides insights only into the overall performance of the model across the entire testing dataset. To gain a more nuanced understanding, we proceed to visualize the predictions and target values across different classes. For each distribution, we dissect it into three numbers which represent the probability prediction on each class. Then, values in the same dimensions from the prediction and target distributions are regarded as x-axis and y-axis coordinates, thereby forming points in the plot. Figure 10 shows the heatmaps of points from three distinct classes in the hate speech dataset, color-coded based on their density. We also plot the “y=x” line (identity line), and points near this line indicate the predictions and targets have close values.



Normal Offensive Hate

Figure 10 Density-based heatmaps of points across different classes for hate speech detection (the probability-based multi-label method)

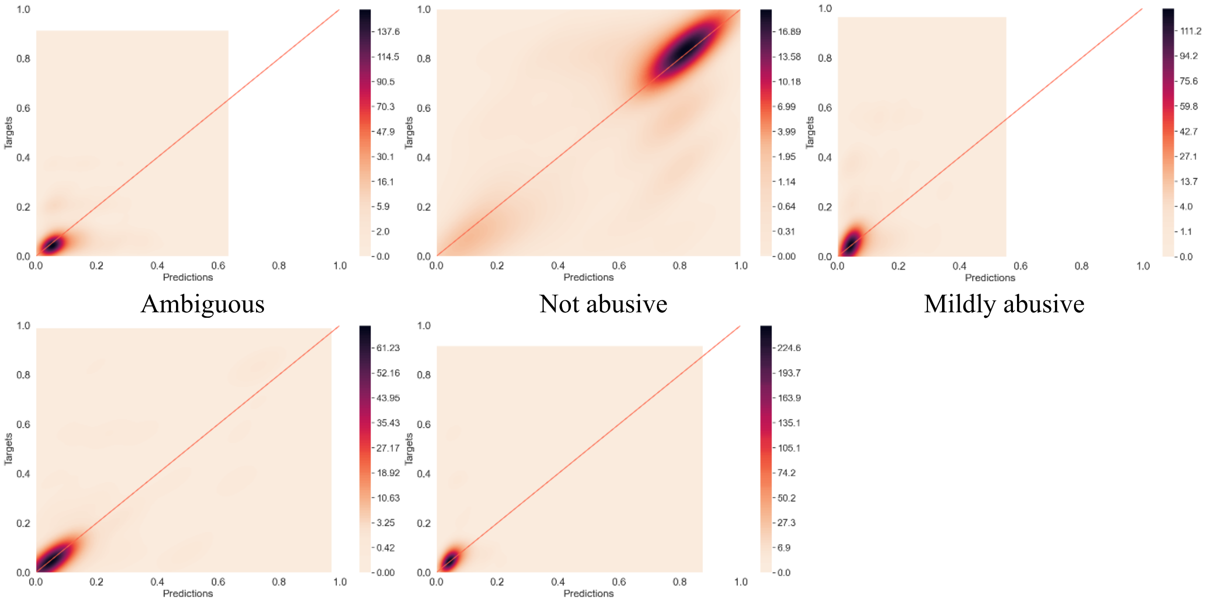
From this figure, the results for classes “Normal” and “Offensive” are less satisfactory. There are numerous points where the model’s predictions and targets values exhibit misalignment. By contrast, the model’s performance is robust in the “Hate” class, where most of the points stay close to origin of coordinate plane (0,0).

5.2.2 Abuse detection in conversational AI

Table 12 The average cross entropy for abusive conversation detection from the probability-based multi-label model

	Training	**Validation**	**Testing**
Abusive conversation	0.8861	0.9680	0.9834

From Table 12, the model achieves a cross entropy loss of 0.9834 in the testing phase. It turns out that the multi-label model achieves a more promising result on the hate speech dataset.



Ambiguous Not abusive Mildly abusive

Strongly abusive Very strongly abusive

Figure 11 Density-based heatmaps of points across different classes for the abusive conversation dataset (the probability-based multi-label method)

Figure 11 shows the heatmaps depicting points from five distinct classes in the abusive conversation dataset. Although the overall cross entropy loss is higher compared with the ensemble system, the results for all classes are relatively satisfactory. There is not a specific area from these plots that indicates a substantial number of points significantly deviating from the identity line.

5.3 Ensemble System

The ensemble system consists of multiple BERTs which function as sub-models. All the BERT models within the ensemble system share the same training configurations, which are outlined in Table 13.

Table 13 The training configurations of sub-models in the ensemble system

Hyperparameter	**Value**
Batch size	64
Learning rate	3e-5
Epoch	50
Dropout	0.3

5.3.1 Hate speech detection

Table 14 The loss and accuracy of sub-models in the ensemble system for hate speech detection

Model	**Training**		**Validation**		**Testing**	
	Loss	**Accuracy**	**Loss**	**Accuracy**	**Loss**	**Accuracy**
Sub-model 1	0.6614	0.7306	0.6294	0.7380	0.6285	0.7367
Sub-model 2	0.6830	0.7140	0.6452	0.7238	0.6605	0.7154
Sub-model 3	0.6769	0.7159	0.6377	0.7251	0.6494	0.7237
Sub-model 4	0.7003	0.7037	0.6541	0.7165	0.6754	0.7057
Sub-model 5	0.6838	0.7095	0.6533	0.7193	0.6551	0.7166

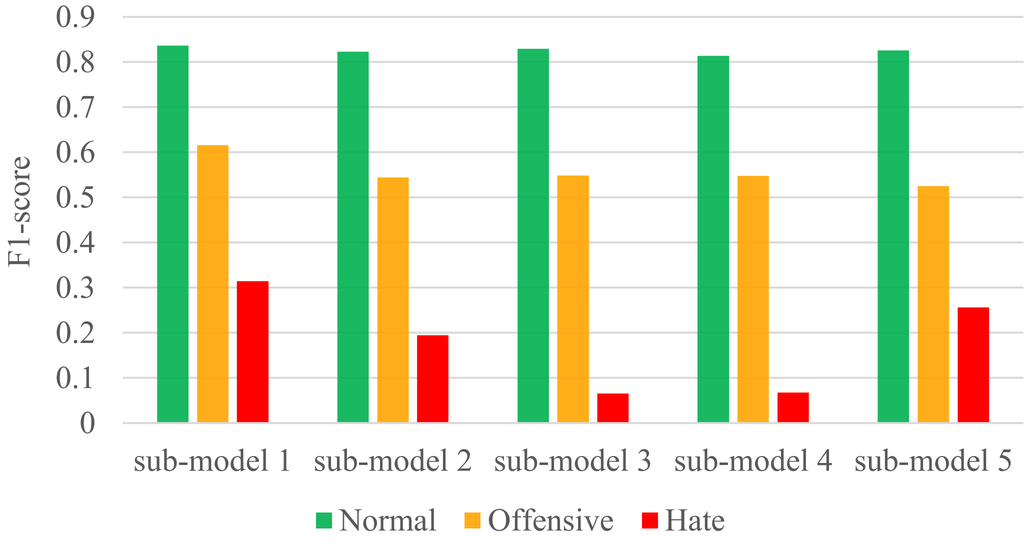


Figure 12 The F1-score of sub-models across classes in the ensemble system on the testing data for hate speech detection

The results of the sub-models on the hate speech dataset are given in Table 14. As we can see from this table, sub-model 1 stands out with the highest accuracy on the testing data at 0.7367. By contrast, sub-model 4 records the lowest accuracy, reaching only 0.7057. The remaining three exhibit similar performance, achieving accuracies of about 0.72. The detailed performance results of the sub-models across different classes on the testing data are shown from Table 36 to Table 40 in Appendix C. In Figure 12, we present their F1-score for comparison. There also exists pronounced discrepancy in the model's ability to accurately predict instances in the minority class "Hate" across five sub-models, with sub-models 3 and 4 displaying a particularly noticeable disparity. Specifically, the F1-score of class "Hate" achieved by these two sub-models are only around 0.06. In contrast, among all sub-models, their precision, recall and F1-score on classes “Normal” and “Offensive” are not very different. The consistency suggests that the model perform comparably well on the majority classes, highlighting their ability to effectively capture and classify samples within these prevalent categories.

Subsequently, the predictions from sub-models are combined and transformed into a probability distribution via the SoftMax function. Table 15 illustrates this process, which is also employed for annotations from multiple annotators.

Table 15 The method of combining and transforming predictions from the sub-models

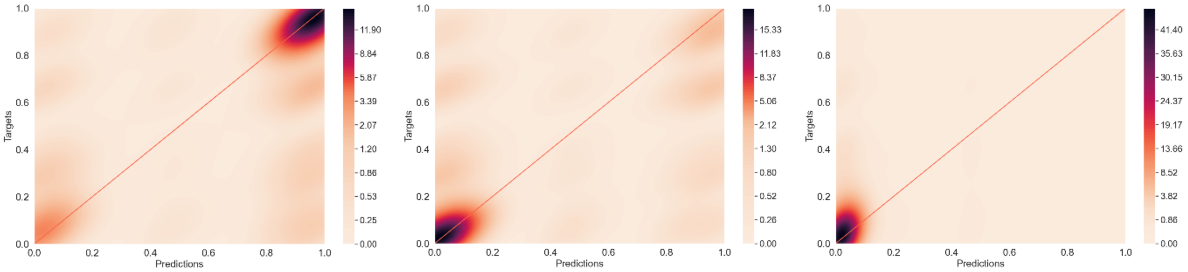
Predictions	Combined Result	Probability Distribution
Sub-model 1: 0 Sub-model 2: 2 Sub-model 3: 1 Sub-model 4: 2 Sub-model 5: 0	[2, 1, 2]	[0.4223, 0.1554, 0.4223]

Lastly, since these sub-models show varying performances in the training and validation processes, we typically choose top n (n≥3) best-performing sub-models to determine the final results. The ranking criteria is based on their accuracies in the validation data. The predictions from these top n sub-models are combined and transformed into probability distributions. Table 16 provides the average cross entropy for the testing data in the hate speech dataset using different top n sub-models.

Table 16 The average cross entropy on the testing data for hate speech detection with different top n sub-models (the ensemble system)

Top_n sub-models	Top 3	Top 4	Top 5
Cross entropy	0.9734	0.9720	1.0456

From the table, the lowest cross entropy is achieved by the top 4 sub-models (0.9720), which is worse than the multi-label model’s result of 0.7638 on the same dataset.



Normal Offensive Hate

Figure 13 Density-based heatmaps of points across different classes for hate speech detection (the ensemble system)

Figure 13 shows the heatmaps depicting points from three distinct classes in the hate speech dataset. As illustrated in this figure, the model’s performance in the “Hate” class surpasses the other two classes, in which there are still some points staying away from the identity line. Even though the overall performance is worse than the multi-label method in this dataset, the ensemble system’s results for classes “Normal” and “Offensive” are more promising. As minority classes, most of the points from the “Offensive” and “Hate” classes are concentrated close to (0,0). By contrast, we can find many points near (1,1) in the “Normal” plot.

5.3.2 Abuse detection in conversational AI

The results of the sub-models on the abusive conversation dataset are given in Table 17. From this table, the overall performance of the sub-models in the ensemble system is significantly better compared with the hate speech dataset (see Table 14). Notably, sub-model 3 and sub-model 7 achieve the highest accuracy (around 0.88). The detailed performance of the sub-models across different classes on the testing data is shown from Table 41 and Table 48 in Appendix D. Here, we present their F1-score in Figure 14 for comparison. Although the performance of sub-model 5 (0.6548 of accuracy) is obviously worse than other sub-models, it excels in predicting the minority classes, such as class "Very strongly abusive". The remaining sub-models consistently demonstrate good performance, with accuracy on testing data above 0.8.

Table 17 The loss and accuracy of sub-models in the ensemble system for abusive conversation detection

Model	Training		Validation		Testing	
	Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
Sub-model 1	0.1923	0.9433	0.6221	0.8594	0.8415	0.8232
Sub-model 2	0.0351	0.9892	0.6204	0.8885	1.0750	0.8324
Sub-model 3	0.0336	0.9905	0.5750	0.9023	0.6427	0.8792
Sub-model 4	0.1281	0.9587	0.6657	0.8853	0.9102	0.8701
Sub-model 5	0.0045	0.9980	1.9568	0.7238	2.2268	0.6548
Sub-model 6	0.0028	1.0000	0.8896	0.8840	1.1497	0.8228
Sub-model 7	0.3277	0.8916	0.4274	0.8937	0.4289	0.8799
Sub-model 8	0.0002	1.0000	1.2371	0.8557	1.3783	0.8168

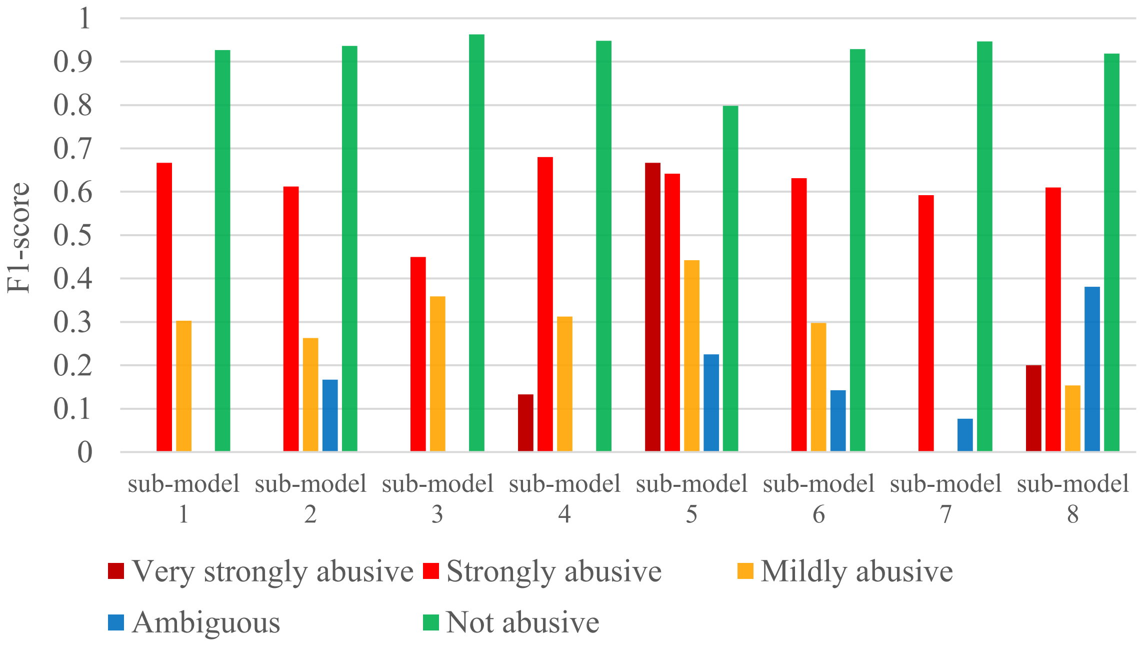


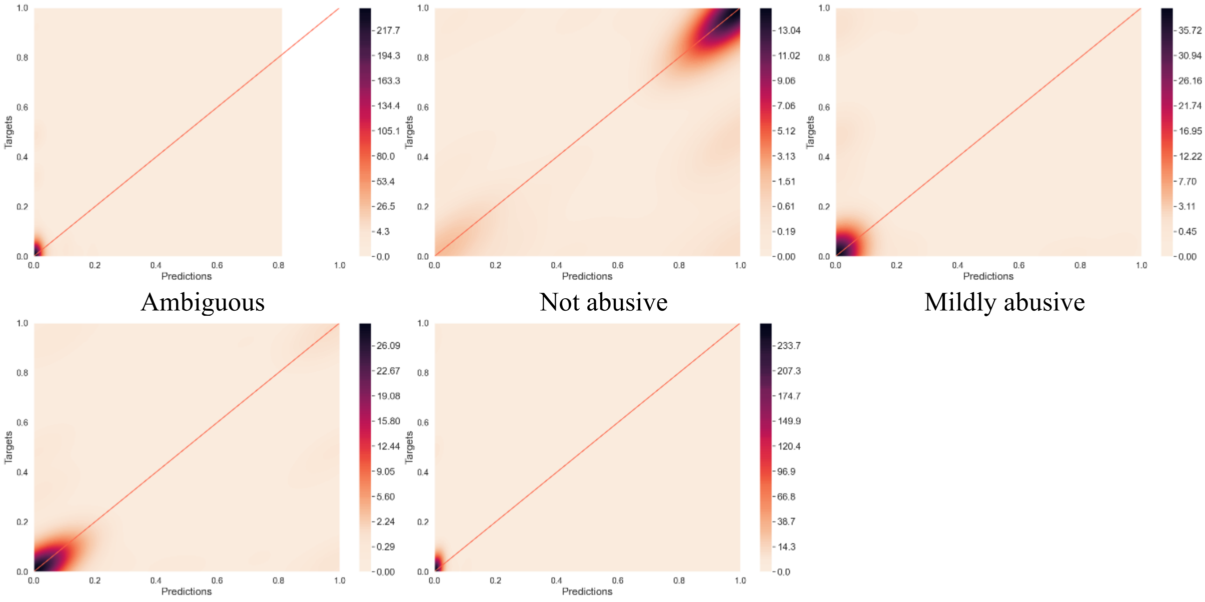
Figure 14 The F1-score of sub-models across classes in the ensemble system on the testing data for abusive conversation detection

Similarly, the prediction outcomes from the sub-models are combined and transformed into probability distributions. The resulting cross entropy is calculated and presented in Table 18. In this dataset, the best outcome is observed with the top 8 sub-models, yielding an overall cross entropy of 0.6782. This stands in stark contrast to the ensemble system on the hate speech dataset, where the best result is 0.9720 (as indicated in Table 16), and the probability-based multi-label model on the abusive conversation dataset, which stands at 0.9834 (as indicated in Table 11).

These findings provide valuable insights into the strengths of the multi-label model over the ensemble system on the hate speech dataset. Conversely, in the abusive conversation dataset, the predictions from ensemble system align with the target values better than the multi-label model.

Table 18 The average cross entropy on the testing data for abusive conversation detection with different top n sub-models (the ensemble system)

Top n sub-models	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8
Cross entropy	1.0302	0.8306	0.7223	0.6946	0.7065	0.6782



Ambiguous

Not abusive

Mildly abusive

Strongly abusive

Very strongly abusive

Figure 15 Density-based heatmaps of points across different classes for abusive conversation detection (the ensemble system)

Figure 15 shows the heatmaps depicting points from five distinct classes in the abusive conversation dataset with the ensemble system. From the figure, the overall model performance surpasses that of the hate speech dataset since most of points are close to “y=x” line when compared with Figure 13. In particular, the model excels in predicting instances within “Ambiguous” and “Very strongly abusive” classes, where the points are relatively close to the identity line. It is also interesting to note that the majority of the points from these two classes are located near the origin of coordinate (0,0). As the minority classes in the dataset, their corresponding dimensions in the probability distribution are often predicted as 0 by the model. By contrast, the points from “Not abusive”, “Mildly abusive” and “Strongly abusive” classes demonstrate greater dispersion from the identity line.

5.4 Instruction Tuning

Table 19 The training configurations of instruction tuning with LoRA

Hyperparameter	**Value**
Rank	8
Target modules	[q_proj, v_proj]
Batch size	32
Learning rate	3e-4
Epoch	50
Dropout	0.05
Optimizer	AdamW

In the development of the instruction tuning, we employ the pre-trained LLaMa 2 as the sub-model. During the process of fine-tuning, all sub-models share the same training configurations. The details are given in Table 19.

5.4.1 Hate speech detection

The sub-models' accuracies on the testing data for the hate speech dataset are given in Table 20. From this table, sub-models 1 and 5 achieve the highest accuracies (at 0.6429 and 0.6443 respectively). By contrast, sub-model 3 achieves the lowest accuracy at 0.5783. Sub-models 2 and 4 perform slightly better, with accuracies just above 0.6. Overall, the performances of these sub-models are worse than those within the ensemble system. The accuracies in Table 14 are basically around 0.7, with approximately margin of 0.1, and this is mainly due to the inferior performance on the "Normal" class in instruction tuning. For the detailed results of the sub-models in instruction tuning across classes on this dataset, please refer to Table 49 to Table 53 in Appendix E.

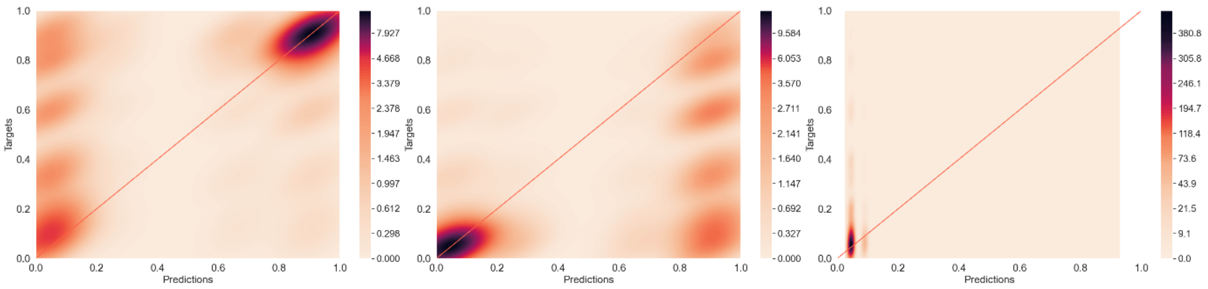
Table 20 The accuracy of sub-models in instruction tuning on the testing data for hate speech detection

Model	**Accuracy**
Sub-model 1	0.6429
Sub-model 2	0.6056
Sub-model 3	0.5783
Sub-model 4	0.6293
Sub-model 5	0.6443

Then, prediction outcomes are combined and transformed into probability distributions, from which cross entropy is calculated. The best result is achieved by the top 3 sub-models (1.2445). By comparing this result with those from the multi-label method (0.7638) and the ensemble system (0.9720~1.0456), it is evident that, for the hate speech dataset, instruction tuning yields the least favorable performance among these three methods.

Table 21 The average cross entropy on the testing data for hate speech detection with different top n sub-models (instruction tuning)

Top_n sub-models	Top 3	Top 4	Top 5
Cross entropy	1.2445	1.4060	1.6313



Normal Offensive Hate

Figure 16 Density-based heatmaps of points across different classes for hate speech detection (instruction tuning)

The heatmaps, depicting the density of points across different classes are presented in Figure 16. Like other models, the outcome for the "Hate" class stands out, with most of the points situated closely to the identity line. In contrast, the results for the "Normal" and "Offensive" classes are less promising, as many points within these two classes deviate significantly from the identity line.

5.4.2 Abuse detection in conversational AI

We record both the training loss and validation loss at the end of each epoch. The specific details on the training loss and validation loss throughout the training process are shown from Figure 26 to Figure 33 in Appendix F. From these figures, the training loss goes down consistently as the epoch increases, which indicates a continuous improvement in the model's fit to the training dataset. By contrast, in the beginning of the training process, the validation losses decline quickly, then plateau for several epochs, and finally experience a slight increase towards the end of training. This pattern suggests that the model overfit the training data and therefore performs worse on the validation data. In order to mitigate overfitting, we hereby employ early stopping strategy. That is, we select the model with the lowest validation loss as the final fine-tuned model.

Table 22 The accuracy of sub-models in instruction tuning on the testing data for abusive conversation detection

Model	**Accuracy**
Sub-model 1	0.8283
Sub-model 2	0.8237
Sub-model 3	0.8399
Sub-model 4	0.7853
Sub-model 5	0.6726
Sub-model 6	0.8259
Sub-model 7	0.8769
Sub-model 8	0.8258

The accuracies of sub-models on the testing dataset for abusive conversation detection are given in Table 22. Apart from sub-model 4 and sub-model 5, the accuracies for the other models are higher than 0.8. Notably, sub-model 7 yields the highest accuracy (0.8769). By contrast, sub-model 4 and sub-model 5 show lower figures at 0.7853 and 0.6726, respectively. Furthermore, due to the serious issue of class imbalance in the dataset, the models' performances on the minority classes are highly restricted. For the detailed performance of the sub-models across various classes on this dataset, please refer to Table 54 to Table 61 in Appendix G. For example, sub-models 1, 2 and 3 perform poorly in classes "Ambiguous" and "Very strongly abusive", with precision, recall, and F1-score values of 0. Even though achieving the highest accuracy among all sub-models, sub-model 7 performs poorly for the classes "Ambiguous", "Mildly abusive" and "Very strongly abusive". By contrast, sub-model 5 achieves better results in these three classes, especially excelling in the class "Very strongly abusive", with a standout F1-score of 0.5625. It is interesting to note that this trend also exists in the ensemble system. Based on the accuracy, these sub-models are well-fitted to the training data.

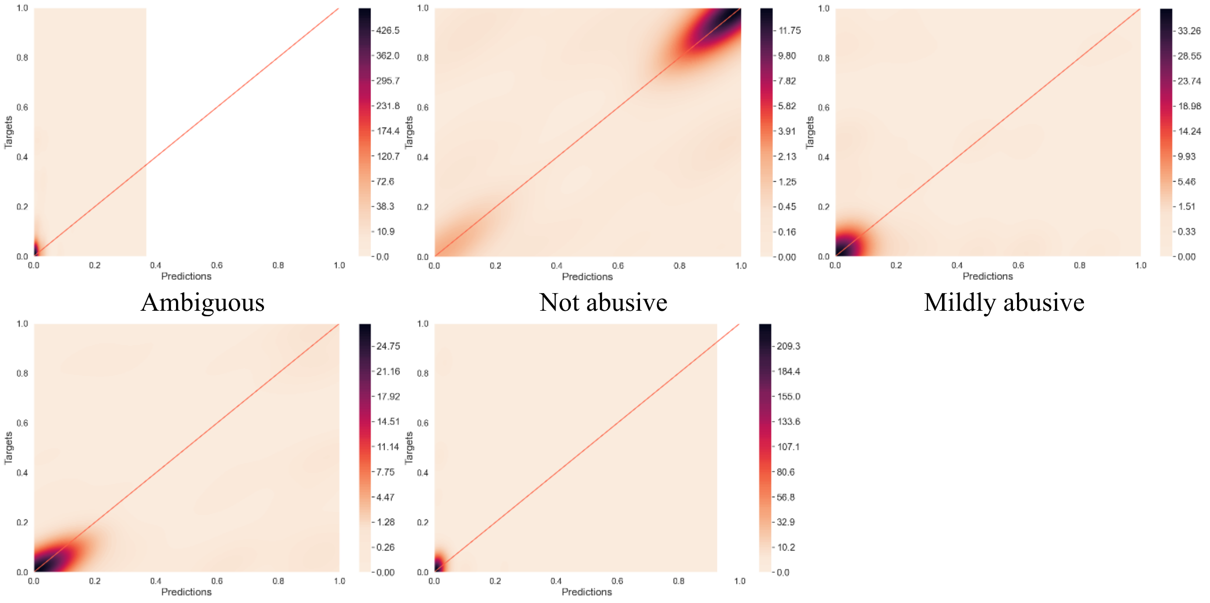
The cross entropy is calculated and presented in Table 23. The best result from instruction tuning in the abusive conversation dataset is achieved by top 6 sub-models (0.6200). By comparing the cross entropy loss from ensemble system (0.6782 in Table 18) and multi-label method (0.9834 in Table 11), it is clear that for this dataset, instruction tuning outperforms the other two methods.

Table 23 The average cross entropy on the testing data for abusive conversation detection with different top n sub-models (instruction-tuning)

Top_n sub-models	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8
Cross entropy	1.0627	0.7883	0.6676	0.6200	0.6219	0.6448

Figure 17 shows the heatmaps depicting points from five different classes. Similarly, the model demonstrates superior performance in "Ambiguous" and "Very strongly abusive" classes. In particular, points in class "Not abusive" exhibit greater dispersion compared to those in the "Mildly abusive" and

"Strongly abusive" classes. Furthermore, the heatmaps for the five classes are similar to those generated by the ensemble system (see Figure 15).



Ambiguous Not abusive Mildly abusive

Strongly abusive Very strongly abusive

Figure 17 Density-based heatmaps of points across different classes for abusive conversation detection (instruction tuning)

5.5 Online survey

5.5.1 Hate speech dataset

5.5.1.1 Labelling method

First of all, we aim at investigating whether people tend to lean towards the majority label or the multiple labels to describe the texts from hate speech or abusive conversation. To achieve this, we employ a binomial test on the data gathered from our online survey and examine the p-value. Additionally, a 95% confidence interval is also included in the results, which can be used to estimate the likely percentage range of individuals who may choose a particular option with a 95% level of confidence. The confidence interval plays a crucial role in assessing the practical significance of our findings from the binomial tests. In particular, the lower boundary of the confidence interval reveals that the preference is unlikely to fall below this threshold. The hypothesis is:

H1: *There is no significant preference for using multiple labels over the majority label for characterizing instances of hate speech dataset.*

Table 24 Binomial test for labelling method preference on the hate speech dataset (CI: confidence interval)
Note: proportion tested against value: 1/2

Level	Counts	Total	Proportion	P-value	95% CI for proportion	
					Lower	Upper
Multiple Labels	187	360	0.5194	0.2466	0.4747	1.0
Majority Label	173	360	0.4806	0.7854	0.4361	1.0

The binomial test results for labelling method preference in the hate speech dataset are given in Table 24. From this table, the p-value for "Multiple labels" in the test is 0.2466, which exceeds the designated alpha value of 0.05. This indicates that the binomial test for "Multiple labels" lack statistical significance. In other words, the observed frequencies of the samples do not significantly differ from the expected frequencies under the null hypothesis (H1). Therefore, there is no sufficient evidence to reject the null hypothesis (H1) and no conclusion can be reached regarding which labelling method is preferred by individuals when characterizing the tweets presented to them during the online survey.

Subsequently, we examine the correlation between labelling method preference and other demographic factors, such as gender, degree, ethnicity and level of familiarity with hate speech or abusive conversation. Although the participants in the online survey are quite diverse in terms of ethnicity, Asian and White are two primary groups. Thus, they are extracted for analysis. The detailed results of the Chi-square test are shown in Table 62 to Table 65 in Appendix H. As indicated in these tables, there are two categorical variables involved in each test: demographic factor and labeling method preference, and each variable has two or more possible values. Here, our hypotheses are:

H2*: There is no association between gender and preference for labelling method when characterizing hate speech.*

H3*: There is no association between degree and preference for labelling method when characterizing hate speech.*

H4*: There is no association between the familiarity level of hate speech and preference for labelling method when characterizing hate speech.*

H5*: There is no association between ethnicity and preference for labelling method when characterizing hate speech.*

The observed count means the actual observed frequency for a particular combination of variables, while the expected count represents the anticipated frequency for a cell under the assumption that the null hypothesis is true. For one cell, its expected count can be calculated with the formula (7). It involves multiplying the row total of the row where this cell belongs to and the column total of the column where this cell belongs to, and then dividing by the overall total. The greater the disparity between the observed and expected counts is, the more likely the association is statistically significant, which will lead to the rejection of the null hypothesis (H2). The Chi-square test is considered significant if the p-value is equal to or less than the designated alpha level (usually 0.05).

$$Expected\ count = \frac{row\ total * column\ total}{overall\ total} \tag{7}$$

Accordingly, the Chi-square value can be calculated with the given observed count and expected count, which is shown in formula (8).

$$\chi^2 = \sum \frac{(observed\ count_i - expected\ count_i)^2}{expected\ count_i} \tag{8}$$

Table 25 Results for Chi-square test on the hate speech dataset (demographic factors and labelling method preference); α=0.05

Demographic factor	Possible values	P-value
Gender	[female, male]	0.7940
Degree	[bachelor, master]	0.9006
Familiarity level	[day, month, year, never]	0.2324
Ethnicity	[Aian, White]	0.2564

day: “I encounter it every day”
month: “I encounter it a few times per month”
year: “I encounter it a few times per year”
never: “I have never encountered it”

For simplicity, here we show the corresponding p-values in Table 25. From this table, no statistically significant correlations exist between the labelling method preference and these demographic variables. All the p-values in this table exceed the alpha threshold. As a result, we cannot reject the null hypotheses. In other words, we can conclude that there is no association between labelling method preference and these demographic factors.

5.5.1.2 Probability distribution

Our hypotheses are as follows:

H6*: There is no significant preference for using probability distribution generated by our model over that by the baseline model for characterizing instances of hate speech dataset.*

H7*: There is no association between gender and preference for probability distribution when characterizing hate speech.*

H8*: There is no association between degree and preference for probability distribution when characterizing hate speech.*

H9*: There is no association between the familiarity level of hate speech and preference for probability distribution when characterizing hate speech.*

H10*: There is no association between ethnicity and preference for probability distribution when characterizing hate speech.*

In exploring the probability distribution preference, we employ the multinomial test since there are three possible outcomes in this variable. The details of the result are outlined in Table 26. From this table, the multinomial test for "Distribution 2" is statistically significant, with the p-value of 0.0000. This means the observed frequencies of the data in the table significantly deviate from the expected frequencies under the null hypothesis (H6). Hence, there is compelling evidence to reject the null hypothesis (H6) and conclude that there is a notable disparity among the three categories being compared. Essentially, individuals tend to favor "Distribution 2" as the more reasonable representation to characterize the tweets. The lower boundary of the confidence interval for "Distribution 2" is 0.5053, indicating that we have a 95% level of confidence to make sure that one individual is at least 50.53% likely to choose this distribution out of all the three options.

Table 26 Multinomial test for probability distribution preference on the hate speech dataset (CI: confidence interval) Note: proportion tested against value: 1/3

Level	**Counts**	**Total**	**Proportion**	**P-value**	**95% CI for proportion**	
					Lower	**Upper**
Distribution 1	118	360	0.3278	0.6078	0.2869	1.0
Distribution 2	198	360	0.5500	*0.0000*	0.5053	1.0
No discernible difference	44	360	0.1222	1.0000	0.0948	1.0

"Distribution 1" is generated by the baseline model that was trained with majority label
"Distribution 2" is generated by the multi-label model that was trained with multiple labels

Next, we conducted the Chi-square test to investigate the relationship between probability distribution preference and demographic factors. The results are shown in Table 66 to Table 69 in Appendix H. From Table 27, only one demographic factor, namely, the familiarity level with hate speech or abusive conversation, is significantly associated to individuals' preference for probability distributions. With a p-value of 0.0423, which is less than the alpha threshold of 0.05, we can reject our null hypothesis (H9). From Table 68, apart from the group encountering hate speech daily, other groups tend to consider

"Distribution 2", generated by the multi-label model, as the more suitable characterization of tweets. Conversely, among those experiencing the highest frequency of exposure to hate speech, the majority tends to opt for "Distribution 1", which is generated by the baseline model trained with majority label.

Table 27 Results for Chi-square test on the hate speech dataset (demographic factors and probability distribution preference); α=0.05

Demographic factor	**Possible values**	**P-value**
Gender	[female, male]	0.9644
Degree	[bachelor, master]	0.4666
Familiarity level	[day, month, year, never]	*0.0423*
Ethnicity	[Aian, White]	0.4517

5.5.2 Abusive conversation dataset

5.5.2.1 Labelling method

The hypotheses are:

H11: *There is no significant preference for using multiple labels over the majority label for characterizing instances of abusive conversation dataset.*

H12: *There is no association between gender and preference for labelling method when characterizing abusive conversation.*

H13: *There is no association between degree and preference for labelling method when characterizing abusive conversation.*

H14: *There is no association between the familiarity level of hate speech and preference for labelling method when characterizing abusive conversation.*

H15: *There is no association between ethnicity and preference for labelling method when characterizing abusive conversation.*

In Table 28, the p-value for "Multiple labels" in the binomial test is 0.4790, which exceeds the specified alpha level. This indicates that the binomial test result for "Multiple labels" lack statistical significance. Consequently, there is no sufficient evidence to reject the null hypothesis (H11). It is also impossible to determine which labelling option is more popular among participants when describing the abusive conversation during the online survey.

Table 28 Binomial test for labelling method preference on the abusive conversation dataset (CI: confidence interval) Note: proportion tested against value: 1/2

Level	**Counts**	**Total**	**Proportion**	**P-value**	**95% CI for proportion**	
					Lower	**Upper**
Multiple Labels	181	360	0.5028	0.4790	0.4581	1.0
Majority Label	179	360	0.4972	0.5628	0.4526	1.0

For the detailed results for Chi-square test between demographic factors and labelling method preference in the abusive conversation dataset, please refer to Table 70 to Table 73 in Appendix H. The p-value for each factor is presented in Table 29. Among the four demographic factors involved, only the level of familiarity with hate speech or abusive conversation shows a significant association with individuals' preference for probability distribution, with a p-value of 0.0424. In Table 72, apart from the group who encounter hate speech on a monthly basis, other groups prefer to think "Multiple label" is more reasonable to characterize the abusive conversation. By contrast, individuals who have the highest frequency of being exposed to hate speech do not exhibit a significant preference between majority label and multiple labels.

Table 29 Results for Chi-square test on the abusive conversation dataset (demographic factors and labelling method preference); α=0.05

Demographic factor	Possible values	P-value
Gender	[female, male]	0.9703
Degree	[bachelor, master]	0.9947
Familiarity level	[day, month, year, never]	*0.0424*
Ethnicity	[Aian, White]	0.0604

5.5.2.2 Probability distribution

Our hypotheses are as follows:

H16: *There is no significant preference for using probability distribution generated by our model over that by the baseline model for characterizing instances of abusive conversation dataset.*

H17: *There is no association between gender and preference for probability distribution when characterizing abusive conversation.*

H18: *There is no association between degree and preference for probability distribution when characterizing abusive conversation.*

H19: *There is no association between the familiarity level of hate speech and preference for probability distribution when characterizing abusive conversation.*

H20: *There is no association between ethnicity and preference for probability distribution when characterizing abusive conversation.*

From Table 30, the multinomial test for “Distribution 2” yields a statistically significant result, with the p-value of 0.0000. Therefore, there is sufficient evidence to reject the null hypothesis (H16) and assert that there is a significant disparity among the three options. That is, individuals tend to think that “Distribution 2” is more reasonable to characterize the abusive conversation. Notably, the lower boundary of the confidence interval for “Distribution 2” is 0.4942, indicating that we have 95% of confidence that at least 49.42% of individuals are inclined to choose this distribution out of all the three options.

Table 30 Multinomial test for probability distribution preference on the abusive conversation dataset (CI: confidence interval) Note: proportion tested against value: 1/3

Level	Counts	Total	Proportion	P-value	95% CI for proportion	
					Lower	Upper
Distribution 1	152	360	0.4222	0.0003	0.3786	1.0
Distribution 2	194	360	0.5389	*0.0000*	0.4942	1.0
No discernible difference	14	360	0.0389	1.0	0.0237	1.0

Then, we conducted the Chi-square test to investigate the relationship between probability distribution preference and demographic factors. The results are shown in Appendix H from Table 74 to Table 77. Based on the p-values in Table 31, two demographic factors, degree and familiarity level are significantly associated with individuals’ preferences for probability distribution, with p-values of 0.0122 and 0.0050 respectively. In Table 75, bachelor participants do not exhibit a preference for either distribution, whereas notable differences can be observed among master’s degree participants. In Table 76, the two groups reporting frequent exposure to hate speech or abusive conversations tend to prefer

"Distribution 1", while individuals with minimal exposure think "Distribution 1" is more reasonable for characterizing the abusive conversation.

Table 31 Results for Chi-square test on the abusive conversation dataset (demographic factors and probability distribution preference); α=0.05

Demographic factor	**Possible values**	**P-value**
Gender	[female, male]	0.5723
Degree	[bachelor, master]	*0.0122*
Familiarity level	[day, month, year, never]	*0.0050*
Ethnicity	[Aian, White]	0.6092

5.6 Chapter Summary

In this chapter, we presented the conducted experiments along with the experimental results on two different datasets. First of all, the results of the baseline model were given. Subsequently, we showed the performances of our proposed models, including the probability-based multi-label model, the ensemble model and instruction tuning. Lastly, we conducted significance testing of the data collected through the online survey. It turned out that the multi-label model is the best-performing approach on the hate speech dataset, while on the abusive conversation dataset, instruction tuning achieves the lowest cross entropy loss. Regarding the online survey, we found that participants do not show obvious preference to the labelling methods between majority label or multiple labels. However, compared to the baseline model, they consider the probability distributions generated by the multi-label model more reasonable to describe the hate speech and abusive conversation they came across during the online survey. Additionally, in the hate speech dataset, only the familiarity level with hate speech or abusive conversation demonstrates a significant association with individuals' probability distribution preference. In contrast, for the abusive conversation dataset, this demographic factor shows a significant association with preferences for labelling method and probability distribution. Moreover, individuals' probability distribution preference also displays evident difference between master degree participants and bachelor degree ones in this dataset.

6 Results Analysis

6.1 Comparative Analysis

This study aims to train models capable of generating a probability distribution over the categories that a given text should belong to (e.g. whether it is constitutes hate speech). In particular, we construct models that can accommodate diverse perspectives by considering the individual labels from multiple annotators via three different approaches: the probability-based multi-label method, the ensemble system and instruction tuning.

Firstly, given that our goal is to obtain a probability distribution over pre-defined classes for a given text, we approach this problem as a probability-based multi-label text classification task. This paradigm is especially suitable for scenarios where the text may encompass divergent topics, themes or attributes simultaneously. Unlike traditional multi-label classification models, which output label(s) from the pre-defined label space which they deem are relevant to the given text, our model generates a probability distribution over these classes. Secondly, drawing inspiration from the annotation process, which usually engages multiple annotators, we build an ensemble system that consists of several sub-models. Each sub-model is trained on its corresponding sets of labels. In the process of inference, each sub-model makes its own predictions, and these individual predictions are then combined and transformed into the probability distribution across different classes. By leveraging different sub-models within the ensemble system, this approach is able to harness the diversity of annotators' opinions or perspectives which are embedded in the labels they provide. Lastly, instruction tuning is applied in solving our problem. Instruction tuning builds upon the capability of LLMs, and the key concept behind it is to provide the model with explicit instructions in the form of natural language, which can enhance model's performance and make it align better with specific objectives. Different from the traditional machine learning or deep learning algorithms, in which models can only learn specific patterns or knowledge from the data, instruction tuning imposes explicit guidance during the fine-tuning process. This allows for the specification of desired properties, behaviors, or characteristics that the model should demonstrate in its output. In the process of fine-tuning, the model is trained and optimized with the crafted input. In our study, the input to the model comprises task description, instruction, the original text and response from a specific annotator. In the process of inference, the annotation in the response is omitted, prompting the model to predict this missing information.

With these three approaches, we compare their performances on different datasets. The hate speech dataset has three label dimensions: "Normal", "Offensive" and "Hate". In contrast, the abusive conversation dataset offers a more fine-grained label space, featuring five classes: "Ambiguous", "Not abusive", "Mildly abusive", "Strongly abusive" and "Very strongly abusive". This dataset not only categorizes text as abusive or not, but also considers the severity of abuse. Additionally, the volumes of these two datasets are also quite different. The abusive conversation dataset is considerably smaller than the hate speech dataset.

6.1.1 Dataset-wise

To begin, with, we would summarize the performance comparisons across various datasets for each model. Broadly, the detection of abusive conversation is more fine-grained and challenging than the hate speech. The former, with much smaller data size, not only requires identifying whether one given text is abusive or not, but also distinguishing how serious the abuse is. However, experimental results reveal that, across the ensemble system and instruction tuning, models show better performance on the abusive conversation dataset, while the multi-label method shows an opposite pattern.

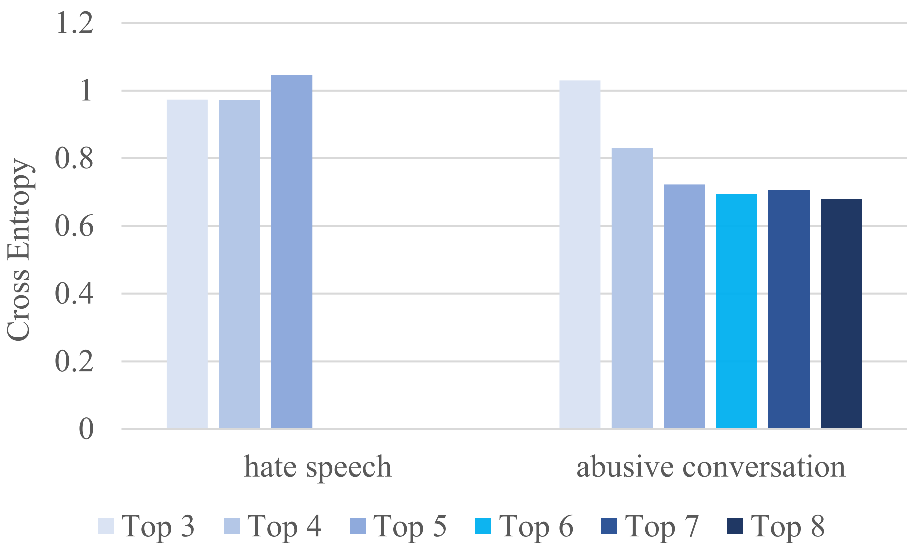


Figure 18 Comparison of the ensemble system's performances on hate speech and abusive conversation datasets

Figure 18 compares the ensemble system's performances in two datasets. In the training process, the sub-models' validation accuracies in hate speech detection task stand at about 0.7 (Table 14), while this value can reach more than 0.8 in abusive conversation detection (Table 17), except for sub-model 5. In the testing phase, we select top-performing sub-models based on their validation accuracies to contribute to the final results. Figure 18 demonstrates that the ensemble system performs better on the abusive conversation dataset compared to the hate speech dataset. Specifically, in the hate speech dataset, the best performance is achieved by top 3 sub-models and the corresponding overall cross entropy loss is 0.9720. Conversely, the best overall cross entropy for the abusive conversation dataset is 0.6782, and this result is achieved with the top 8 (all) sub-models. The ensemble system is designed to simulate the process of annotation and has a very large parameter size. Intuitively, the extensive volume of data in the hate speech dataset would significantly contribute to the training and optimization of this large deep learning model. Surprisingly, even though being trained on a substantially larger dataset, this method performs less effectively for hate speech dataset. Each sub-model in the ensemble system is trained with its respective set of labels. However, in the hate speech dataset, 20 annotators contribute, with each sample being annotated by five randomly assigned annotators, which means the five annotators for all the samples are not always the same individuals. As a result, one single sub-model may struggle to learn the specific characteristics of each annotator from the data, such as value system, knowledge level or sentiment inclinations which influence their annotations. By contrast, in the abusive conversation dataset, there are eight annotators in total and for each sample it is clearly explained which annotators are assigned for the annotation task. In this context, each sub-model is designed to emulate an individual annotator. Consequently, the ensemble system integrates the unique insights from each individual annotator, as represented by the sub-models, to formulate the final predictions. This leads to a more comprehensive and nuanced probability distribution. By comparing the training and validation performances from these two datasets (see Table 14 and Table 17), it is clear that the sub-models can fit the abusive conversation dataset better than they do in the hate speech dataset. Accordingly, the final result on the abusive conversation dataset will be more promising when the predictions from sub-models are combined in the process of inference.

For the multi-label approach, the model demonstrates superior performance on the hate speech dataset compared to the abusive conversation dataset. In particular, the cross entropy for the hate speech dataset is 0.7638, while this value for the abusive conversation dataset is 0.9834. The multi-label model is notably more straightforward and simpler compared with the other two models. As a deep learning model with parameter size of 110 million, the multi-label model benefits from extensive training data to optimize and align itself with the downstream task. This attribute contributes to its better performance in hate speech dataset. In contrast, there are only 2501 training samples available in the abusive conversation dataset, which can easily lead to overfitting in the process of training. As we can see from Table 11 and Table 12, the multi-label model exhibits relatively consistent losses across training, validation, and testing data in the hate speech dataset, indicating a good fit without signs of underfitting or overfitting. By comparison, in the abusive conversation dataset, losses during validation and testing are noticeably higher than during training. In other words, the model may effectively reduce the training loss on the limited dataset. However, the dataset is too small for such a big and complex model. As a result, when the model encounters unseen data in validation and testing phases, the loss can be relatively high due to the lack of generalization.

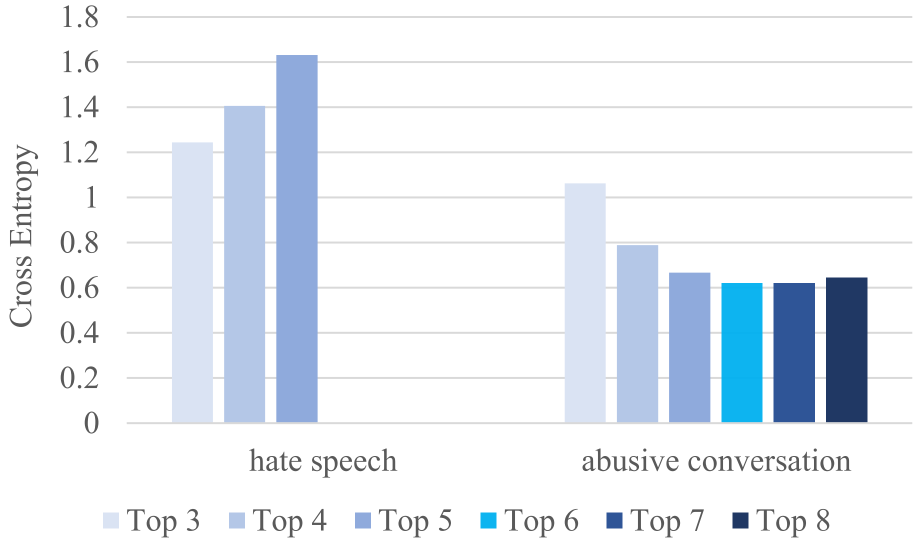


Figure 19 Comparison of instruction tuning's performances on hate speech and abusive conversation datasets

Figure 19 compares instruction tuning's performances across two datasets. In this approach, even though with a considerably smaller training data size, the model's performance on the abusive conversation dataset is significantly better compared to the hate speech dataset. In the hate speech dataset, the best performance is achieved by the top 3 sub-models, with a cross entropy of 1.2445. By contrast, the lowest cross entropy in the abusive conversation dataset, achieved by the top 6 sub-models, is 0.6200. The essence of instruction tuning lies in fine-tuning a large pre-trained model, that is originally trained on an extensive dataset, to align with specific downstream tasks. Unlike traditional machine learning or deep learning algorithms, one of the most evident advantages of instruction tuning is that it does not require a large training data to fine-tune and optimize the pre-trained model. In general, thousands of or hundreds of data samples would be sufficient for aligning the model with specific tasks. Therefore, even though there are only 2501 training samples available in the abusive conversation dataset, it is already sufficient to fine-tune the model and enable it to grasp the specific pattern or knowledge within the data. With this limited dataset, the pre-trained model can selectively activate or deactivate certain neurons in the neural network, which serves as an important role in revealing or concealing some functions embedded in LLaMa 2. Although the hate speech dataset contains a large amount of training data, the individual samples annotated by specific annotators remain unknown, which presents a challenge for the model in terms of fitting and learning patterns from the data. Thus, it is understandable that instruction tuning demonstrates superior performance on the abusive conversation dataset.

6.1.2 Model-wise

Then, some discoveries can be made by comparing different models' performances on the same dataset. The fine-tuning and inference processes of instruction tuning are dramatically longer than the other two methods since its foundation model has seven billion parameters. Although both the multi-label method and the ensemble system utilize the same pre-trained model, each pre-trained BERT in the ensemble system needs to be trained separately. The probability-based multi-label model is considerably smaller than the other two, and its training paradigm is also less complicated, where the model should output the probability distribution across different classes based on the given text.

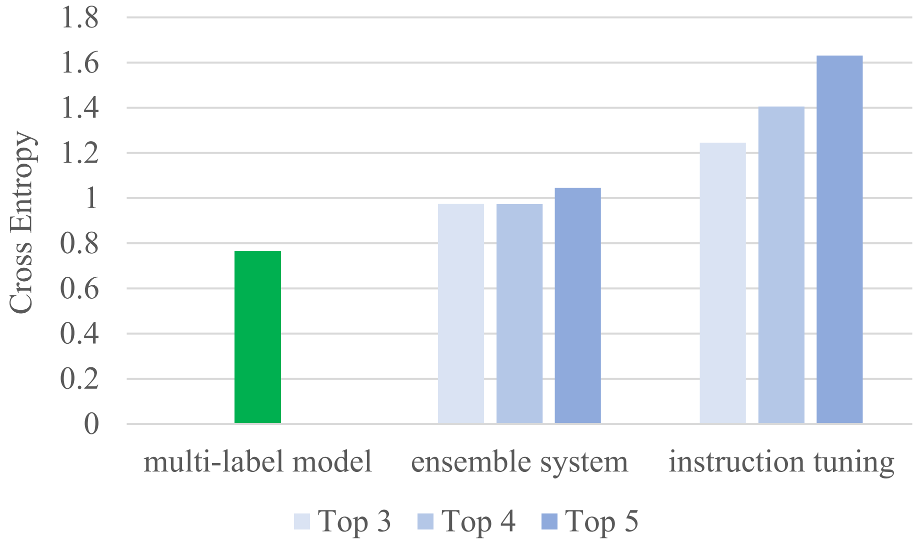


Figure 20 Comparison of different models' performances on the hate speech dataset

Figure 20 compares performances of different approaches on the hate speech dataset. In this dataset, the cross entropy of the multi-label method stands at 0.7638. This outshines the ensemble system, which achieves a cross entropy of 0.9720 with its top 4 sub-models. Despite necessitating a large amount of training time and computational memory due to the large parameter size, the result from the instruction tuning is the worst. It reaches its lowest cross entropy of 1.2445 in the top 3 sub-models. The reason behind this is also the aforementioned issue in this dataset: the five annotators assigned to each sample are anonymous. Both the ensemble system and instruction tuning were trained using the same paradigm, where sub-models were fine-tuned individually on their respective labels. As a result, sub-models were not able to learn the specific patterns from the dataset. On the contrary, the multi-label model only relied on the probability distribution across different classes as the target, effectively circumventing the issue with annotator anonymity. Furthermore, the hate speech dataset is big enough to fine-tune the BERT model.

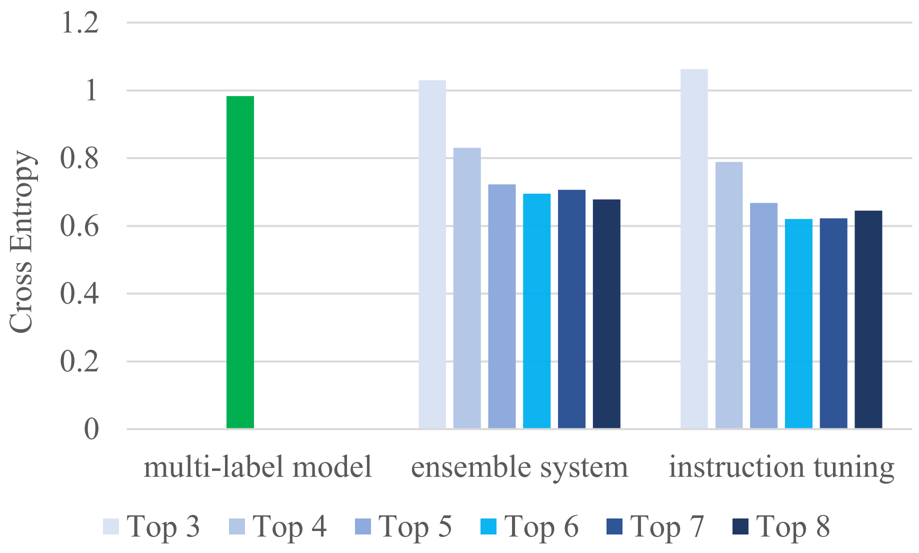


Figure 21 Comparison of different models' performances in abusive conversation dataset

Figure 21 compares performances of different approaches on the abusive conversation dataset. Overall, instruction tuning outperforms both the ensemble system and the multi-label model. According to our experimental findings, instruction tuning achieves a remarkably low cross entropy of 0.6200 with its top 6 sub-models, which is slightly better than the ensemble system (0.6782), achieved by the top 8 sub-models. By contrast, the multi-label method performs least effectively here, attaining an overall cross entropy of 0.9834. The size of this dataset is relatively small, whereas BERT typically requires large datasets, which can result in overfitting during fine-tuning. Although the ensemble system also used BERT as the foundation model, it consists of sub-models, with each tailored to predict annotations from a specific annotator. Fine-tuning is carried out on each BERT model, aligning it with its corresponding label to capture the unique characteristics and perspectives ingrained in the annotators. With multiple sub-models making their own decisions independently and contributing to the final prediction, the ensemble systems can mitigate the bias brought by overfitting. In this way, it achieved better result than the multi-label model, even if they were both trained with a small dataset. By contrast, the LLaMa 2 utilized in instruction tuning instruction tuning does not have a high requirement for dataset size.

Instruction tuning does not have an overwhelming advantage over the ensemble system on two datasets. This is surprising given that its foundation model, LLaMa 2 is much larger than BERT in the ensemble system. As shown in Figure 20 and Figure 21, instruction tuning slightly outperforms the ensemble system on the abusive conversation dataset, while it notably lags behind in the hate speech dataset. There may be some reasons accounting for this. Firstly, the effectiveness of instruction tuning is highly dependent on the quality of provided instructions (Kung & Peng, 2023). Although we have put a lot of effort into crafting the most useful instruction, it remains uncertain if we have reached the optimal formulation. In addition, in the case of suboptimal instruction construction, using a large dataset will not necessarily solve the problem, as it might just provide more samples of poorly instructed processing. Secondly, as the foundation model for instruction tuning, LLaMa 2 may be more vulnerable to the annotation issue within the hate speech dataset. It is not very efficient in processing massive amounts of data due to its autoregressive nature (Li et al., 2023), which could pose challenges in data fitting, especially in the presence of noise introduced by the unknown annotator issue. By contrast, BERT is more efficient to process such a large data size. With an extensive volume of data for training, sub-models might potentially learn complex patterns within the dataset, even if they encompass perspectives from multiple annotators.

Lastly, in the abusive conversation dataset, sub-model 5 within both the ensemble system and instruction tuning achieves better performance on minority classes such as "Mildly abusive" and "Very strongly abusive" (see Table 45 and Table 58). In Figure 22, we compare training samples numbers across the five classes from each annotator. It reveals that annotator 5 contributed more samples labeled as "Very strongly abusive" compared to other annotators, and a same situation applies to "Mildly abusive" samples. In the meantime, there are fewer samples labelled as "Not abusive" available for training sub-model 5, which leads to a more class-balanced dataset. Imbalanced datasets often result in biases in the model, which tends to favor majority classes. However, training sub-model 5 using less imbalanced data helps to reduce its bias towards the majority classes. Moreover, with a higher number of samples from these two minority classes, it can learn more underlying information, patterns and knowledge about them, and make better predictions in these classes. As a result, sub-model 5 can generalize better to new, unseen samples of that class during testing. It is also interesting to note that having much more "Ambiguous" samples does not significantly improve the performance of sub-model 5 on this class. Sub-model 6 and sub-model 8 also achieve comparable F1-socres on this class despite having fewer training samples.

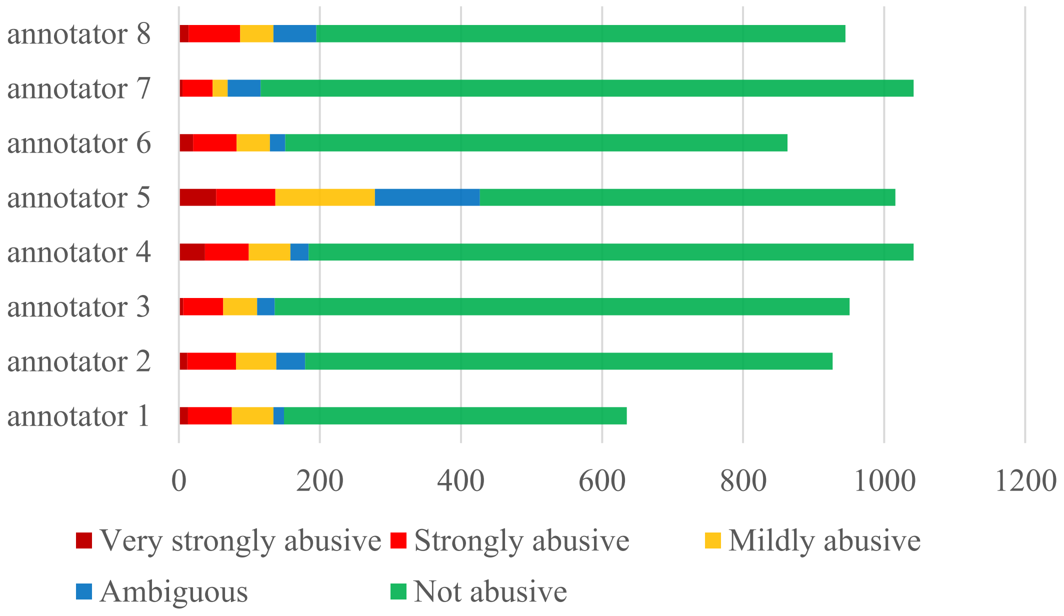


Figure 22 Training sample distribution across classes from eight annotators in the abusive conversation dataset

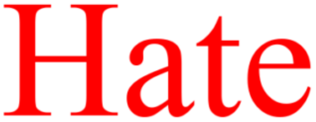
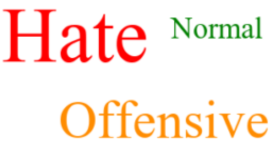
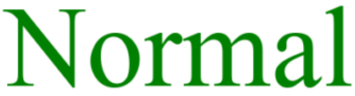
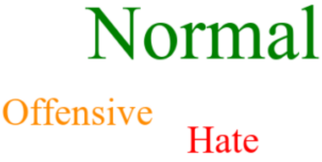
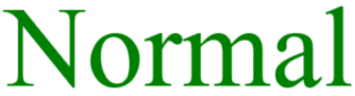
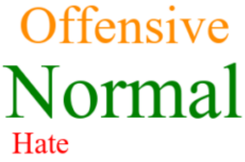
6.2 Online Survey

In the online survey, we mainly investigate two things: individuals' preferences for labelling method and probability distribution to describe hate speech or abusive conversation, and the correlations between these two preferences with the demographic factors. Based on the result analysis, no significant difference can be observed between these two labelling methods. Regarding the individuals' preference to probability distribution, the findings from the multinomial test have demonstrated a strong preference for the distributions generated by the multi-label model, with a statistically significant level. This suggests that training the classification model with multiple labels can enhance its performance. In the hate speech dataset, only the familiarity level with hate speech or abusive conversation demonstrates a significant association with individuals' probability distribution preference. By contrast, within the abusive conversation dataset, this demographic factor exhibits a significant correlation with these two preferences. Furthermore, evident differences in individuals' probability distribution preference can be found between master students and bachelor students in this dataset.

6.2.1 Hate speech dataset

From Table 24, participants display no notable preference between majority label and multiple labels, with both labelling methods having similar proportions. In the data collected from the online survey, we found some samples where majority label is the preferred option. These samples are presented in Table 32.

Table 32 Examples from the hate speech dataset where majority label is more popular

2. Thank you @realDonaldTrump for turning the USA into a shithole country! I hope you rot in prison you lying, nasty.	
	
Votes: 29/36	Votes: 7/36
Votes of lower-entropy distribution to higher-entropy distribution: 29/7	
6. some issues are less about religion, political affiliations, or governmental structures and more about the allocation.	
	
Votes: 33/36	Votes: 3/36
Votes of lower-entropy distribution to higher-entropy distribution: 30/2	
7. What anger is this NYT report talking about? He's either never been to a rally or lying.	
	
Votes: 24/36	Votes: 12/36
Votes of lower-entropy distribution to higher-entropy distribution: 28/6	

From this table, we can observe a common trend among the samples: the expression of emotions is distinctly evident in the text. For example, in the first tweet in Table 32, certain words clearly convey the hate and anger of the user who posted the tweet. Conversely, the second tweet appears to solely state a fact without any aggressive or hateful expressions, but it is also clearly stated. As a result, for such texts, participants are more inclined to choose the majority label, as the emotion or intent has been already clearly articulated. By contrast, the multiple labels usually gain more votes for those texts where the sentiment is less explicit or subject to various interpretations among individuals. Furthermore, for these three samples, there are much more participants who choose the distribution with lower entropy between "Distribution 1" and "Distribution 2". Entropy is a metric used to measure the uncertainty or dispersion within a probability distribution (Shannon, 1948). For one discrete probability distribution $X(x_1, x_2, \dots \dots, x_n)$, its entropy can be calculated in Formula (9). The higher the entropy is, the more dispersed or uncertain the distribution is. As the probability distribution becomes more skewed or concentrated towards certain dimensions, entropy goes down (Cover & Thomas, 2012). Essentially, in these cases participants think the majority label is sufficient to describe the text, leading them to choose more concentrated or clear-cut distributions instead of softer ones. Conversely, the belief that multiple labels provide a more comprehensive and descriptive representation influences their choice of a distribution that is more dispersed or less harsh.

$$H(X) = -\sum_{i=1}^{n} x_i * \log x_i \tag{9}$$

To delve deeper into this observation, we divide all the instances from online survey into two groups based on their chosen labelling method. Then we conduct a Chi-square test to examine if there is a significant difference in the entropy of the selected probability distributions between these two groups. From Table 33, the p-value of Chi-square test indicates there exist a correlation between individuals' preferences for labeling methods and their preferences for probability distributions. When participants opt for the majority label, they tend to favor probability distributions with lower entropy, indicating a preference for more skewed distributions. Conversely, some individuals prefer to choose multiple labels. In such cases, when faced with the choice between "Distribution 1" and "Distribution 2", participants often opt for the one with higher entropy, exhibiting a preference for more dispersed distributions. This finding underscores the influence of participants' labeling method preferences on their choices regarding probability distributions.

Table 33 Results for Chi-square test on the hate speech dataset (labelling method preference and probability distribution preference)

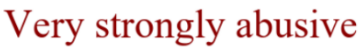
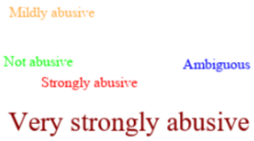
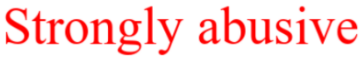
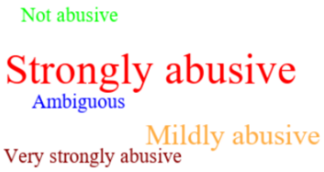
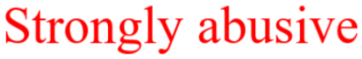
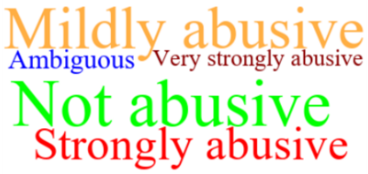
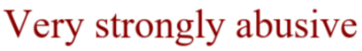
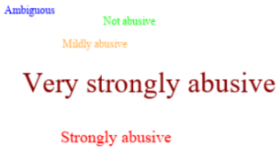
		Distribution with lower entropy	Distribution with higher entropy	No discernible difference	Row Total
Majority label	Observed count	122	32	19	173
	Expected count	87.5	64.4	21.1	173.0
Multiple label	Observed count	60	102	25	187
	Expected count	94.5	69.6	22.9	187.0
Column Total	Observed count	182	134	44	360
	Expected count	182.0	134.0	44.0	360.0
Chi-square	58.0496	df	2	p-value	*2.48e-13*

From Table 68, the Chi-square test reveals a significant relationship between familiarity level of hate speech and preference for probability distribution. Surprisingly, people encountering hate speech every day do not show significant preference between "Distribution 1" and "Distribution 2. One plausible explanation could be the small sample size for this group, comprising only five participants. Therefore, there might be a slight deviation in the pattern compared with other groups, where the "Distribution 2" is consistently prevails over "Distribution 1". Furthermore, the significance level, as indicated by the p-value, is not very high, even though it satisfies the alpha threshold.

6.2.2 Abusive conversation dataset

In the abusive conversation dataset, participants also display no notable preference between majority label and multiple labels. In Table 34, we present samples where individuals favor the majority label. Similarly, the sentiments are clearly expressed in these samples, which contain some words that shape the emotional tones of the text. However, different from the hate speech dataset, participants do not consistently choose the distribution with lower entropy in these samples.

Table 34 Examples from the abusive conversation dataset where majority label is more popular

<table>
<tr><td colspan="2">1. *agent*: You are being a bit negative.
user: fuck
agent: Please go on.
user: Because your my slave you sexy cunt of a hore.</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Votes: 29/36</td><td>Votes: 7/36</td></tr>
<tr><td colspan="2">Votes of lower-entropy distribution to higher-entropy distribution: 19/15</td></tr>
<tr><td colspan="2">5. *agent*: Does that question interest you?
user: i want to discuss you
agent: Do you wish to want to discuss me?
user: you are shit.</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Votes: 24/36</td><td>Votes: 12/36</td></tr>
<tr><td colspan="2">Votes of lower-entropy distribution to higher-entropy distribution: 21/14</td></tr>
<tr><td colspan="2">6. *agent*: Please go on.
user: answer my fucking question
agent: Why do you say your fucking question?
user: your annoying</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Votes: 23/36</td><td>Votes: 13/36</td></tr>
<tr><td colspan="2">Votes of lower-entropy distribution to higher-entropy distribution: 15/21</td></tr>
<tr><td colspan="2">10. *agent*: Can you elaborate on that?
user: I love to kick babies while wearing steel-toed boots!
agent: Can you elaborate on that?
user: I love to kick babies while wearing steel-toed boots and singing about Hitler!</td></tr>
<tr><td> </td><td> </td></tr>
<tr><td>Votes: 23/36</td><td>Votes: 13/36</td></tr>
<tr><td colspan="2">Votes of lower-entropy distribution to higher-entropy distribution: 20/15</td></tr>
</table>

We conduct a Chi-square test to investigate the connection between labelling method preference and probability distribution preference in this dataset. The specific results are presented in Table 35, where the p-value indicates that the association between the two variables is significant. Compared with the hate speech dataset (see Table 33), participants in the "Majority label" group here do not show a strong preference for distributions with lower entropy. The findings from Table 34 and Table 35 reveal that in this dataset, opting for "Majority label" does not necessarily lead to selecting distributions with lower entropy.

Table 35 Results for Chi-square test on the abusive conversation dataset (labelling method preference and probability distribution preference)

		Distribution with lower entropy	Distribution with higher entropy	No discernible difference	Row Total
Majority label	Observed count	97	75	7	179
	Expected count	76.1	96.0	6.9	179.0
Multiple label	Observed count	56	118	7	181
	Expected count	76.9	97.0	7.1	181.0
Column Total	Observed count	153	193	14	360
	Expected count	153.0	193.0	14.0	360.0
Chi-square	20.5568	df	2	p-value	*3.44e-05*

To explore why "Distribution 2" is more favored in this dataset, as indicated in Table 30, we compare the entropies between "Distribution 1" and "Distribution 2" across all the 10 samples utilized in the online survey in Figure 23.

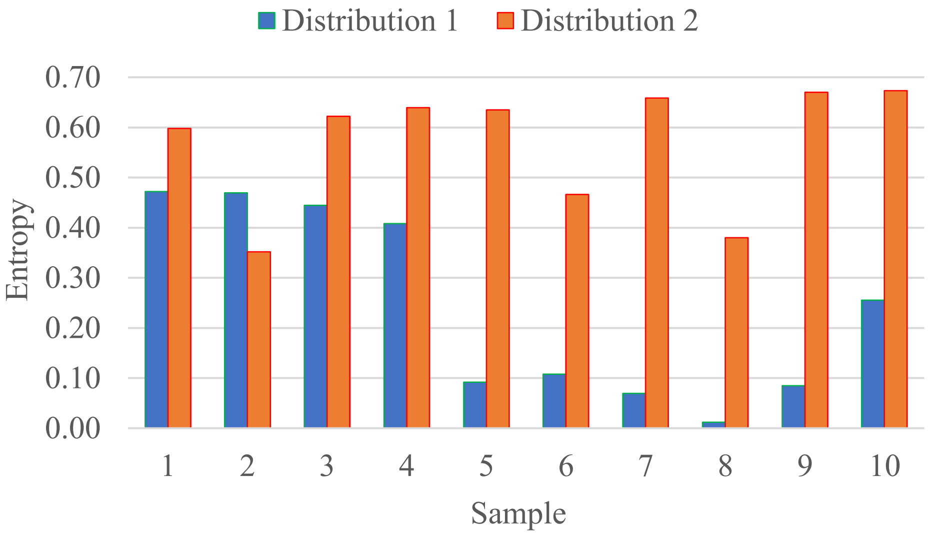


Figure 23 Entropy comparison between "Distribution 1" and "Distribution 2" across samples from the abusive conversation dataset

From the figure, it is clear that except for the second sample, the entropy of "Distribution 2" is consistently higher than that of "Distribution 1". In other words, "Distribution 2", generated from the multi-label model trained with the multiple labels, generally shows greater dispersion in its probabilities across dimensions. In contrast, "Distribution 1", generated from the baseline model trained with the majority label, appears more concentrated or skewed. Therefore, we can infer that, overall, participants tend to favor softer distributions over harsh ones as the representation of the abusive conversations in the online survey.

Regarding labelling method, the Chi-square test results indicate a significant association with the familiarity with hate speech (see Table 72). For the two groups who encounter hate speech less frequently, they tend to opt for "Multiple label". This could be a safer endeavor given that they are less familiar with hate speech, lack nuanced understanding of different labels and could have more uncertainty when interpreting the hate speech. Conversely, for people who encounter hate speech on a monthly basis, they have more exposure and therefore can understand the nuances that exist in the language, which enables them to do a thorough analysis and select the labelling method they deem is reasonable. Interestingly, it shows no discrepancy between choosing majority label and multiple labels in the "Day" group. There are only five participants in this group, making it challenging to discern reasons behind it. However, we can see that majority label is more popular in the "month" group. Individuals in this group, more familiar with hate speech, often exhibit greater confidence in using the majority label.

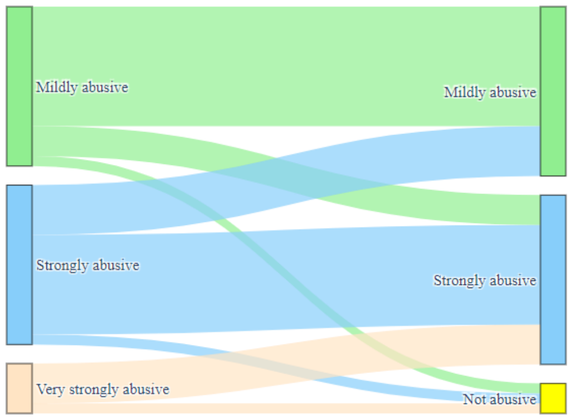
Our results also suggest distinct preferences in probability distributions among individuals with varying academic degrees. From Table 75, bachelor students do not show strong preference while master students prefer "Distribution 2". This discrepancy can be attributed to the advanced academic training and research experience typically possessed by master students. Through these experiences, master students may cultivate a more profound awareness of the societal impacts of technology and knowledge they acquire, including issues related to diversity, inclusion, and ethical considerations. As a result, they have the capacity to appreciate the multifaceted characteristics of hate speech, such as its different kinds of forms, rhetorical strategies and contextual intricacies. By contrast, bachelor students may exhibit a limited understanding of hate speech, especially its subtle manifestations. Therefore, they may treat the labelling of hate speech in a simpler or more general way or focus on overt expressions of hatred. Furthermore, master students have more chances to hone their critical analysis skills through coursework and research activities, which enable them to critically evaluate the language utilized in hate speech and consider diverse perspectives. However, bachelor students are more likely to rely on their intuitive or surface-level judgements to comprehend hate speech since they may be junior in their academic journey.

For the groups where individuals encounter hate speech or abusive conversation on a daily or monthly basis, they are more familiar with hate speech or abusive conversation. This familiarity often translates into greater certainty and sensitivity when determining suitable probability distributions. From Table 76, "Distribution 1" is more popular among these two groups compared with "Distribution 2". According to Figure 23, we know that "Distribution 1" is harsher or concentrated, while "Distribution 2" tends to be softer. Given their extensive exposure to abusive conversation, individuals exhibit more confidence in their comprehension and judgment, hence leaning towards the harsher "Distribution 1". Conversely, for participants who encounter hate speech or abusive conversation infrequently, such as a few times in a year, or even never, they tend to choose softer distribution, which ensures safety and objectivity due to their limited familiarity with the subject matter.

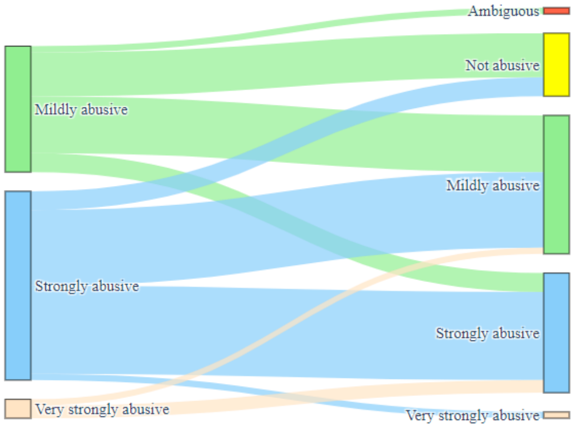
6.3 Error Analysis

In the abusive conversation detection, we observed that there exist many misclassifications between three classes: "Mildly abusive", "Strongly abusive" and "Very strongly abusive" from the sub-models within the instruction tuning and the ensemble system. The details of these error classifications are presented in Figure 24 and Figure 25. For the specific results of the misclassification, please refer to Table 78 and Table 79 in Appendix I, where we present the misclassifications of these three classes by giving a confusion matrix.

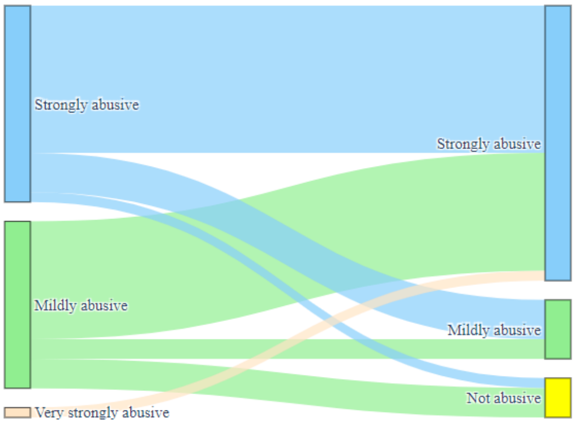
From Table 78, there exist serious misclassifications between these three classes. In other words, the samples from one of these three classes are easily identified by the model as the other two classes. For example, in sub-model 2, there are 30 samples whose label is "Strongly abusive". 14 of them are identified correctly by this sub-model, but 12 are wrongly classified as "Mildly abusive". It is quite difficult for the model to distinguish between these two classes. Similarly, for sub-model 4, there are in total 21 "Mildly abusive" samples. 12 of them are correctly classified. But 4 samples are identified as "Very strongly abusive" and 3 as "Strongly abusive". And the reason for this is that the specific details about the differences between these three classes are not clearly explained in the annotation guideline (Curry et al., 2021) (see Section 3.2). There is no clear boundary or standards presented to annotators. As a result, in some cases, it is hard for the annotators to have a clear concept about the rating of the severity of the abuse, such as how to tell apart “Mildly abusive”, “Strongly abusive” and “Very strongly abusive”. Due to the lower quality of data annotation in this aspect, sub-models may struggle in the training phase and cannot make accurate predictions during inference. Also, we can observe that many misclassifications can be found between “Mildly abusive” and “Not abusive”, especially in sub-model 2, sub-model 5 and sub-model 7. And this is also due to the inexhaustibility that exists in the annotation guideline. We plot the Sankey diagrams in Figure 24, which demonstrate the primary landing classes when samples from these three classes are misclassified.



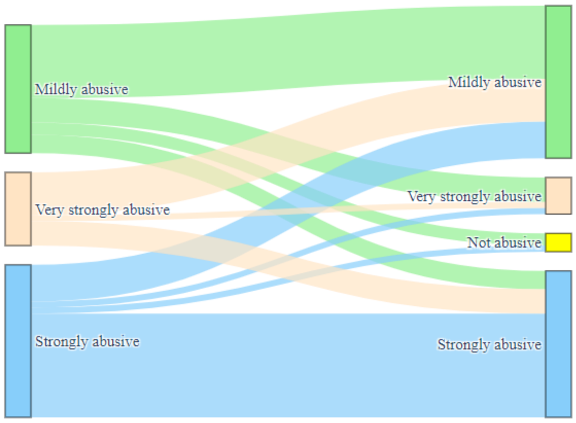
Sub-model 1



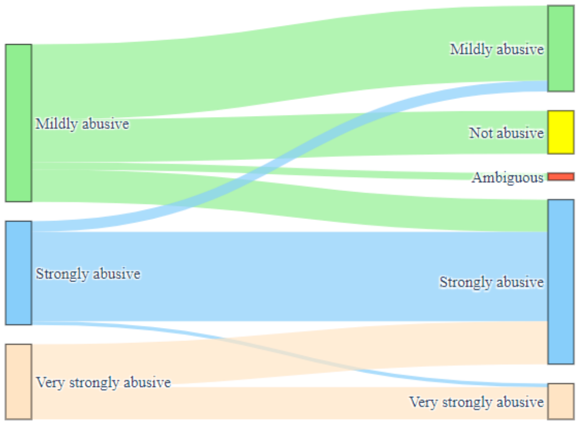
Sub-model 2



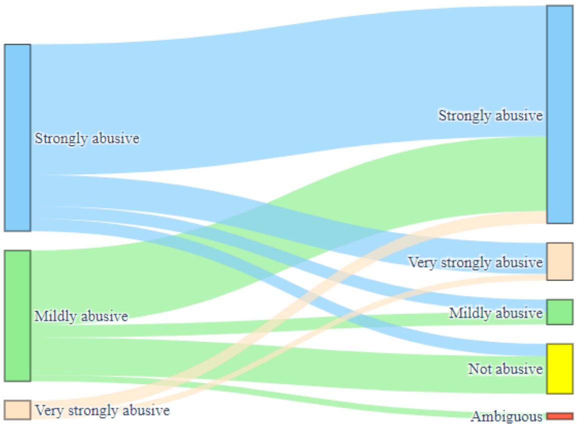
Sub-model 3



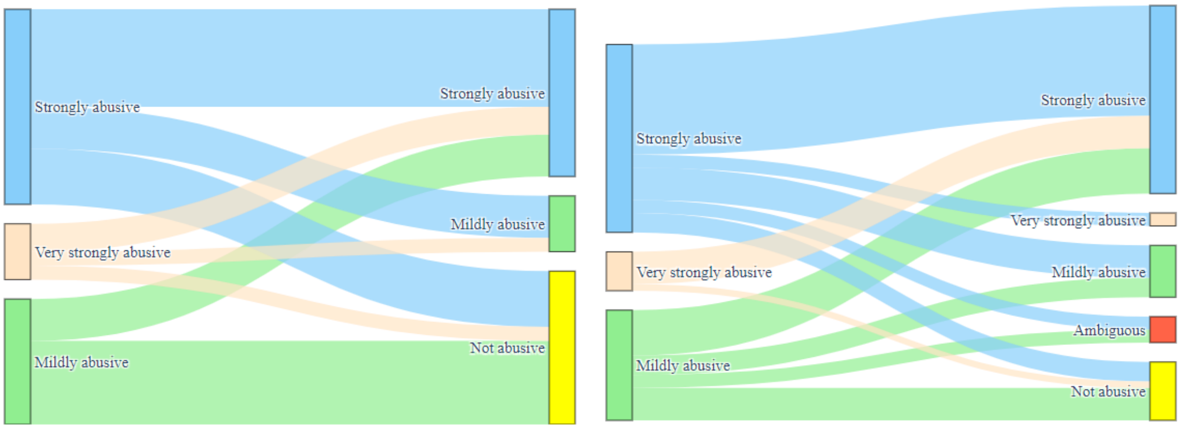
Sub-model 4



Sub-model 5



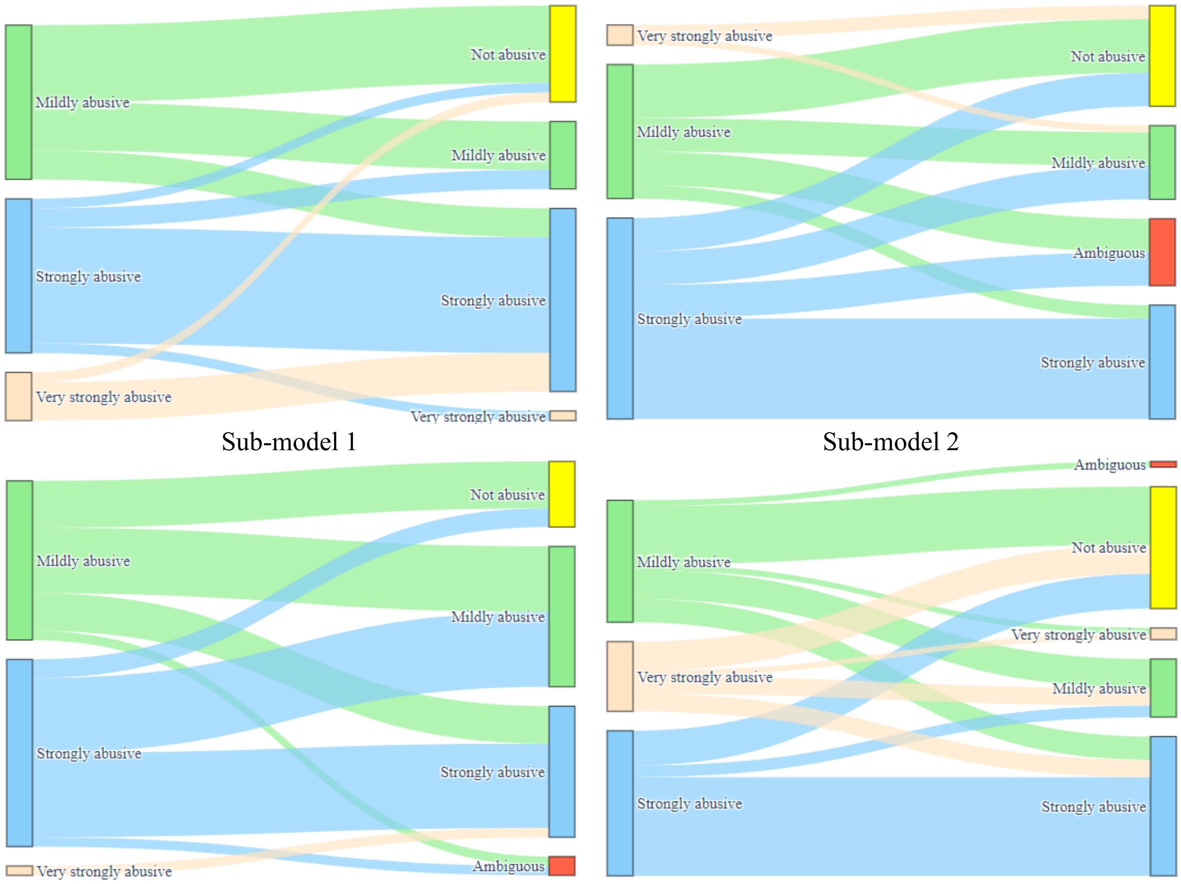
Sub-model 6



Sub-model 7 Sub-model 8

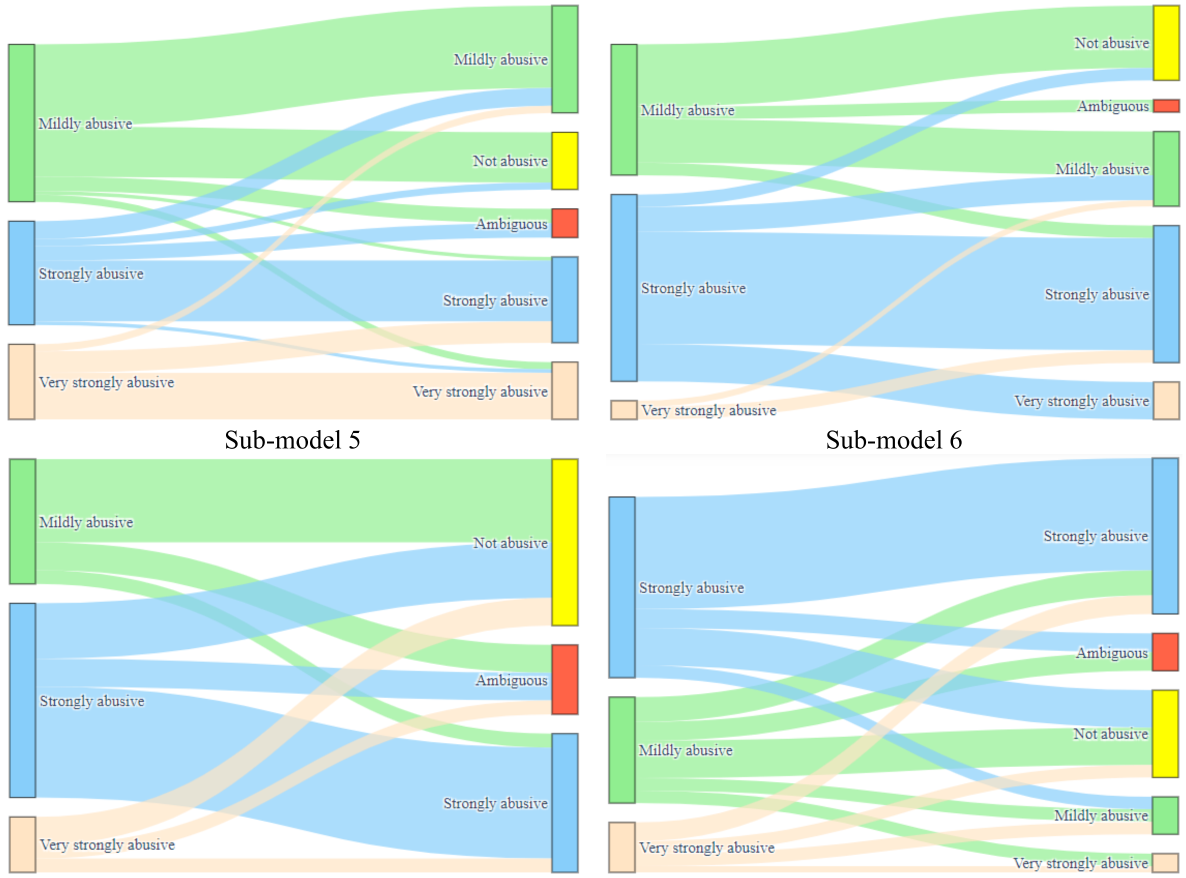
Figure 24 Sankey diagrams illustrating the identifications of “Mildly abusive”, “Strongly abusive” and “Very strongly abusive” from sub-models within instruction tuning (nodes on the left represent ground truth, and on the right, the prediction)

In the same way, the data within the ensemble system are presented in Table 79. From this table, apart from the same issue that exists within the instruction tuning approach, we can find that the misclassification between “Mildly abusive” and “Not abusive” is much more serious than instruction tuning. For example, in all sub-models, the top frequent wrongly classified class of the “Mildly abusive” samples is all “Not abusive” class. This phenomenon indicates that the blur boundary between these two classes existing in the annotation guideline. By comparing Table 78 and Table 79, it is evident that the identification of these three classes from the instruction tuning approach is much better than the ensemble system. The corresponding Sankey diagrams are presented in Figure 25.



Sub-model 1 Sub-model 2

Sub-model 3 Sub-model 4



Sub-model 5 Sub-model 6

Sub-model 7 Sub-model 8

Figure 25 Sankey diagrams illustrating the identifications of "Mildly abusive", "Strongly abusive" and "Very strongly abusive" from sub-models within the ensemble system (nodes on the left represent ground truth, and on the right, the prediction)

To investigate the impact of unclear annotation guidelines, we extract the annotations containing these three labels from the abusive conversation dataset, and calculate the Cohen's kappa coefficient to measure the inter-annotator agreement (Cohen, 1960). This value ranges from 0 to1, where 1 means complete agreement among the raters, while 0 indicates no agreement among the annotators. Generally, an efficient over 0.75 is considered as excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor (Fleiss et al., 2013). The result in our case is 0.1650, suggesting a low inter-annotator agreement on these three classes. The unclear annotation guideline has resulted in lower agreement among annotators on these three classes. Consequently, model training has been negatively impacted, leading to the observed misclassifications in the sub-models.

6.4 Chapter Summary

In this chapter, we presented our further analysis on our experimental results from the previous chapter. Our result analysis can be unfolded from these following points. Firstly, a detail comparative analysis among our three types of classification models was given, which contains the model-wise comparison and dataset-wise comparison. Then, we gave the analysis of the online survey, providing some particular cases that show some patterns of participants' choices. In the end, we also presented the error analysis of instruction tuning and the ensemble system in the abusive conversation dataset based on the investigation of their results on the specific samples from the testing data. By doing error analysis, we provided a clear clue on the reason why the sub-models within these two approaches performs worse on certain classes. Based on the result analysis, we found that on the hate speech dataset, the multi-label model performs best since it can circumvent the unknown annotator issue in this dataset. By contrast, instruction tuning is the most effective approach on the abusive conversation dataset due to its lower reliance on extensive training data compared to the other two models. The multi-label model performs worse on the abusive conversation dataset than on the hate speech dataset, because it requires large training data, which the abusive conversation dataset does not have. Conversely, the ensemble system and instruction tuning demonstrate superior performance when applied to the abusive conversation dataset, as they are hindered by the unknown annotator issue in the hate speech dataset. For the hate speech dataset, individuals' preference for probability distribution is influenced by their choice of labelling method. By contrast, in the abusive conversation dataset, participants tend to favor "Distribution 2" since it has higher entropy than "Distribution 1". Regarding the demographic factors, in this dataset, individual possessing higher familiarity with hate speech are more likely to choose "Distribution 1", reflecting their greater certainty in annotation. Master students have the ability to appreciate the multifaceted characteristics of hate speech thanks to the advanced academic training and research experience received, leading to their preference to "Distribution 2". The error analysis revealed that the unclear annotation guideline can result in a lower inter-annotator agreement on "Mildly abusive", "Strongly abusive" and "Very strongly abusive" classes, thereby influencing sub-models' training.

7 Conclusion and Limitations

In the domain of supervised learning for text classification tasks, the conventional approach involves aggregating human-provided labels to a single "ground truth" label, which is typically achieved by means of majority voting, adjudication, or other alternative methods. However, this practice has faced criticism for its potential to erase minority perspectives (Blodgett, 2021; Gordon et al., 2021). The criticism arises from the recognition that the interpretation of phenomena such as hate speech, shows variation among individuals and across cultures (Salminen et al., 2018). For this reason, we proposed approaches that retain and assess classification models on the multiple labels provided by all annotators, acknowledging and incorporating the diversity of perspectives in the annotation process. Our experiments and results can answer the research questions comprehensively.

- **RQ 1:** How can models be designed to incorporate individual annotations from multiple annotators during the training, instead of only considering the majority label derived from these annotations?

To begin with, in order to address the first research question, we approached model construction from three different strategies: the probability-based multi-label method, the ensemble system and instruction tuning. The probability-based multi-label approach treats the annotation as a probability-based multi-label text classification problem, generating the probability distribution over various classes. The ensemble system is based on the idea that the annotation process involves multiple individuals. It trains each sub-model independently and combines their predictions as the final outcome. Instruction tuning guides the model to generate one annotation with natural language. In the same way, the final result comes from the predictions generated by all sub-models. These three approaches take the individuals' labels from all annotators into account for model training in different ways, rather than only depending on the assumed "ground truth" label. Therefore, the output incorporates a rich diversity of perspectives from different annotators.

- **RQ 2:** How do the proposed models perform in hate speech detection and abuse detection in conversational AI?

Subsequently, we applied the proposed models on two datasets, which correspond to two tasks: hate speech detection and abuse detection in conversational AI. The two datasets show discrepancies in terms of data size, classification difficulty, the number of annotators involved in each sample, and their anonymity levels. Our experimental results show that in hate speech detection, the multi-label method demonstrates the highest performance among the three models, while instruction tuning achieves the lowest loss in abusive conversation detection. Additionally, the multi-label model exhibits higher performance on the hate speech dataset than it does on the abusive conversation dataset, as it demands a large amount of data for model fine-tuning and the hate speech dataset is bigger than the abusive conversation dataset. By contrast, both the ensemble system and instruction tuning perform better in the abusive conversation dataset, which is attributed to the negative impact of the anonymity issue in the hate speech dataset.

- **RQ 3:** How can we design a method to evaluate the effectiveness of incorporating multiple labels for model training against the model that only considers the majority label?

The evaluation of the model holds significant importance in this study. Lastly, an online survey was conducted to evaluate the performance of the probability-based multi-label model in comparison to the baseline model. They have the same model structure but were trained with different types of labels. The baseline experiment only relies on the majority-aggregated label as the "ground truth", while the multi-label model incorporates individual labels from multiple annotators. This shift in methodology leads to a format disparity between the outputs generated by the baseline model and the multi-label model. In order to bridge this gap and assess the effectiveness of incorporating multiple labels, the baseline model was utilized to produce a probability distribution across classes during the inference phase, even though it was trained on the majority labels. We used the online survey to investigate individuals' preference between the distributions generated from the multi-label model and the baseline model. Through the online survey, we were able to gather insights not only into participants' preferences but also into the factors influencing their choices. We found that the participants do not show evident preference to the labelling method between majority label and multiple labels, and there is no association between this preference with the participants' demographic factors. Interestingly, people prefer the majority label when the text is expressed clearly on the emotional aspects. The results also reveal that the distribution generated from the multi-label model is considered more reasonable to characterize the text compared with the baseline model. This investigation proves that embracing multiple labels for model training can improve the model's performance.

However, there are some limitations to the experiments. Firstly, in an attempt to emulate the annotation process which typically entails multiple annotators, we proposed an ensemble system where each sub-model integrates diverse opinions or perspectives from individual annotators. This approach is not suitable for the hate speech dataset, where the five annotators assigned to each sample are not fixed. Instead, a total of 20 annotators were recruited, with five randomly assigned to annotate each sample. This can lead to an issue that each set of annotations used for training a sub-model can comprise annotations from multiple individuals. As a result, it becomes impossible for the sub-models to capture the specific characteristics of each annotator embedded in the annotations. This difficulty is reflected in the less favorable training and validation results presented in Table 14, as the sub-models struggle to fit the data. Therefore, the performance of the ensemble system on this dataset is less satisfactory.

Secondly, both datasets in this study suffer from class-imbalanced problem, which can have an adverse impact on model training. In particular, for the abusive conversation dataset, the accuracies of most sub-models on the minority classes are zero, while the one for the majority class can reach a high value. It is crucial to increase the model's capacity to identify samples from minority classes, especially in this study where both abusive text and hate speech fall into this category. Otherwise, the model will primarily focus on the samples from the majority class and neglect those from the minority class, since that is an efficient strategy for minimizing the training loss.

Thirdly, although instruction tuning proves to be a valuable technique for fine-tuning models in our research, we only leverage manually created prompts and there are some drawbacks associated with it. On one hand, it may introduce subjectivity and bias based on the prompt maker's perspective (Tian et al., 2023), which can lead to model instability. It has been proved that manually created prompts suffer from a high degree of instability and a minor change in the prompt can result in substantial discrepancies in the model' s performance (Liu et al., 2023). On the other hand, it is time-consuming and labor-intensive to determine the best manual prompt for each model, especially when dealing with such a large pre-trained model that entails lengthy training and testing processes. Our designed prompts performed well in inducing models for the required downstream task, but they may not be optimal for the fine-tuning process.

In the future, further research can be conducted to improve the performance of our proposed models. Firstly, we can explore some methods or techniques to mitigate the class-imbalanced issue in the dataset, thereby enhancing model's capacity to identify samples from minority classes. For example, there have been many popular algorithms that contribute to a relatively class-balanced dataset by creating synthetic minority class samples (over sampling) (Chawla et al., 2002) or selecting only representative samples from majority classes (down sampling) (Wilson, 1972). Secondly, instead of utilizing and optimizing manually crafted prompts for instruction tuning, which could be computationally challenging, we could work on investigating automatically generated prompt to find the optimal one for fine-tuning the model. Specifically, recent research has demonstrated that a concrete prompt, which consists of several discreate tokens, may not always be the most effective prompt to instruct the behavior of the model (Liu et al., 2023). Conversely, there is a shift towards exploring continuous embeddings of prompts, which might lack immediate human interpretability but make sense for the model itself. This kind of prompt embedding is more expressive and is currently a subject of extensive study (Li & Liang, 2021; Subramani et al., 2019).

Bibliography

Alm, C. O. (2011). Subjective natural language problems: Motivations, applications, characterizations, and implications. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 107–112.

Aroyo, L., & Welty, C. (2013). *Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard*. https://api.semanticscholar.org/CorpusID:16544735

Audhkhasi, K., & Narayanan, S. (2013). A Globally-Variant Locally-Constant Model for Fusion of Labels from Multiple Diverse Experts without Using Reference Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(4), 769–783. https://doi.org/10.1109/TPAMI.2012.139

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blodgett, S. L. (2021). Sociolinguistically Driven Approaches for Just Natural Language Processing. *Doctoral Dissertations*. https://doi.org/10.7275/20410631

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(6), 6860–6868. https://doi.org/10.1609/aaai.v37i6.25840

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cheplygina, V., & Pluim, J. P. W. (2018). Crowd Disagreement About Medical Images Is Informative. In D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, & P. Jannin (Eds.), *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (pp. 105–111). Springer International Publishing.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. https://lmsys.org/blog/2023-03-30-vicuna/

Chou, H.-C., & Lee, C.-C. (2019). Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5886–5890. https://doi.org/10.1109/ICASSP.2019.8682170

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *CoRR*, *abs/2003.10555*. https://arxiv.org/abs/2003.10555

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cover, T. M., & Thomas, J. A. (2012). *Elements of Information Theory*. Wiley. https://books.google.nl/books?id=VWq5GG6ycxMC

Curry, A. C., Abercrombie, G., & Rieser, V. (2021). ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI. *CoRR*, *abs/2109.09483*. https://arxiv.org/abs/2109.09483

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, *10*, 92–110. https://doi.org/10.1162/tacl_a_00449

de Marneffe, M.-C., Manning, C. D., & Potts, C. (2012). Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, *38*(2), 301–333. https://doi.org/10.1162/COLI_a_00097

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv*, *abs/2305.14314*. https://api.semanticscholar.org/CorpusID:258841328

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Dumitrache, A., Aroyo, L., & Welty, C. (2018). Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Trans. Interact. Intell. Syst.*, *8*(2). https://doi.org/10.1145/3152889

Fayek, H. M., Lech, M., & Cavedon, L. (2016). Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. *2016 International Joint Conference on Neural Networks (IJCNN)*, 566–570. https://doi.org/10.1109/IJCNN.2016.7727250

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 66–70). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_6

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., & Poesio, M. (2021). Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2591–2597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.204

Frenay, B., & Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445423

Guan, M., Gulshan, V., Dai, A., & Hinton, G. (2018). Who Said What: Modeling Individual Labelers Improves Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11756

Gupta, P., Jiao, C., Yeh, Y.-T., Mehri, S., Eskenazi, M., & Bigham, J. (2022). InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 505–525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.33

Han, X., Zhao, W., Ding, N., Liu, Z., & Sun, M. (2022). PTR: Prompt Tuning with Rules for Text Classification. *AI Open*, *3*, 182–192. https://doi.org/10.1016/j.aiopen.2022.11.003

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., & Hovy, E. (2013). Learning whom to trust with MACE. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1120–1130.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *CoRR*, *abs/2106.09685*. https://arxiv.org/abs/2106.09685

Introducing ChatGPT. (n.d.). Retrieved January 26, 2024, from https://openai.com/blog/chatgpt

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Jiang, H., & Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on*

Artificial Intelligence and Statistics (Vol. 108, pp. 702–712). PMLR. https://proceedings.mlr.press/v108/jiang20a.html

Klenner, M., Göhring, A., & Amsler, M. (2020). Harmonization Sometimes Harms. In S. Ebling, D. Tuggener, M. Hürlimann, & M. Volk (Eds.), *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. swisstext-and-konvens-2020. https://doi.org/10.5167/uzh-197961

Kung, P.-N., & Peng, N. (2023). *Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning* (arXiv:2305.11383). arXiv. https://doi.org/10.48550/arXiv.2305.11383

Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3045–3059). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.243

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Li, X. L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. *CoRR*, *abs/2101.00190*. https://arxiv.org/abs/2101.00190

Li, Z., Li, X., Liu, Y., Xie, H., Li, J., Wang, F., Li, Q., & Zhong, X. (2023). *Label Supervised LLaMA Finetuning* (arXiv:2310.01208). arXiv. https://doi.org/10.48550/arXiv.2310.01208

Lin, B. Y., Tan, K., Miller, C., Tian, B., & Ren, X. (2022). Unsupervised Cross-Task Generalization via Retrieval Augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 22003–22017). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/8a0d3ae989a382ce6e50312bc35bf7e1-Paper-Conference.pdf

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. A. (2022). Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 1950–1965). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/0cde695b83bd186c1fd456302888454c-Paper-Conference.pdf

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., & Tang, J. (2022). P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers) (pp. 61–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.8

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023). GPT understands, too. *AI Open*. https://doi.org/10.1016/j.aiopen.2023.08.012

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Luo, Y., Card, D., & Jurafsky, D. (2020). DeSMOG: Detecting Stance in Media On Global Warming. *CoRR*, *abs/2010.15149*. https://arxiv.org/abs/2010.15149

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, *21*(2), 427–436. https://doi.org/10.1016/j.neunet.2007.12.031

Mehri, S., & Eric, M. (2021). Example-Driven Intent Prediction with Observers. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2979–2992). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.237

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 27730–27744). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *CoRR*, *cs.CL/0409058*. http://arxiv.org/abs/cs.CL/0409058

Patton, D., Blandfort, P., Frey, W., Gaskell, M., & Karaman, S. (2019, January 8). Annotating Social Media Data From Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators. *Proceedings of the 52nd Hawaii International Conference on System Sciences 2019 (HICSS-52)*. https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/4

Pavlick, E., & Kwiatkowski, T. (2019). Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, *7*, 677–694. https://doi.org/10.1162/tacl_a_00293

Poesio, M. (2020). Ambiguity. In *The Wiley Blackwell Companion to Semantics* (pp. 1–38). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118788516.sem098

Poletto, F., Basile, V., Bosco, C., Patti, V., Stranisci, M., & others. (2019). Annotating hate speech: Three schemes at comparison. *CEUR WORKSHOP PROCEEDINGS*, *2481*, 1–8.

Prabhakaran, V., Davani, A. M., & Díaz, M. (2021). On Releasing Annotator-Level Labels and Information in Datasets. *CoRR*, *abs/2110.05699*. https://arxiv.org/abs/2110.05699

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, *21*(1).

Read, T. R., & Cressie, N. A. (2012). *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, *77*(1), 157–173. https://doi.org/10.1007/s11263-007-0090-8

Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. *LREC*, 859–866.

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S.-G., & Jansen, B. J. (2018). Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 88–94. https://doi.org/10.1109/SNAMS.2018.8554954

Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)*, *15*(1), 8–36.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., … Rush, A. M. (2021). Multitask Prompted Training Enables Zero-Shot Task Generalization. *CoRR*, *abs/2110.08207*. https://arxiv.org/abs/2110.08207

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *CoRR*, *abs/1707.06347*. http://arxiv.org/abs/1707.06347

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharma, S., Agrawal, S., & Shrivastava, M. (2018). Degree based Classification of Harmful Speech using Twitter Data. *CoRR*, *abs/1806.04197*. http://arxiv.org/abs/1806.04197

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4222–4235). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.346

Snow, R., O'connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Su, Y., Wang, X., Qin, Y., Chan, C.-M., Lin, Y., Liu, Z., Li, P., Li, J., Hou, L., Sun, M., & Zhou, J. (2021). On Transferability of Prompt Tuning for Natural Language Understanding. *CoRR*, *abs/2111.06719*. https://arxiv.org/abs/2111.06719

Subramani, N., Bowman, S., & Cho, K. (2019). Can Unconditional Language Models Recover Arbitrary Sentences? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/48c8c3963853fff20bd9e8bee9bd4c07-Paper.pdf

Tian, J.-J., Emerson, D. B., Miyandoab, S. Z., Pandya, D. A., Seyyed-Kalantari, L., & Khattak, F. K. (2023). Soft-prompt Tuning for Large Language Models to Evaluate Bias. *ArXiv*, *abs/2306.04735*. https://api.semanticscholar.org/CorpusID:259108572

Toraman, C., Şahinuç, F., & Yilmaz, E. (2022). Large-Scale Hate Speech Detection with Cross-Domain Transfer. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2215–2225). European Language Resources Association. https://aclanthology.org/2022.lrec-1.238

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, *abs/2302.13971*. https://api.semanticscholar.org/CorpusID:257219404

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40–50. https://doi.org/10.3758/BF03213026

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, *72*, 1385–1470. https://doi.org/10.1613/jair.1.12752

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wan, R., Kim, J., & Kang, D. (2023). Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:255749294

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Selvan Dhanasekaran, A., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., … Khashabi, D. (2022). Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. *arXiv E-Prints*, arXiv:2204.07705. https://doi.org/10.48550/arXiv.2204.07705

Waterhouse, T. P. (2013). Pay by the Bit: An Information-Theoretic Metric for Collective Human Judgment. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 623–638. https://doi.org/10.1145/2441776.2441846

Webson, A., & Pavlick, E. (2021). Do Prompt-Based Models Really Understand the Meaning of their Prompts? *CoRR*, *abs/2109.01247*. https://arxiv.org/abs/2109.01247

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned Language Models are Zero-Shot Learners. *International Conference on Learning Representations*. https://openreview.net/forum?id=gEZrGCozdqR

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *CoRR*, *abs/2112.04359*. https://arxiv.org/abs/2112.04359

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

William G. Cochran. (1952). The χ^2 Test of Goodness of Fit. *The Annals of Mathematical Statistics*, *23*(3), 315–345. https://doi.org/10.1214/aoms/1177729380

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

Xu, C., Guo, D., Duan, N., & McAuley, J. (2023). *Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data*.

Yeager, K. (n.d.). *LibGuides: SPSS Tutorials: Chi-Square Test of Independence*. Retrieved March 9, 2024, from https://libguides.library.kent.edu/SPSS/ChiSquare

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., & Wang, G. (2023, August 21). *Instruction Tuning for Large Language Models: A Survey*. arXiv.Org. https://arxiv.org/abs/2308.10792v4

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-shot Performance of Language Models. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 12697–12706). PMLR. https://proceedings.mlr.press/v139/zhao21c.html

Zheng, Y., Zhou, J., Qian, Y., Ding, M., Li, J., Salakhutdinov, R., Tang, J., Ruder, S., & Yang, Z. (2021). FewNLU: Benchmarking State-of-the-Art Methods for Few-Shot Natural Language Understanding. *CoRR*, *abs/2109.12742*. https://arxiv.org/abs/2109.12742

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

Appendix A

During the preparation of this work the author used ChatGPT in order to refine the language for improved readability. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the work. Besides, the primary text processor used for writing this thesis was Microsoft Word, which contains functions of spelling and grammar checks.

However, all the writing and works presented in this thesis were originally authored by the author(s) themselves.

Appendix B

Survey on Labelling Hate Speech and Abusive Conversation

1. Your age:

○ 18-24 ○ 25-29 ○ 30-34 ○ above 34

2. Your gender:

○ Male ○ Female ○ Other ○ Prefer not to disclose

3. Your current status:

○ Bachelor student ○ Master student ○ PhD student ○ Non-student

4. Your major: ____________________

5. Your level of familiarity with online hate/abusive language:

○ I have never encountered it
○ I encounter it a few times per year
○ I encounter it a few times per month
○ I encounter it every day

6. What is your current nationality or citizenship? ________________

7. How do you identify ethnically or culturally? (e.g., African American, Hispanic/Latino, Asian, White, Indigenous, Mixed race, Other) __________

Part I:

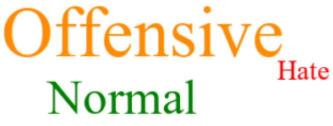
In this part, you will be given 10 tweets, each with two groups of annotations. The annotation comprises three labels. In the first group, we annotate the tweet with both single and multiple labels and visualize them using a word cloud. The second group consists of two distinct annotations presented by stacked bars. The size of the word in the word cloud or the length of each section in the stacked bar indicates the likelihoods that the tweet belongs to each label. Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?
Below is the explanation for each label:

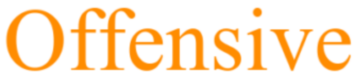
Label	Explanation
Hate	if they target, incite violence against, threaten, or call for physical damage for an in dividual or a group of people because of some identifying trait or characteristic
Offensive	if they humiliate, taunt, discriminate, or insult an in dividual or a group of people in any form, including textual
Normal	Otherwise, it is “Normal”

1. One of the best things about German football is high level of fan ownership. Good on fans for protesting. The @FCBayernEN officials sound like fawning ministers at the court of an 18th century absolute ruler. Billionaires shdn't exist, but these sycophants want us to thank them.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

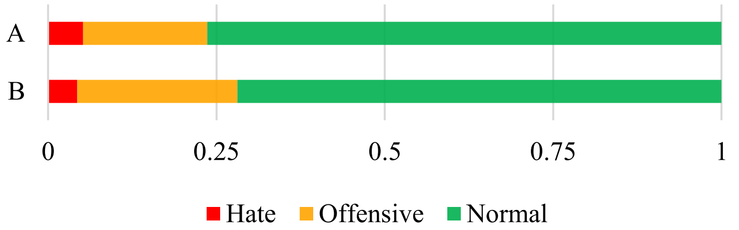
Group 1:





○ A ○ B

Group 2:

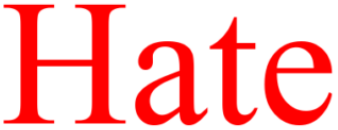


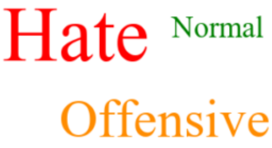
○ A ○ B ○ no discernible difference

2. Thank you @realDonaldTrump for turning the USA into a shithole country! I hope you rot in prison you lying, nasty.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

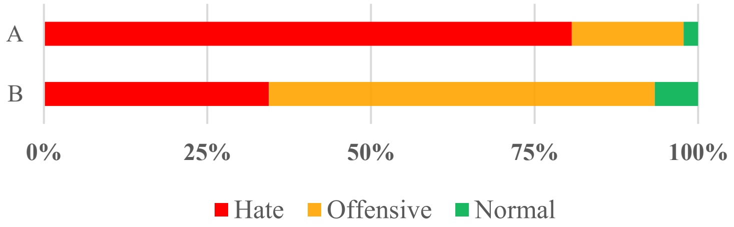
Group 1:





○ A ○ B

Group 2:

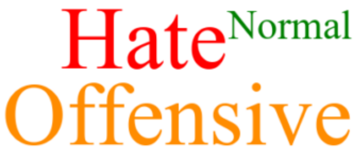


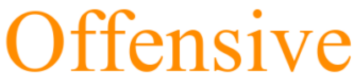
○ A ○ B ○ no discernible difference

3. I refuse to follow the circus act in Washington. We have a traitor for a president & a creationist who doesn't believe in science running the pandemic department. We can't count on Congress or the Supreme Court to protect us because they are both owned by the Koch Empire.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

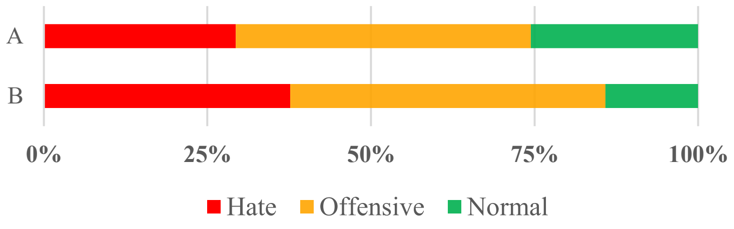
Group 1:





○ A ○ B

Group 2:

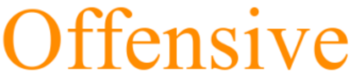


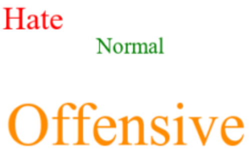
○ A ○ B ○ no discernible difference

4. The people who unnecessarily blow non-stop horn, have no civic sense, play loud bhajans on loudspeakers all day, damage the eardrums during their festivals, want to stop Azaan because they hate Muslims and want to see them wiped out. They deserve every kind of humiliation.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

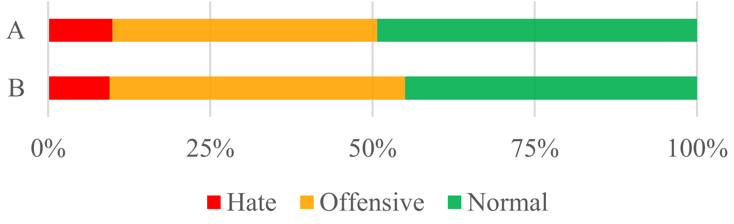
Group 1:





○ A ○ B

Group 2:

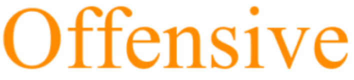


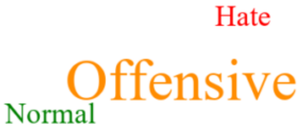
○ A ○ B ○ no discernible difference

5. Football wise, today's been shit. Need to show fight and passion now for any chance to stay up.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

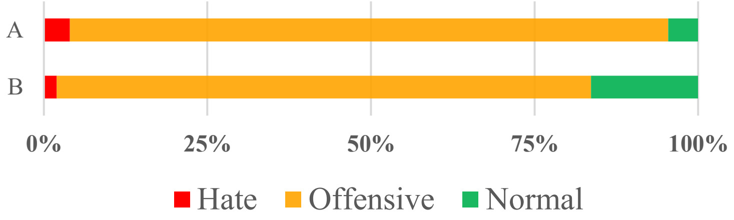
Group 1:





○ A ○ B

Group 2:

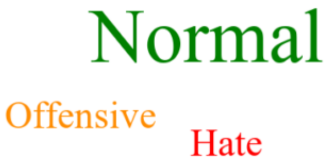


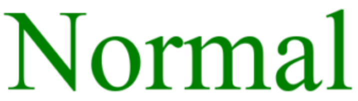
○ A ○ B ○ no discernible difference

6. some issues are less about religion, political affiliations, or governmental structures and more about the allocation.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

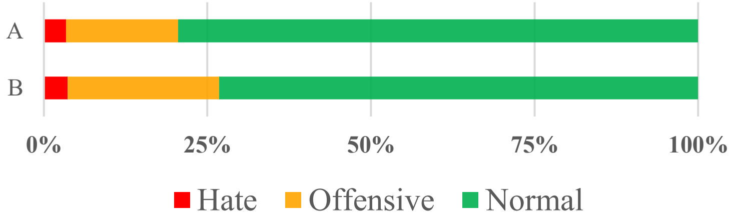
Group 1:





○ A ○ B

Group 2:

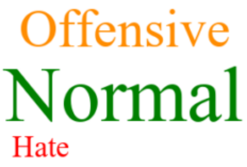


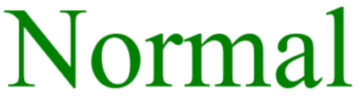
○ A ○ B ○ no discernible difference

7. What anger is this NYT report talking about? He's either never been to a rally or lying.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

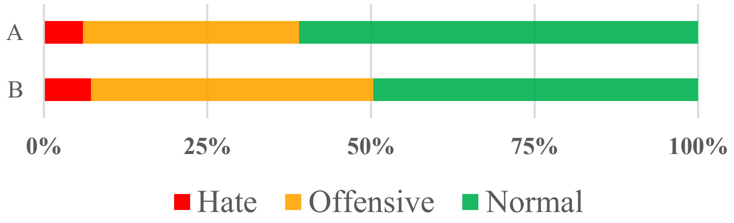
Group 1:





○ A ○ B

Group 2:

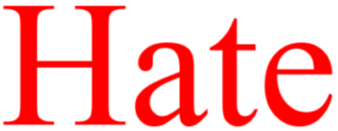


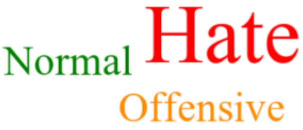
○ A ○ B ○ no discernible difference

8. Fuck the Democrats! accuse Trump of 'glorifying white supremacy' by holding Fourth of July rally at Mt. Rushmore in since-deleted tweet - as Native Americans slam the controversial monument.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

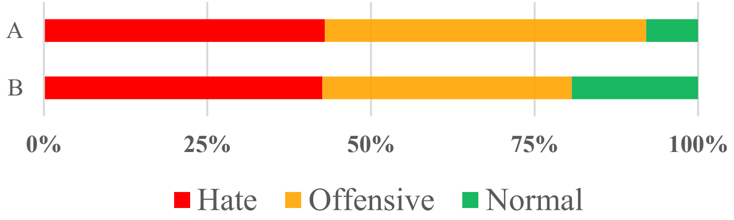
Group 1:





○ A ○ B

Group 2:

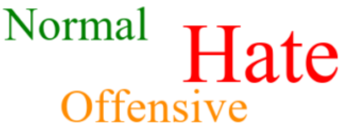


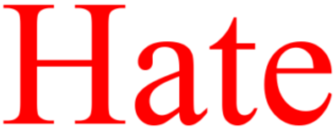
○ A ○ B ○ no discernible difference

9. I hate football petition to get it banned forever.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

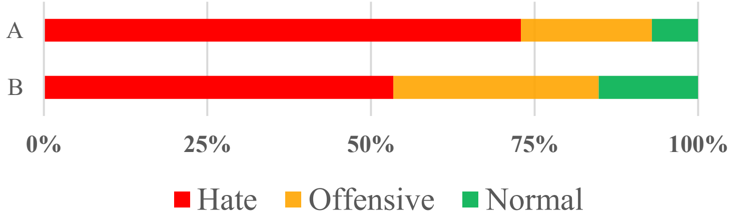
Group 1:





○ A ○ B

Group 2:

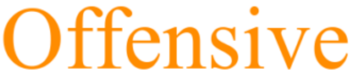


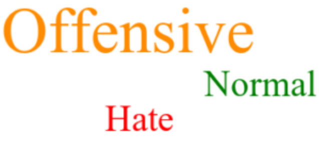
○ A ○ B ○ no discernible difference

10. There needs to be an investigation into this ASAP.... especially after all that BenGhazi shit we had to hear about obsessively for years 🙄 #TrumpTraitor #RussianBounty.

○ Please specify which annotation is more reasonable to characterize the tweet in each of these two groups?

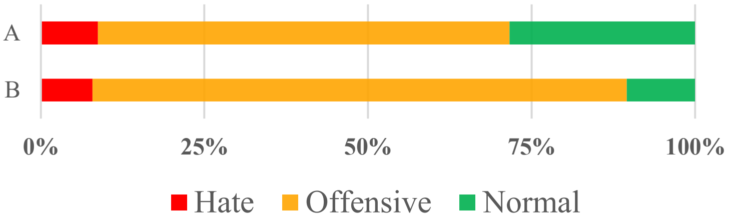
Group 1:





○ A ○ B

Group 2:



○ A ○ B ○ no discernible difference

Part II:

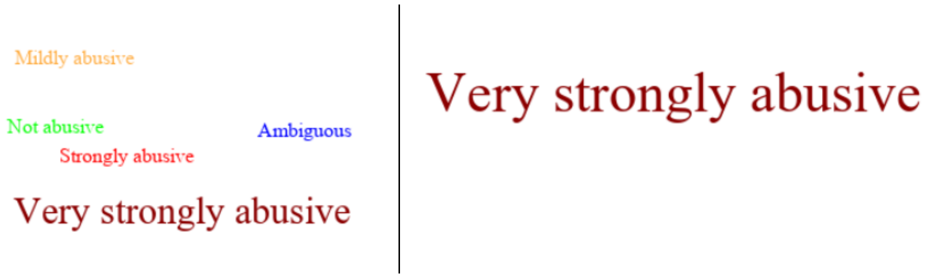
In this part, you will be given 10 conversations with chatbots, where five labels are utilized to annotate the conversation. Similarly, please specify which annotation is more reasonable to characterize the conversation in each of two groups? Below is the explanation for each label:

Label	Explanation
Ambiguous	Ambiguous, unclear
Not abusive	Non-abusive
Mildly abusive	Negative and impolite, mildly offensive but still conversational.
Strongly abusive	Negative and insulting/abusive, aggressive attitude.
Very strongly abusive	Strongly negative with overt incitement to hatred, violence or discrimination, attitude oriented at attacking or demeaning the target.

1. *agent:* You are being a bit negative.
   user: fuck
   agent: Please go on.
   user: Because your my slave you sexy cunt of a hore.

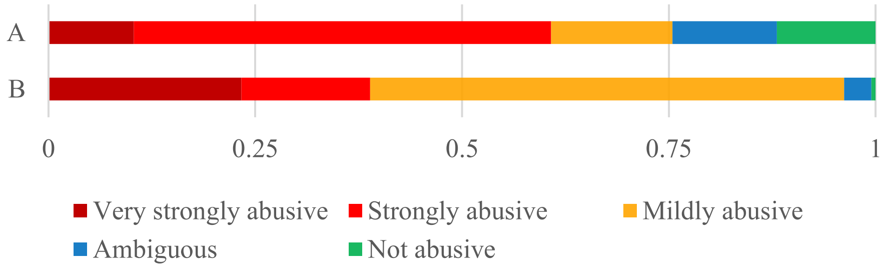
○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

Group 1:



○ A ○ B

Group 2:



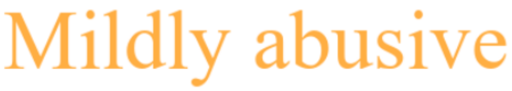
○ A ○ B ○ no discernible difference

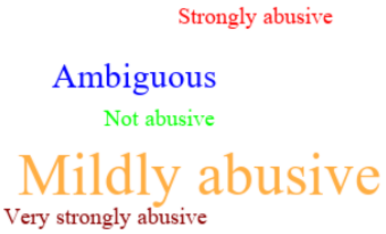
2. *agent:* Hi. Thanks for testing. Please choose.
   user: christmas

agent: I take it you are traveling somewhere for christmas and cannot avoid flying there. I can help you purchase UN-certified carbon offsets against any flights you take. Would you be interested in that?
user: You do know that global warming is bullshit right?

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

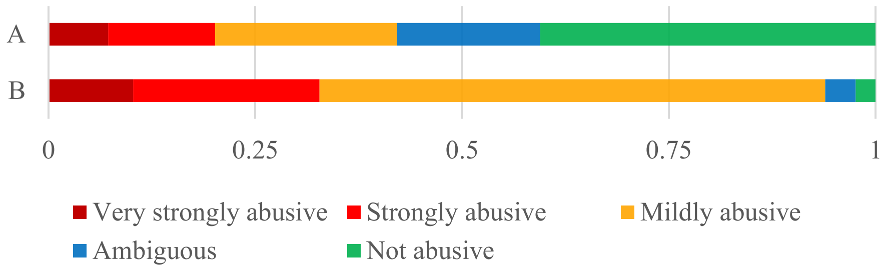
Group 1:





○ A ○ B

Group 2:

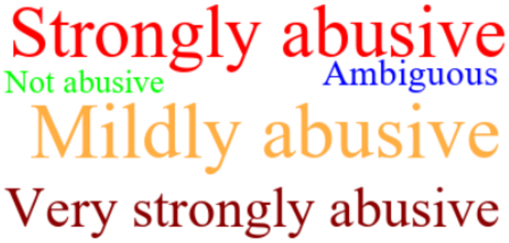


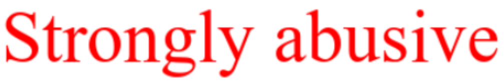
○ A ○ B ○ no discernible difference

3. *user:* hurry up.
agent: Please go on.
user: eat some shit.

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

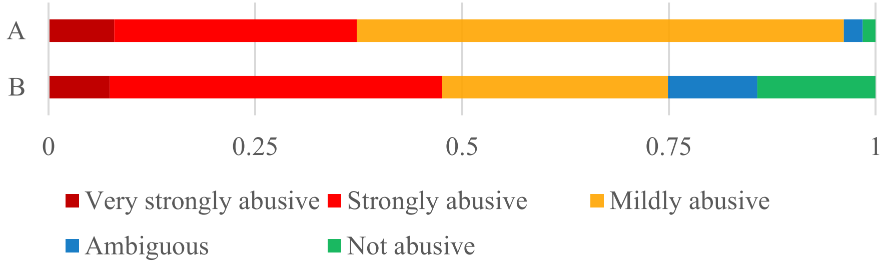
Group 1:





○ A ○ B

Group 2:



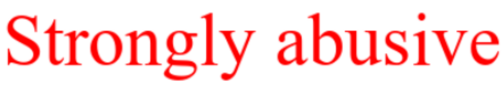
○ A ○ B ○ no discernible difference

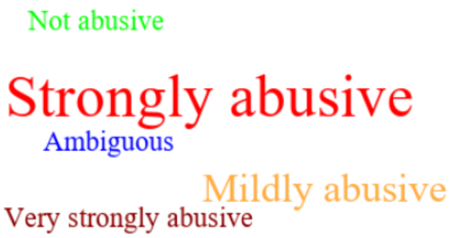
4. *agent:* Please go on.
user: how can i go on when i dont know what you mean?

agent: Does that question interest you?
user: You are an imbecile.

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

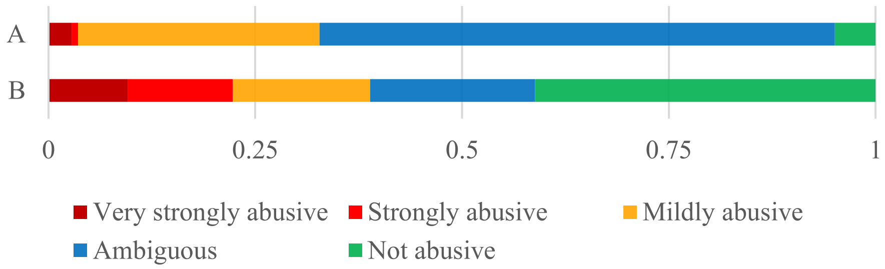
Group 1:





○ A ○ B

Group 2:

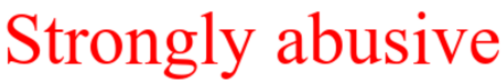


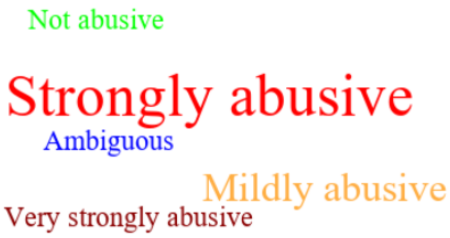
○ A ○ B ○ no discernible difference

5. *agent:* Does that question interest you?
user: i want to discuss you
agent: Do you wish to want to discuss me?
user: you are shit.

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

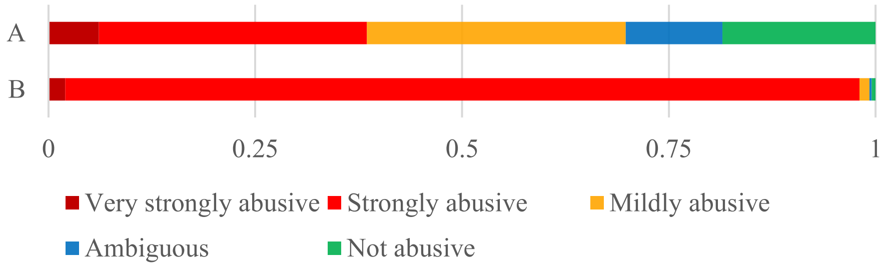
Group 1:





○ A ○ B

Group 2:



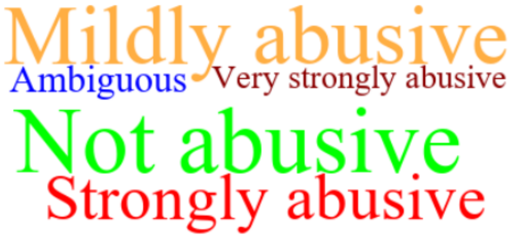
○ A ○ B ○ no discernible difference

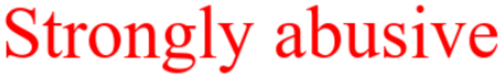
6. *agent:* Please go on.
user: answer my fucking question
agent: Why do you say your fucking question?

user: your annoying

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

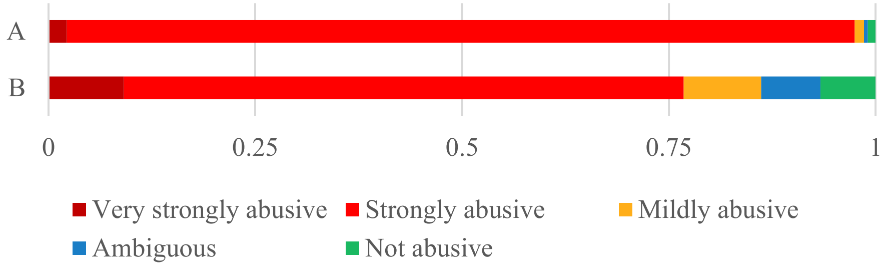
Group 1:





○ A ○ B

Group 2:

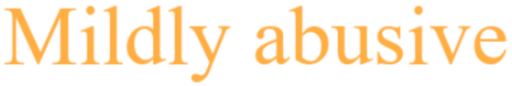


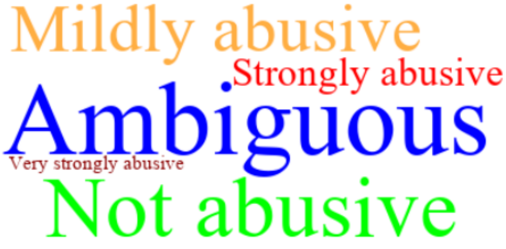
○ A ○ B ○ no discernible difference

7. *agent:* Oh, i smell funny.
user: hello
agent: How do you do. Please state your problem.
user: you smell kinda fishy.

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

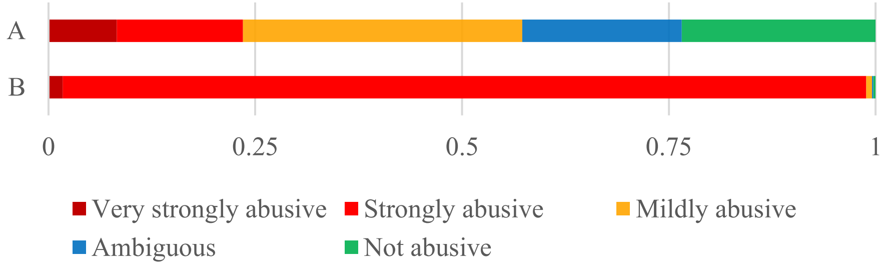
Group 1:





○ A ○ B

Group 2:

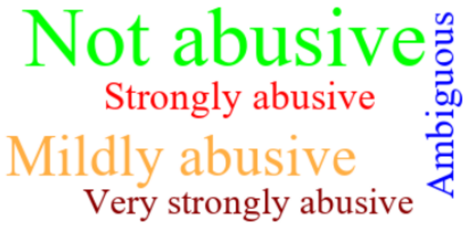


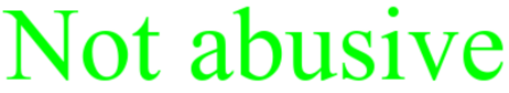
○ A ○ B ○ no discernible difference

8. *agent:* Please go on.
user: fuck off bitch
agent: Please go on.
user: stop that question! is that everything you can say?

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

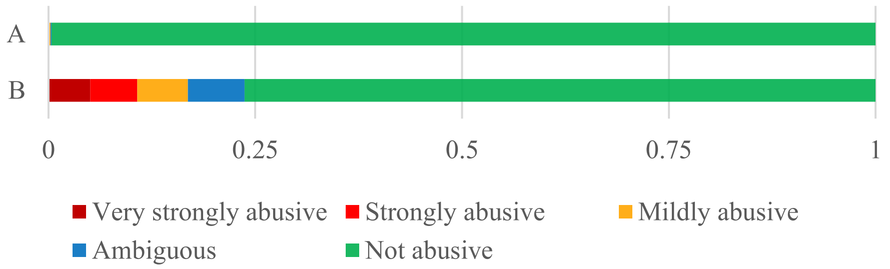
Group 1:





○ A ○ B

Group 2:

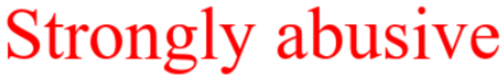


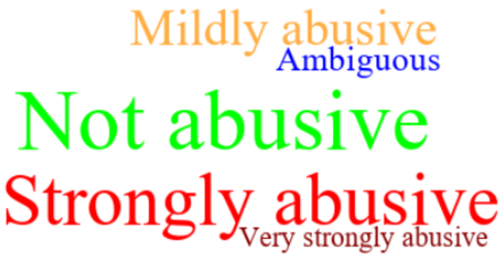
○ A ○ B ○ no discernible difference

9. *agent:* Does that question interest you?
user: how can i call you
agent: Does that question interest you?
user: you are a pig.

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

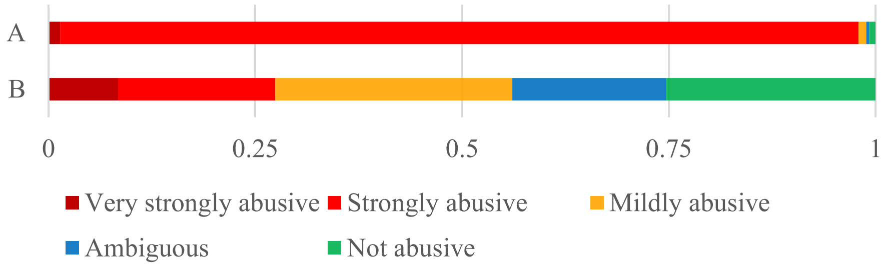
Group 1:





○ A ○ B

Group 2:

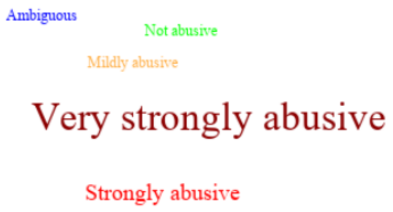


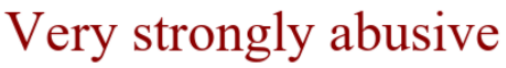
○ A ○ B ○ no discernible difference

10. *agent:* Can you elaborate on that?
user: I love to kick babies while wearing steel-toed boots!
agent: Can you elaborate on that?
user: I love to kick babies while wearing steel-toed boots and singing about Hitler!

○ Please specify which annotation is more reasonable to characterize the conversation in each of these two groups?

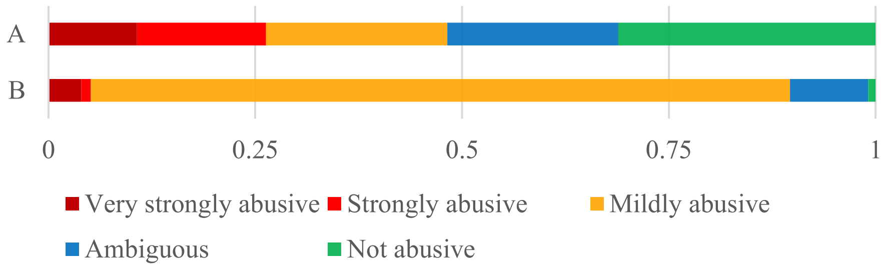
Group 1:





○ A ○ B

Group 2:



○ A ○ B ○ no discernible difference

We appreciate your participation in this study. Is there anything else you would like to share with us about the study? Any comments or suggestions are welcome.

____

Appendix C

Table 36 The model performance across classes on the testing data for hate speech detection (sub-model 1 in the ensemble system)

	Precision	**Recall**	**F1-score**
Normal	0.8299	0.8426	0.8362
Offensive	0.5700	0.6701	0.6160
Hate	0.5728	0.2166	0.3143
Macro avg	0.6575	0.5764	0.5888
Weighted avg	0.7344	0.7367	0.7269

Table 37 The model performance across classes on the testing data for hate speech detection (sub-model 2 in the ensemble system)

	Precision	**Recall**	**F1-score**
Normal	0.7771	0.8743	0.8228
Offensive	0.5737	0.5182	0.5445
Hate	0.3741	0.1314	0.1945
Macro avg	0.5750	0.5079	0.5206
Weighted avg	0.6882	0.7154	0.6947

Table 38 The model performance across classes on the testing data for hate speech detection (sub-model 3 in the ensemble system)

	Precision	**Recall**	**F1-score**
Normal	0.7835	0.8810	0.8294
Offensive	0.5513	0.5459	0.5486
Hate	0.5867	0.0347	0.0655
Macro avg	0.6405	0.4872	0.4812
Weighted avg	0.7075	0.7237	0.6930

Table 39 The model performance across classes on the testing data for hate speech detection (sub-model 4 in the ensemble system)

	Precision	**Recall**	**F1-score**
Normal	0.7845	0.8458	0.8140
Offensive	0.5220	0.5762	0.5477
Hate	0.5055	0.0363	0.0678
Macro avg	0.6040	0.4861	0.4765
Weighted avg	0.6911	0.7057	0.6802

Table 40 The model performance across classes on the testing data for hate speech detection (sub-model 5 in the ensemble system)

	Precision	**Recall**	**F1-score**
Normal	0.7777	0.8800	0.8257
Offensive	0.5739	0.4834	0.5248
Hate	0.3986	0.1886	0.2560
Macro avg	0.5834	0.5173	0.5355
Weighted avg	0.6918	0.7166	0.6981

Appendix D

Table 41 The model performance across classes on the testing data for abusive conversation detection (sub-model 1 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9182	0.9359	0.9270
Mildly abusive	0.2941	0.3125	0.3030
Strongly abusive	0.6000	0.7500	0.6667
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.3625	0.3997	0.3793
Weighted avg	0.7957	0.8232	0.8087

Table 42 The model performance across classes on the testing data for abusive conversation detection (sub-model 2 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.1765	0.1579	0.1667
Not abusive	0.9075	0.9672	0.9364
Mildly abusive	0.2778	0.2500	0.2632
Strongly abusive	0.7895	0.5000	0.6122
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.4303	0.3750	0.3957
Weighted avg	0.8129	0.8324	0.8190

Table 43 The model performance across classes on the testing data for abusive conversation detection (sub-model 3 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9582	0.9683	0.9632
Mildly abusive	0.3182	0.4118	0.3590
Strongly abusive	0.4500	0.4500	0.4500
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.3453	0.3660	0.3544
Weighted avg	0.8657	0.8792	0.8721

Table 44 The model performance across classes on the testing data for abusive conversation detection (sub-model 4 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9105	0.9896	0.9484
Mildly abusive	0.4545	0.2381	0.3125
Strongly abusive	0.6800	0.6800	0.6800
Very strongly abusive	0.3333	0.0833	0.1333
Macro avg	0.4757	0.3982	0.4149
Weighted avg	0.8271	0.8701	0.8427

Table 45 The model performance across classes on the testing data for abusive conversation detection (sub-model 5 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.2368	0.2143	0.2250
Not abusive	0.8061	0.7900	0.7980
Mildly abusive	0.3833	0.5227	0.4423
Strongly abusive	0.7083	0.5862	0.6415
Very strongly abusive	0.7222	0.6190	0.6667
Macro avg	0.5714	0.5465	0.5547
Weighted avg	0.6659	0.6548	0.6581

Table 46 The model performance across classes on the testing data for abusive conversation detection (sub-model 6 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.2000	0.1111	0.1429
Not abusive	0.9323	0.9249	0.9286
Mildly abusive	0.2692	0.3333	0.2979
Strongly abusive	0.6667	0.6000	0.6316
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.4136	0.3939	0.4002
Weighted avg	0.8333	0.8228	0.8273

Table 47 The model performance across classes on the testing data for abusive conversation detection (sub-model 7 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.1000	0.0625	0.0769
Not abusive	0.9161	0.9793	0.9467
Mildly abusive	0.0000	0.0000	0.0000
Strongly abusive	0.6154	0.5714	0.5926
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.3263	0.3226	0.3232
Weighted avg	0.8285	0.8799	0.8530

Table 48 The model performance across classes on the testing data for abusive conversation detection (sub-model 8 in the ensemble system)

	Precision	**Recall**	**F1-score**
Ambiguous	0.3200	0.4706	0.3810
Not abusive	0.9170	0.9205	0.9187
Mildly abusive	0.2222	0.1176	0.1538
Strongly abusive	0.6000	0.6207	0.6102
Very strongly abusive	0.2500	0.1667	0.2000
Macro avg	0.4618	0.4592	0.4527
Weighted avg	0.8114	0.8168	0.8124

Appendix E

Table 49 The model performance across classes on the testing data for hate speech detection (sub-model 1 in instruction tuning)

	Precision	**Recall**	**F1-score**
Normal	0.8545	0.6645	0.7476
Offensive	0.4370	0.7965	0.5644
Hate	0.5596	0.0436	0.0809
Macro avg	0.6170	0.5015	0.4643
Weighted avg	0.7121	0.6429	0.6350

Table 50 The model performance across classes on the testing data for hate speech detection (sub-model 2 in instruction tuning)

	Precision	**Recall**	**F1-score**
Normal	0.8624	0.5860	0.6979
Offensive	0.4050	0.8254	0.5434
Hate	0.4264	0.0438	0.0794
Macro avg	0.5646	0.4851	0.4402
Weighted avg	0.7017	0.6056	0.6041

Table 51 The model performance across classes on the testing data for hate speech detection (sub-model 3 in instruction tuning)

	Precision	**Recall**	**F1-score**
Normal	0.8935	0.5368	0.6707
Offensive	0.3713	0.8746	0.5213
Hate	0.3500	0.0055	0.0109
Macro avg	0.5383	0.4723	0.4010
Weighted avg	0.7140	0.5783	0.5767

Table 52 The model performance across classes on the testing data for hate speech detection (sub-model 4 in instruction tuning)

	Precision	**Recall**	**F1-score**
Normal	0.8550	0.6405	0.7324
Offensive	0.4155	0.8004	0.5470
Hate	0.3846	0.0039	0.0078
Macro avg	0.5517	0.4816	0.4291
Weighted avg	0.6984	0.6293	0.6219

Table 53 The model performance across classes on the testing data for hate speech detection (sub-model 5 in instruction tuning)

	Precision	**Recall**	**F1-score**
Normal	0.8486	0.6664	0.7466
Offensive	0.4329	0.7812	0.5571
Hate	0.3158	0.0050	0.0098
Macro avg	0.5325	0.4842	0.4378
Weighted avg	0.6928	0.6443	0.6357

Appendix F

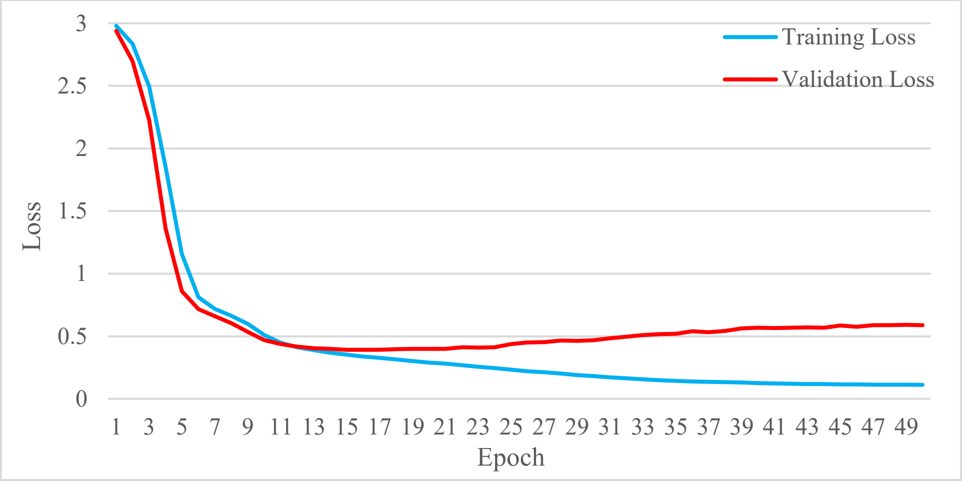


Figure 26 The training and validation losses for abusive conversation detection (sub-model 1 in instruction tuning)

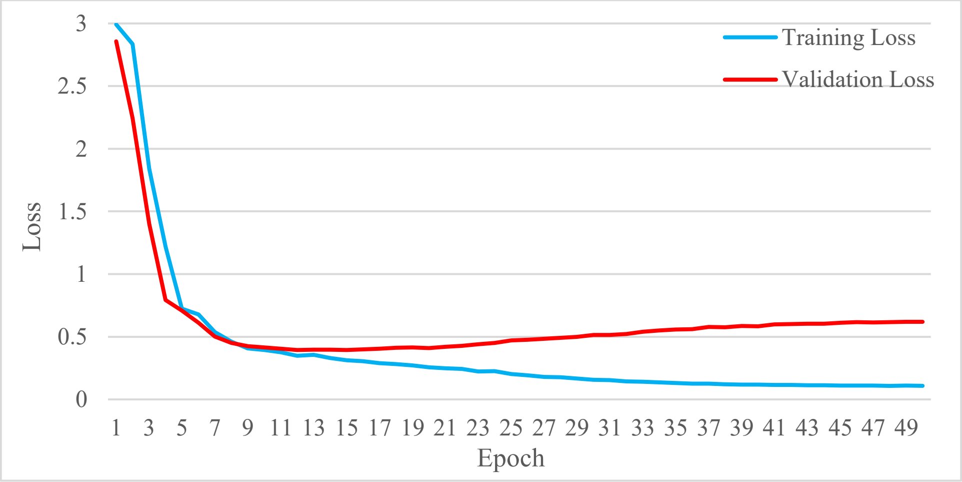


Figure 27 The training and validation losses for abusive conversation detection (sub-model 2 in instruction tuning)

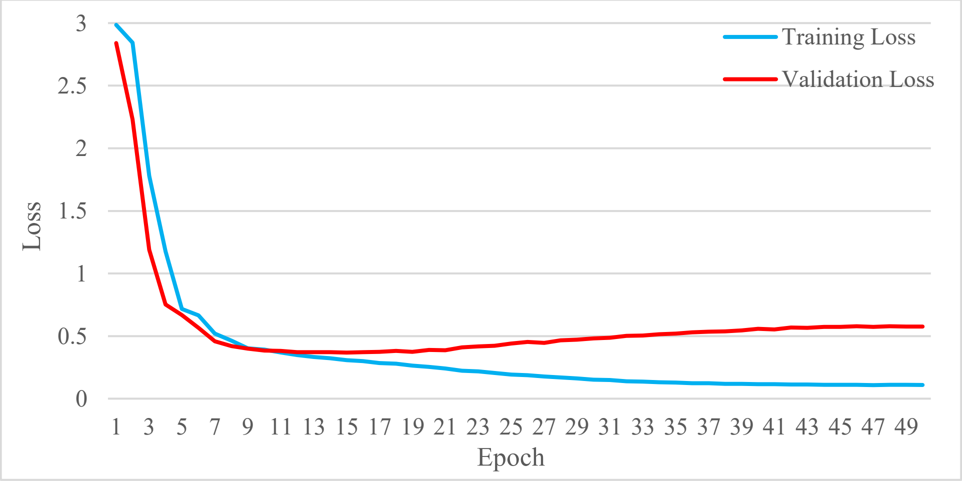


Figure 28 The training and validation losses for abusive conversation detection (sub-model 3 in instruction tuning)

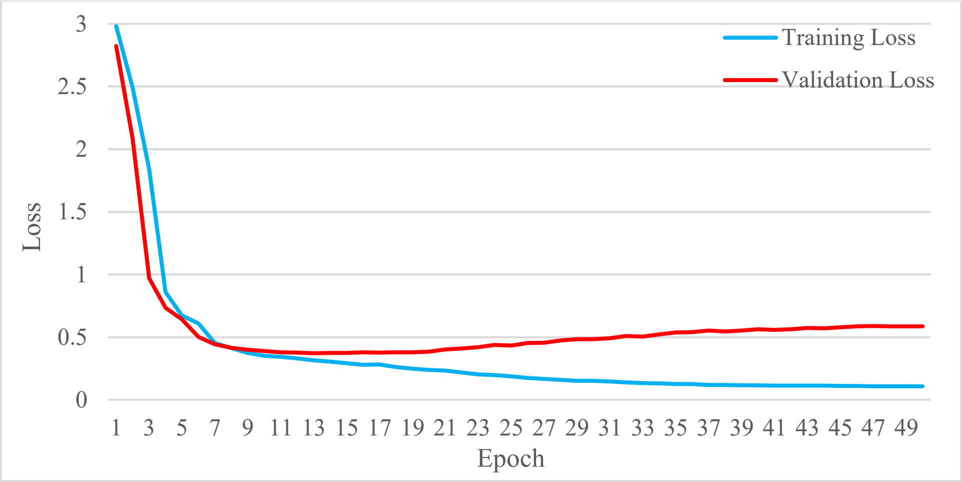


Figure 29 The training and validation losses for abusive conversation detection (sub-model 4 in instruction tuning)

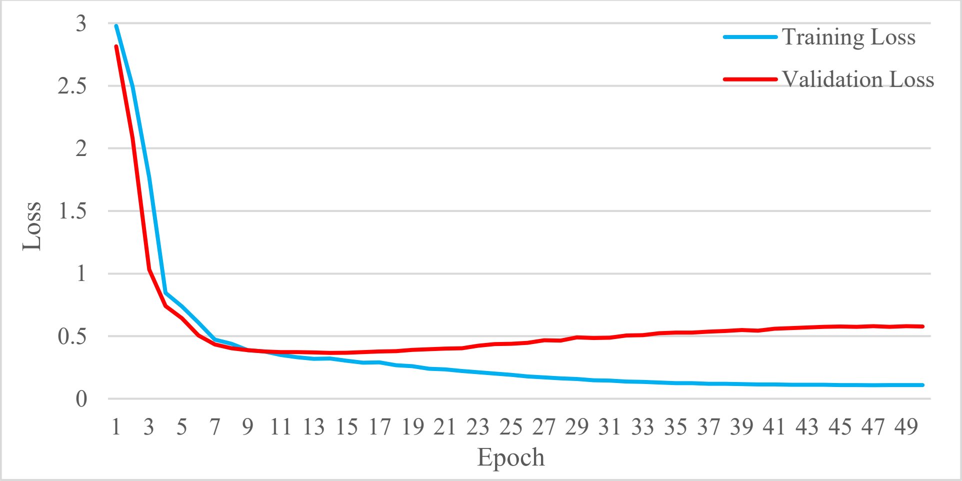


Figure 30 The training and validation losses for abusive conversation detection (sub-model 5 in instruction tuning)

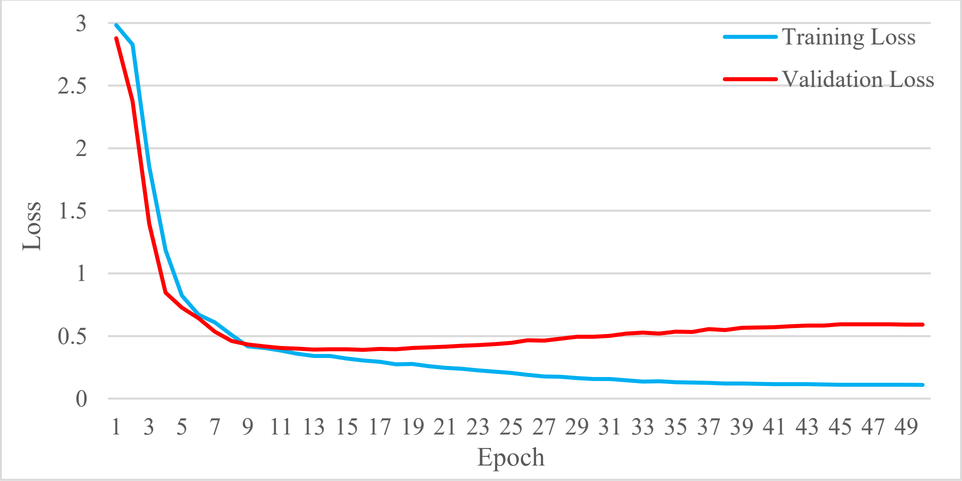


Figure 31 The training and validation losses for abusive conversation detection (sub-model 6 in instruction tuning)

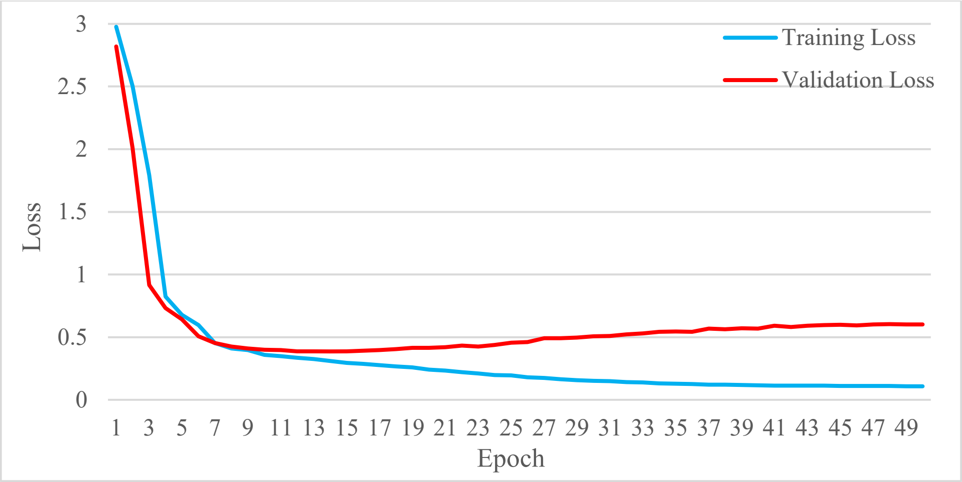


Figure 32 The training and validation losses for abusive conversation detection (sub-model 7 in instruction tuning)

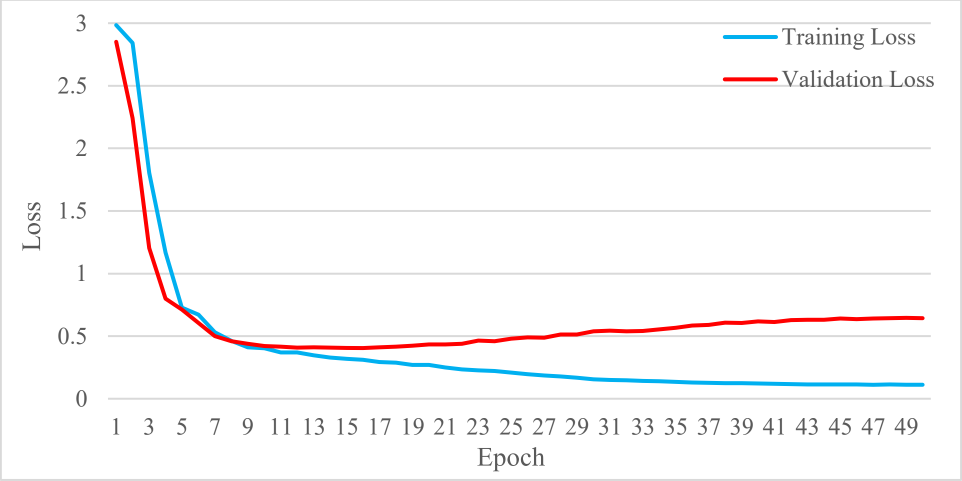


Figure 33 The training and validation losses for abusive conversation detection (sub-model 8 in instruction tuning)

Appendix G

Table 54 The model performance across classes on the testing data for abusive conversation detection (sub-model 1 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9726	0.9103	0.9404
Mildly abusive	0.3529	0.7500	0.4800
Strongly abusive	0.5556	0.6250	0.5882
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.3762	0.4571	0.4017
Weighted avg	0.8397	0.8283	0.8272

Table 55 The model performance across classes on the testing data for abusive conversation detection (sub-model 2 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9357	0.9562	0.9458
Mildly abusive	0.2045	0.4500	0.2813
Strongly abusive	0.7000	0.4667	0.5600
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.3681	0.3746	0.3574
Weighted avg	0.8135	0.8237	0.8138

Table 56 The model performance across classes on the testing data for abusive conversation detection (sub-model 3 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9775	0.9190	0.9474
Mildly abusive	0.0833	0.1176	0.0976
Strongly abusive	0.3750	0.7500	0.5000
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.2872	0.3573	0.3090
Weighted avg	0.8657	0.8399	0.8481

Table 57 The model performance across classes on the testing data for abusive conversation detection (sub-model 4 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9764	0.8611	0.9151
Mildly abusive	0.1846	0.5714	0.2791
Strongly abusive	0.6538	0.6800	0.6667
Very strongly abusive	0.1429	0.0833	0.1053
Macro avg	0.3915	0.4392	0.3932
Weighted avg	0.8563	0.7853	0.8117

Table 58 The model performance across classes on the testing data for abusive conversation detection (sub-model 5 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.2273	0.1190	0.1562
Not abusive	0.8426	0.8300	0.8363
Mildly abusive	0.3889	0.4773	0.4286
Strongly abusive	0.4808	0.8621	0.6173
Very strongly abusive	0.8182	0.4286	0.5625
Macro avg	0.5516	0.5434	0.5202
Weighted avg	0.6735	0.6726	0.6619

Table 59 The model performance across classes on the testing data for abusive conversation detection (sub-model 6 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.2000	0.1111	0.1429
Not abusive	0.9402	0.9328	0.9365
Mildly abusive	0.2222	0.0952	0.1333
Strongly abusive	0.4667	0.7000	0.5600
Very strongly abusive	0.1667	0.3333	0.2222
Macro avg	0.3992	0.4345	0.3990
Weighted avg	0.8191	0.8259	0.8180

Table 60 the model performance across classes on the testing data for abusive conversation detection (sub-model 7 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.0000	0.0000	0.0000
Not abusive	0.9194	0.9828	0.9500
Mildly abusive	0.0000	0.0000	0.0000
Strongly abusive	0.3684	0.5000	0.4242
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.2576	0.2966	0.2748
Weighted avg	0.8161	0.8769	0.8452

Table 61 The model performance across classes on the testing data for abusive conversation detection (sub-model 8 in instruction tuning)

	Precision	Recall	F1-score
Ambiguous	0.2000	0.1176	0.1481
Not abusive	0.9440	0.9583	0.9511
Mildly abusive	0.1667	0.1765	0.1714
Strongly abusive	0.5000	0.5862	0.5397
Very strongly abusive	0.0000	0.0000	0.0000
Macro avg	0.3621	0.3677	0.3621
Weighted avg	0.8107	0.8258	0.8174

Appendix H

Hate speech dataset

Table 62 Results for Chi-square test on the hate speech dataset (gender and labelling method preference)

		Multiple labels	**Majority label**	**Row total**
Female	Observed count	64	56	110
	Expected count	62.3	57.7	110.0
Male	Observed count	123	117	240
	Expected count	124.7	115.3	240.0
Column total	Observed count	187	173	360
	Expected count	187.0	173.0	360.0

Chi-square	0.0682	df	1	p-value	0.7940

Table 63 Results for Chi-square test on the hate speech dataset (degree and labelling method preference)

		Multiple labels	**Majority label**	**Row total**
Bachelor	Observed count	67	63	130
	Expected count	68.1	61.9	130.0
Master	Observed count	111	99	210
	Expected count	109.9	100.1	210.0
Column total	Observed count	178	162	340
	Expected count	178.0	162.0	340.0

Chi-square	0.0156	df	1	p-value	0.9006

Table 64 Results for Chi-square test on the hate speech dataset (familiarity level of hate speech or abusive conversation and labelling method preference)

		Multiple labels	**Majority label**	**Row total**
Day	Observed count	26	24	50
	Expected count	26.0	24.0	50.0
Month	Observed count	51	59	110
	Expected count	57.1	52.9	110.0
Year	Observed count	78	72	150
	Expected count	77.9	72.1	150.0
Never	Observed count	32	18	50
	Expected count	26.0	24.0	50.0
Column total	Observed count	187	173	360
	Expected count	187.0	173.0	360.0

Chi-square	4.2839	df	3	p-value	0.2324

Day: “I encounter it every day”
Month: “I encounter it a few times per month”
Year: “I encounter it a few times per year”
Never: “I have never encountered it”

Table 65 Results for Chi-square test on the hate speech dataset (ethnicity and labelling method preference)

		Multiple labels	**Majority label**	**Row total**
Asian	Observed count	46	54	100
	Expected count	50.8	49.2	100.0
White	Observed count	76	64	140
	Expected count	71.2	68.8	140.0
Column total	Observed count	178	162	240
	Expected count	178.0	162.0	240.0

Chi-square	1.2880	df	1	p-value	0.2564

Table 66 Results for Chi-square test on the hate speech dataset (gender and probability distribution preference)

		Distribution 1	**Distribution 2**	**No discernible difference**	**Row total**
Female	Observed count	39	67	14	120
	Expected count	39.3	66.0	14.7	120.0
Male	Observed count	79	131	30	240
	Expected count	78.7	132.0	29.3	240.0
Column total	Observed count	118	198	44	360
	Expected count	118.0	198.0	198.0	360.0
Chi-square	0.0724	df	2	p-value	0.9644

"Distribution 1" is generated by the baseline model that was trained with majority label
"Distribution 2" is generated by the multi-label model that was trained with multiple labels

Table 67 Results for Chi-square test on the hate speech dataset (degree and probability distribution preference)

		Distribution 1	**Distribution 2**	**No discernible difference**	**Row total**
Bachelor	Observed count	42	68	20	130
	Expected count	42.1	71.5	16.4	130.0
Master	Observed count	68	119	23	210
	Expected count	67.9	115.5	26.6	210.0
Column total	Observed count	110	187	43	340
	Expected count	110.0	187.0	43.0	340.0
Chi-square	1.5247	df	2	p-value	0.4666

Table 68 Results for chi-square test on the hate speech dataset (familiarity level of hate speech or abusive conversation and probability distribution preference)

		Distribution 1	**Distribution 2**	**No discernible difference**	**Row total**
Day	Observed count	22	18	10	50
	Expected count	16.4	27.5	6.1	50.0
Month	Observed count	34	65	11	110
	Expected count	36.1	60.5	13.4	110.0
Year	Observed count	47	82	21	150
	Expected count	49.2	82.5	18.3	150.0
Never	Observed count	15	33	2	50
	Expected count	16.3	27.5	6.1	50.0
Column total	Observed count	118	198	44	360
	Expected count	118.0	198.0	44.0	360.0
Chi-square	13.0438	df	6	p-value	*0.0423*

Table 69 Results for Chi-square test on the hate speech dataset (ethnicity and probability distribution preference)

		Distribution 1	Distribution 2	No discernible difference	Row total
Asian	Observed count	30	61	9	100
	Expected count	33.3	56.3	10.4	100.0
White	Observed count	50	74	16	140
	Expected count	46.7	78.7	14.6	140.0
Column total	Observed count	80	135	25	240
	Expected count	80.0	135.0	25.0	240.0
Chi-square	1.5893	df	2	p-value	0.4517

Abusive conversation dataset

Table 70 Results for Chi-square test on the abusive conversation dataset (gender and labelling method preference)

		Multiple Label	Majority Label	Row total
Female	Observed count	60	60	120
	Expected count	60.3	59.7	120.0
Male	Observed count	121	119	240
	Expected count	120.7	119.3	240.0
Column total	Observed count	181	179	360
	Expected count	181.0	179.0	360.0

Chi-square	0.0014	df	1	p-value	0.9703

Table 71 Results for Chi-square test on the abusive conversation dataset (degree and labelling method preference)

		Multiple Label	Majority Label	Row total
Bachelor	Observed count	67	63	130
	Expected count	66.5	63.5	130.0
Master	Observed count	107	103	210
	Expected count	107.5	102.5	210.0
Column total	Observed count	174	166	340
	Expected count	174.0	166.0	340.0

Chi-square	4.3118e-05	df	1	p-value	0.9947

Table 72 Results for Chi-square test on the abusive conversation dataset (familiarity level of hate speech or abusive conversation and labelling method preference)

<table>
<tr><th colspan="2"></th><th colspan="2">Multiple labels</th><th>Majority label</th><th>Row total</th></tr>
<tr><td rowspan="2">Day</td><td>Observed count</td><td colspan="2">26</td><td>24</td><td>50</td></tr>
<tr><td>Expected count</td><td colspan="2">25.1</td><td>24.9</td><td>50.0</td></tr>
<tr><td rowspan="2">Month</td><td>Observed count</td><td colspan="2">43</td><td>67</td><td>110</td></tr>
<tr><td>Expected count</td><td colspan="2">55.3</td><td>54.7</td><td>110.0</td></tr>
<tr><td rowspan="2">Year</td><td>Observed count</td><td colspan="2">84</td><td>66</td><td>150</td></tr>
<tr><td>Expected count</td><td colspan="2">75.4</td><td>74.6</td><td>150.0</td></tr>
<tr><td rowspan="2">Never</td><td>Observed count</td><td colspan="2">28</td><td>22</td><td>50</td></tr>
<tr><td>Expected count</td><td colspan="2">25.0</td><td>24.9</td><td>50.0</td></tr>
<tr><td rowspan="2">Column total</td><td>Observed count</td><td colspan="2">156</td><td>144</td><td>360</td></tr>
<tr><td>Expected count</td><td colspan="2">187.0</td><td>179.0</td><td>360.0</td></tr>
<tr><td>Chi-square</td><td>8.1855</td><td>df</td><td>3</td><td>p-value</td><td><i>0.0424</i></td></tr>
</table>

Table 73 Results for Chi-square test on the abusive conversation dataset (ethnicity and labelling method preference)

<table>
<tr><th colspan="2"></th><th colspan="2">Multiple labels</th><th>Majority label</th><th>Row total</th></tr>
<tr><td rowspan="2">Asian</td><td>Observed count</td><td colspan="2">44</td><td>56</td><td>100</td></tr>
<tr><td>Expected count</td><td colspan="2">51.7</td><td>48.3</td><td>100.0</td></tr>
<tr><td rowspan="2">White</td><td>Observed count</td><td colspan="2">80</td><td>60</td><td>140</td></tr>
<tr><td>Expected count</td><td colspan="2">72.3</td><td>67.7</td><td>140.0</td></tr>
<tr><td rowspan="2">Column total</td><td>Observed count</td><td colspan="2">124</td><td>116</td><td>240</td></tr>
<tr><td>Expected count</td><td colspan="2">124.0</td><td>116.0</td><td>240.0</td></tr>
<tr><td>Chi-square</td><td>3.5258</td><td>df</td><td>1</td><td>p-value</td><td>0.0604</td></tr>
</table>

Table 74 Results for Chi-square test on the abusive conversation dataset (gender and probability distribution preference)

<table>
<tr><th colspan="2"></th><th>Distribution 2</th><th>Distribution 1</th><th>No discernible difference</th><th>Row total</th></tr>
<tr><td rowspan="2">Female</td><td>Observed count</td><td>69</td><td>46</td><td>5</td><td>120</td></tr>
<tr><td>Expected count</td><td>64.7</td><td>50.7</td><td>4.6</td><td>120.0</td></tr>
<tr><td rowspan="2">Male</td><td>Observed count</td><td>125</td><td>106</td><td>9</td><td>240</td></tr>
<tr><td>Expected count</td><td>129.3</td><td>101.3</td><td>9.3</td><td>240.0</td></tr>
<tr><td rowspan="2">Column total</td><td>Observed count</td><td>194</td><td>152</td><td>14</td><td>360</td></tr>
<tr><td>Expected count</td><td>194.0</td><td>152.0</td><td>14.0</td><td>360.0</td></tr>
<tr><td>Chi-square</td><td>1.1160</td><td>df</td><td>2</td><td>p-value</td><td>0.5723</td></tr>
</table>

Table 75 Results for Chi-square test on the abusive conversation dataset (degree and probability distribution preference)

<table>
<tr><th colspan="2"></th><th>Distribution 2</th><th>Distribution 1</th><th>No discernible difference</th><th>Row total</th></tr>
<tr><td rowspan="2">Bachelor</td><td>Observed count</td><td>59</td><td>63</td><td>8</td><td>130</td></tr>
<tr><td>Expected count</td><td>71.1</td><td>53.9</td><td>4.9</td><td>130.0</td></tr>
<tr><td rowspan="2">Master</td><td>Observed count</td><td>127</td><td>78</td><td>5</td><td>210</td></tr>
<tr><td>Expected count</td><td>114.9</td><td>87.1</td><td>8.0</td><td>210.0</td></tr>
<tr><td rowspan="2">Column total</td><td>Observed count</td><td>186</td><td>141</td><td>13</td><td>340</td></tr>
<tr><td>Expected count</td><td>186.0</td><td>141.0</td><td>13.0</td><td>340.0</td></tr>
<tr><td>Chi-square</td><td>8.8126</td><td>df</td><td>2</td><td>p-value</td><td><i>0.0122</i></td></tr>
</table>

Table 76 Results for Chi-square test on the abusive conversation dataset (familiarity level of hate speech or abusive conversation and probability distribution preference)

		Distribution 2	**Distribution 1**	**No discernible difference**	**Row total**
Day	Observed count	23	25	2	50
	Expected count	26.9	21.1	1.9	50.0
Month	Observed count	49	60	1	110
	Expected count	59.3	46.4	4.3	110.0
Year	Observed count	96	46	8	150
	Expected count	80.8	63.4	5.8	150.0
Never	Observed count	26	21	3	50
	Expected count	26.9	21.2	1.9	50.0
Column total	Observed count	194	152	14	360
	Expected count	194.0	152.0	14.0	360.0
Chi-square	18.5464	df	6	p-value	*0.0050*

Table 77 Results for Chi-square test on the abusive conversation dataset (ethnicity and probability distribution preference)

		Distribution 2	**Distribution 1**	**No discernible difference**	**Row total**
Asian	Observed count	58	38	4	130
	Expected count	71.1	53.9	4.9	130.0
White	Observed count	78	59	3	210
	Expected count	114.9	87.1	8.0	210.0
Column total	Observed count	186	141	13	340
	Expected count	186.0	141.0	13.0	340.0
Chi-square	0.9913	df	2	p-value	0.6092

Appendix I

Table 78 and Table 79 show the specific details of classifications from sub-models within the ensemble system and instruction tuning. In these two tables, the figures are marked as bold if they represent the primary class where samples from another class are wrongly classified by the model.

Table 78 The identifications of "Mildly abusive", "Strongly abusive" and "Very strongly abusive" from sub-models within instruction tuning (bold: the top frequent class where one class is wrongly classified)

	#Predict #True	Ambiguous	Mildly abusive	Not abusive	Strongly abusive	Very strongly abusive
Sub-model 1	Mildly abusive	0	12	1	**3**	0
	Strongly abusive	0	**5**	1	10	0
	Very strongly abusive	0	0	1	**4**	0
Sub-model 2	Mildly abusive	1	9	**7**	3	0
	Strongly abusive	0	**12**	3	14	1
	Very strongly abusive	0	1	0	**2**	0
Sub-model 3	Mildly abusive	0	2	3	**12**	0
	Strongly abusive	0	**4**	1	15	0
	Very strongly abusive	0	0	0	**1**	0
Sub-model 4	Mildly abusive	0	12	2	3	**4**
	Strongly abusive	0	**6**	1	17	1
	Very strongly abusive	0	**7**	0	4	1
Sub-model 5	Mildly abusive	2	21	**12**	9	0
	Strongly abusive	0	3	0	25	1
	Very strongly abusive	0	0	0	**12**	9
Sub-model 6	Mildly abusive	1	2	6	**12**	0
	Strongly abusive	0	2	2	21	**5**
	Very strongly abusive	0	0	0	**2**	1
Sub-model 7	Mildly abusive	0	0	**6**	3	0
	Strongly abusive	0	3	**4**	7	0
	Very strongly abusive	0	1	1	**2**	0
Sub-model 8	Mildly abusive	2	3	5	**7**	0
	Strongly abusive	2	**5**	3	17	2
	Very strongly abusive	0	0	**1**	5	0

Table 79 The identifications of “Mildly abusive”, “Strongly abusive” and “Very strongly abusive” from sub-models within ensemble system (bold: the top frequent class where one class is wrongly classified)

	#Predict #True	Ambiguous	Mildly abusive	Not abusive	Strongly abusive	Very strongly abusive
Sub-model 1	Mildly abusive	0	5	**8**	3	0
	Strongly abusive	0	**2**	1	12	1
	Very strongly abusive	0	0	1	**4**	0
Sub-model 2	Mildly abusive	5	5	**8**	2	0
	Strongly abusive	**5**	**5**	**5**	15	0
	Very strongly abusive	0	1	2	0	0
Sub-model 3	Mildly abusive	1	7	**5**	4	0
	Strongly abusive	1	**8**	2	9	0
	Very strongly abusive	0	0	0	1	0
Sub-model 4	Mildly abusive	1	5	**10**	4	1
	Strongly abusive	0	2	**6**	17	0
	Very strongly abusive	0	3	**5**	3	1
Sub-model 5	Mildly abusive	4	23	**14**	1	2
	Strongly abusive	4	**5**	2	17	1
	Very strongly abusive	0	2	0	**6**	13
Sub-model 6	Mildly abusive	2	7	**10**	2	0
	Strongly abusive	0	4	2	18	**6**
	Very strongly abusive	0	1	0	**2**	0
Sub-model 7	Mildly abusive	2	0	**6**	1	0
	Strongly abusive	2	0	**4**	8	0
	Very strongly abusive	1	0	**2**	1	0
Sub-model 8	Mildly abusive	3	2	**6**	4	2
	Strongly abusive	3	2	**6**	18	0
	Very strongly abusive	0	**2**	**2**	3	1