Utrecht University
Faculty of Science
Department of Information and Computing Sciences
MSc Artificial intelligence

Uncovering Annotator Dynamics in Hate Speech Detection: A Multi-Model Investigation

A Thesis By

Marco Mikulin
6600018

Project Supervisor Prof. Dr. Massimo Poesio
Second Examiner Dr. Dong Nguyen

Utrecht
University

Abstract

This thesis navigates the complex environment of Natural Language Processing (NLP), focusing on hate speech detection—a task that notably faces various difficulties, such as annotator disagreement [2]. This variance across annotators throughout the labelling process is a typical stumbling block in NLP. Traditionally, annotator discordance is addressed by rejecting conflicting annotations, aggregating a "ground" truth, or utilising softmax for label assignment [46]. On the other hand, a study by Davani and colleagues (2022) [9] argued for the use of multi-annotator models that treat each annotator's judgement as a different subtask while sharing a common representation. An example of this model is presented by Davani and colleagues (2022) [9], their research in addition proposed a technique for estimating prediction uncertainty, which is critical in circumstances requiring careful prediction withholding. Overall, their study is a crucial reference work for the present project, which will draw from it while simultaneously attempting to addresses its weak point, which is resource consumption. Indeed, training on each annotator is not always a viable choice, depending on the volume of the data.

The goal of this research project is to develop a cluster-centric technique based on annotators' metadata or response patterns. This strategy entails training separate models for each identified cluster. The process will involve a careful assessment of the performance of models trained for individual annotators against those optimised for annotator clusters. Evaluation measures, such as cross-entropy, will be used to determine any significant discrepancies in the efficacy of hate speech identification.

Thus, the primary question driving this work is to investigate how the suggested multi-model approach, which takes into account clustered individual annotator dynamics, performs in comparison to previous approaches in hate speech identification. This study aims to enhance NLP approaches to better address real-world problems. In addition, it will examine the time necessary to train each model, offering insights into the relationship between training length and results gained. Overall, the proposed project will contribute to the optimisation of hate speech detection models by providing a comprehensive knowledge of the relationship between annotator dynamics and model performance.

Contents

List of abbreviations and terms

NLP	Natural Language Processing
NN	Neural Network
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformers

1. Introduction

1.1 Problem statement

The task of hate speech detection within the field of Natural Language Processing (NLP) is notably complex, especially due to variance in annotator judgments (Leonardelli et al., 2021 [23]; Waseem, 2016 [48]). This variance translates into significant difficulties for achieving consensus on what constitutes hate speech and what does not. Thus, the goal of training of robust detection models needs to inevitably face issues of variance between annotators. Traditional and widespread practices to deal with annotator disagreement involve discarding conflicting annotations, or consolidating a consensus "ground truth" – for an overview of the intrinsic limitations of these approaches, see Uma and colleagues (2021) [46]. Indeed, one of the most problematic limitations of these methods is that they tend to overlook the nuanced perspectives that different individual annotators offer. On the other hand, recent research endeavors, such as proposed by Davani and colleagues(2022) [9], advocate for multi-annotator models that are able to take into account the diversity of annotators' judgements by treating each annotator's judgment as a distinct subtask within a unified model framework. This novel approach has shown efficacy, particularly in binary classification scenarios, but it does have significant limitations as well. Crucially, the task of training individualized models for each annotator entails a substantial resource consumption. Building on these premises, this research is committed to exploring a cluster-centric approach that would be able to both give justice to the variance in annotator judgement and simultaneously minimize the computational and resource overload that is usually associated with multi-annotator models. The proposed solution is to cluster annotators based on metadata or response patterns, seeking to group annotators with similar socio-economic backgrounds or similar labelling behaviors, and thereby reaching the goal of reducing the number of models required, without significantly compromising the nuanced viewpoints of different annotators. This approach not only should alleviate the resource- intensive nature of multi-annotator models but it also embodies a viable solution for the annotator variance dilemma that is ever-present in hate speech detection tasks. The critical examination of this clustering strategy's impact on model performance and resource efficiency is essentially the core of the problem that this research aims to address.

1.2 Research question

The research question of this project is based on the problem outlined above, and it can be summarized as follows:

RQ: *Is clustering of annotators a good approach to reduce resource consumption and what is its effect on the evaluation score metrics on a multi-annotator model?*

To answer this research question, it is necessary to compare the effectiveness of different clustering methodologies, focusing particularly on whether clustering based on annotator metadata (such as ethnicity, age, or political ideology) yields superior model performance and resource efficiency compared to clustering derived from machine learning algorithms that analyze response patterns. This investigation unfolds in two inter-related steps. First, it seeks to measure the extent to which clustering annotators based on inherent demographic characteristics or observed labelling behaviors can reduce the computational complexity necessary to train separate models for each annotator. Second, it aims to evaluate the impact of these clustering strategies on the accuracy and reliability of hate speech detection, thereby assessing the trade-off between model performance and resource consumption. Overall, it is expected that clustering annotators based on demographic aspects should reduce the computational resources needed to train the model, as compared to the ensemble model reported in Davani and colleagues' (2022) article. For what concerns the evaluation score metrics, one might argue that the clustering approach tested in this thesis could also yield a better performace –as measured by the cross-entropy score (Hui & Belkin, 2020) [20] – due to the fact that it is less subject to the idiosyncratic patterns of individual annotators.

The exploration of these clustering methodologies is pivotal for the advancement of NLP, particularly for what concerns applications that require a nuanced understanding and interpretation of textual data, such as in the case of hate speech detection. By assessing the efficacy of different clustering approaches systematically, this research should produce valuable insights for the optimization of hate speech detection models, thus enhancing their applicability and effectiveness.

2. Literature Review

2.1 Hate Speech Detection

Hate speech refers to the use of linguistic expressions that conveys hate or violence based on race, sex, religion or other sensitive characteristics (Warner & Hirschberg, 2012) [47]. Since the advent of the digital era, with the widespread use of the Internet, and particularly social media, hate speech has always been a strong (and unfortunate) component of the digital world. Nowadays, hate speech detection is a critical and challenging task within the field of Natural Language Processing (Poletto et al. [32], 2021; MacAvaney et al., 2019 [25]), and it is gaining significant attention due to the increasing prevalence of hate speech on digital media and its harmful societal impacts. The purpose of hate speech detection is to identify the presence of text that is offensive, promotes violence, or expresses hate towards individuals or groups based on characteristics such as race, religion, gender, or sexual orientation. The following literature review synthesizes key findings, methodologies, and challenges within this methodology, highlighting the evolution of approaches and the current research directions.

2.1.1 Early Approaches and Challenges

The first attempts to develop automated strategies for the detection of hate speech were proposed in the early 2010s, soon after the launch of social media such as Facebook and Twitter. These initial efforts in hate speech detection primarily relied on keyword-based methods and manually curated blacklists in order to flag potentially harmful content (Warner and Hirschberg, 2012) [47]. Although these approaches were quite straightforward, they evidently suffered from high false positive rates. Crucially, they also lacked the ability to understand context or the subtleties of language, such as sarcasm or idiomatic expressions. The reliance on explicit keywords also made these systems easy to evade through simple lexical substitutions.

2.1.2 Machine Learning and Feature Engineering

The advent of machine learning (ML) algorithms brought more sophisticated techniques, which were adopted for the study of hate speech detection. ML algorithms such as Support

Vector Machines (SVMs), Naive Bayes, and Random Forest classifiers were among the first ones to be employed. Their characteristic is utilizing a range of linguistic features such as n-grams, syntactic patterns, and sentiment analysis (Nobata et al., 2016) [28]. These methods marked a significant improvement over keyword-based approaches (Abro et al., 2020) [1]. However, it is important to note that they were also heavily dependent on the quality and representativeness of the feature sets, requiring extensive manual effort in feature engineering.

2.1.3 Deep Learning and Neural Networks

The introduction of deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), represented a paradigm shift in hate speech detection (Zhang et al., 2018) [50]. These models are capable of learning hierarchical representations of textual data, thus demonstrating superior performance for what concerns capturing semantic and contextual nuances (Pitsilis et al., 2018) [30]. Furthermore, the use of word embeddings, such as Word2Vec and GloVe, additionally enhanced model capabilities by providing dense and meaningful representations of words based on their usage in large corpora.

2.1.4 Transfer Learning and Pre-trained Language Models

The newest development in the field of hate speech detection came from the adoption of transfer learning approaches. Pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformers), and their variants setting new benchmarks in hate speech detection (Devlin et al., 2019 [11], Radford et al. [33], 2018 and Caselli et al., 2020 [8]). These models, which are pre-trained on vast text corpora, are able to capture a deep understanding of language, which can be fine-tuned on smaller hate speech datasets to achieve remarkable accuracy. The ability of these models to understand context, ambiguity, and even implicit expressions of hate speech has significantly advanced the field.

Drawing from these recent advancements in the field, the present project will take inspiration from the work of (Davani et al. 2022) [9]. Indeed, in their study, the main problem affecting the algorithm was the extended period of training for an increased number of annotators. The model uses a pre-trained layer based on BERT and then trains a single classifier for each individual annotator. The mentioned method showed promising results, although it must be noted that only accuracy rate and F score were provided for the experiment. This study will seek to address the gaps in Davani and colleagues' 2022 study, namely the exponential training time in accordance with the number of annotators and more sophisticated performance scores.

2.2 Disagreement

In the realm of annotated data for machine learning, various pieces of the literature show that a single ground truth is often unrealistic (Aroyo et al., 2014) [3] (Geva et. al 2019) [15], but the disagreement among annotators is often regarded as mere noise that should be filtered out to achieve a clean dataset. However, this perspective overlooks the rich insights that such disagreements can offer, particularly in complex tasks like hate speech detection, where subjective interpretation plays a crucial role [22]. Annotator disagreements can reflect genuine ambiguities in the content, diverse perspectives, and the complex nature of language, rather than simple human errors or inconsistencies. The consensus on the most recent literature is to take disagreement and ambiguity into account (Massimo Poesio & Ron Artstein, 2005) [31].

One common method used to resolve annotator disagreement is majority voting, which is based on selecting the label with the most annotator support. The problem with this method is that it may not always yield the most accurate or nuanced understanding of the data. The most critical risk is to oversimplify complex issues and even silencing minority perspectives that may actually be crucial for understanding subtle or context-dependent instances of hate speech. Treating disagreement as an asset rather than an annoying obstacle encourages a deeper exploration of the data and leads to the development of more sophisticated models that can handle the inherent complexity of human language and judgment. As Aroyo and Welty state, “truth is a lie” and disagreement is inevitable as much as it is valuable and representative, since “measuring annotations on the same objects of interpretation [. . .] across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations” (Aroyo & Welty, 2015, p. 21) [4].

Recent examples of research on annotator disagreement, such as the work of Basile and colleagues (2021) [6], challenge the traditional reliance on a single "ground truth" in NLP evaluation. Indeed, they advocate for a reevaluation of current practices in light of the fact that annotator disagreement is impossible to avoid. This perspective is further explored by Leonardelli and colleagues (2023) [24], who explore the potential of learning from disagreements to enhance model training, thus suggesting that understanding the nature and reasons behind annotator variances can lead to more robust models. Similarly, Sandri and colleagues (2023) [39] contribute to this dialogue by proposing a taxonomy to categorize the roots of disagreement, emphasizing that such differences can stem from subjective interpretations, contextual ambiguities, or even the annotators’ backgrounds. These insights collectively argue against the oversimplification of complex linguistic phenomena through majority voting, highlighting the need for models that can navigate and incorporate the variety of human judgment and language interpretation.

The concept of perspectivism (Cabitza et al., 2023) [7], has been proposed to advocate for a paradigm that values and preserves the diversity of opinions in the annotation process. This approach recognizes that annotator disagreements often stem from legitimate variances in perspective and subjectivity, rather than from mere errors. By integrating these multiple perspectives into the "ground truthing" process, perspectivism aims to develop ML models that are more nuanced and representative of the complex nature of human language and cognition. This shift towards acknowledging and leveraging annotator diversity aligns with the solutions proposed by other researchers (Basile et al., 2021 [6]; Leonardelli et al., 2023 [24]; Sandri et al., 2023 [39]). This consensus suggests that there is an acknowledged need in the field of NLP to develop models that can accommodate the inherent subjectivity in tasks like hate speech detection.

Indeed, the state-of-the-art methods for addressing annotator disagreement in classification tasks have become increasingly nuanced. Techniques such as Bayesian models (Fornaciari & Poesio, 2021) [14] take into account the reliability and expertise of each annotator, providing a probabilistic framework to weigh their annotations differently. Additionally, machine learning algorithms that factor in annotator characteristics into the training process, as explored by Rodrigues and Pereira (2018) [36], offer a way to personalize model learning to account for individual biases and interpretation styles, thus enriching the model's understanding and handling of diverse data interpretations. Overall, these approaches signify a shift towards more adaptive and annotator-aware methodologies for dealing with disagreement, emphasizing the complexity and subjectivity inherent in human-labeled data.

For what concerns the evaluation metrics used to assess the performance of the types of models used to study annotator disagreement, the consensus found in recent literature indicates the need for a soft-score approach (Uma et al., 2021)[45] and (Rizzi et al., 2023) [35]. Therefore, the model evaluated in this thesis will be assessed using the cross-entropy measure, which calculates the difference between the calculated probability distribution of a classification model and the values that it predicts.

2.2.1 Challenges and Ethical Considerations

Despite the reviewed technological advancements, the path towards achieving accurate hate speech detection methods yet remains filled with challenges. The ambiguity of language, the evolving nature of forms of hate speech itself, and contextual dependence make automated detection difficult. Furthermore, issues of bias and fairness in models are another crucial point to keep in mind, and they are due to imbalanced or non-representative training data, which pose significant ethical concerns (Dixon et al., 2018) [12]. The risk of over-policing certain dialects, languages, or demographic groups highlights the need for careful consideration of ethical implications in model development and deployment.

2.3 Clustering

A multi-task classification on hate speech can be a potentially very time-consuming task that requires a lot of resources, particularly if - for the sake of a high metrics score - each annotator is considered as a single task and a classification layer is trained for each of them. A way to overcome this issue by not classifying each annotator would be to create different clusters of annotators instead, considering clusters as representative for each group of annotators [9]. With this approach, it should be possible or increase to preserve a high score metric without using a heavy computational power, thus reducing time and resources for the classification.

2.3.1 Introduction

Clustering methodologies play a pivotal role in annotator behavior analysis, especially in tasks requiring subjective judgment, such as hate speech detection. These methodologies aim to identify patterns within annotator responses, thus yielding a more nuanced understanding of variability and consensus among annotators. The literature review reported below describes two primary clustering approaches: metadata-based clustering, which focuses on demographic and background information, and algorithm-based clustering, which focuses on annotator choices and response patterns.

2.3.2 Metadata-Based Clustering

Metadata-based clustering involves grouping annotators based on their demographic characteristics such as age, ethnicity, gender, educational background, or cultural context. This approach is based on the premise that annotators' backgrounds influence their perception and interpretation of content and language [18], and that therefore it could potentially lead to systematic differences in their annotations (Hovy and Søgaard, 2015) [16]. For instance, (Davidson et al., 2019) [10] spotted a clear difference in annotation based on race, while (Waseem & Hovy 2016) [17] demonstrated how annotators' cultural and social backgrounds could affect their judgments on what constitutes hate speech, suggesting that metadata-based clustering could help in understanding these biases and improving the generalizability of NLP models by incorporating diverse perspectives. Furthermore, Sap and colleagues' research (2019) [41] state that annotations might be racially biased, basing their claim on the finding that phrases in the African American English dialect (AAE) tend to be identified as more toxic than their general American English equivalent when toxicity detection tools are used, compared to when they are evaluated by African American English speakers themselves (Sap et al., 2019; Spears 1988) [41] [44].

The work of Fleisig and colleagues (2023) [13] shows that is also possible to predict individual annotator ratings based on their demographic and social background and in doing so increase the performance score of the classifier. Additionally, the study of Sap and colleagues (2021) [40] highlights the link between the annotator's choice and their social background, contextualizing toxicity labels with social variables.

Despite its potential, metadata-based clustering inevitably faces challenges related to privacy concerns, the availability and reliability of metadata, and the risk of reinforcing stereotypes or introducing bias by assuming homogeneous behaviour within demographic groups (Sen and Wasow, 2016) [42]. This possible bias reinforcement is cause for ethical concern. Moreover, the complexity of human perception and the influence of implicit biases imply that demographic metadata may not always provide a complete picture of annotator behaviour patterns.

This last point is reinforced by the study of Orlikowski and colleagues (2023) [29], which showed that individual annotation behaviour depends on much more than just sociodemographics.

2.3.3 Voting-Behaviour-Based Clustering

Algorithm-based clustering, on the other hand, focuses on the analysis of annotator choices and labelling patterns, and this analysis is independent of their demographic characteristics. This approach employs machine learning algorithms to identify clusters of annotators who exhibit similar behaviours in their responses, potentially reflecting shared understanding, biases, or interpretation strategies. Techniques such as K-means clustering, hierarchical clustering, and model-based clustering have been applied to capture the latent structure in annotator responses (Raykar et al., 2010) [34].

A significant advantage of algorithm-based clustering is its ability to dynamically adapt to the data, uncovering patterns that may not be apparent from metadata alone. For instance, Raykar and colleagues (2010) [34] investigated a model that identifies reliable annotators and estimates true labels from noisy annotations, effectively revealing underlying patterns of agreement among annotators. This approach highlights the diversity in interpretation and the potential for clustering annotators based on their labeling consistency, even within groups that might seem homogeneous from a demographic perspective. Insights from such methodologies are invaluable for developing more robust and adaptive NLP models, as they facilitate the integration of annotations based on nuanced agreement patterns rather than relying exclusively on static demographic categories.

2.3.4 Comparative Analysis and Challenges

Comparing metadata-based and voting behavior-based clustering reveals advantages and disad- vantages between the two approaches. Metadata-based clustering offers an intuitive way to account for diversity and potential biases but may oversimplify the complex influences on annotator behavior. In contrast, voting behavior-based clustering provides a data-driven perspective that captures dynamic patterns in annotator responses but may lack interpretability and the ability to directly account for known sources of bias. Overall, both approaches face challenges related to the representativeness of the data, the risk of introducing or overlooking biases, and the ethical implications of clustering decisions. Furthermore, the effectiveness of each clustering method can vary significantly depending on the task, the nature of the data, and the specific objectives of the analysis. In this study, both approaches will be used and a comparison between their effectiveness will be carried out, since both cluster-resulting datasets will be trained into the model. From a theoretical point of view, a voting-behavior approach might seem preferable since it should better capture and understand similarities on the data, However, it is important to note that the performance of this method highly depends on the amount, quality and sparsity of the data. Indeed, it requires a substantial number of both annotations per annotator and annotations per comment. As described in the methodology section below, the dataset used in this study does present some challenges in this respect, it is plausible that the metadata clustering approach might prove to be more efficient, as it does not rely on specific conditions, but only on the existence of metadata information.

In summary, clustering techniques in annotator behavior analysis offer powerful tools for understanding the variability and consensus among annotators in tasks such as hate speech detection. While metadata-based clustering provides a straightforward means to incorporate diversity and background influences, voting behavior-based clustering offers a flexible, data-driven approach to uncovering latent patterns in annotator responses. The choice between these approaches — or a hybrid strategy that combines elements of both — should be informed by the specific research questions, the available data, and the ethical considerations inherent in the analysis of annotator behaviour. Future research goals in this area involve further refining these methodologies, enhancing their applicability and efficacy in the development of NLP models that are both robust and sensitive to the nuances of human judgment.

3. Data

3.1 Dataset

This project will employ a dataset presented in Kennedy and colleagues (2020) [21] and Sachdeva and colleagues (2022) [38]. The Dataset was downloaded from HuggingFace, a notorious repository for dataset and models [19]. It constitutes an expansive corpus of annotated social media comments, designed to facilitate a detailed examination of hate speech dynamics online. Comprising 39,565 comments, each subjected to annotation by one of 7,912 distinct annotators, the dataset has a total of 135,556 data points sourced from YouTube, Twitter, and Reddit and labelled by US-based Amazon Mechanical Turk workers. Central to this study is the "hatespeech" column in particular, a categorical variable that quantifies the existence of hate speech within a comment, but the dataset also includes different other additional parameters, for example: sentiment, respect, insult, humiliate, violence, genocide, dehumanize and status. As explained by Kennedy and colleagues (2020) each annotator received a random sample of comments to annotate and there was no specified aim to annotate certain comments (such as particularly ambiguous ones) more than others.
Each entry in the dataset is uniquely identified by a comment ID and has been evaluated by an annotator, who in turn has been assigned a unique ID. The dataset also provides a rich array of metadata for each comment, including the text of the social media post. Additionally, detailed demographic information for each annotator is provided, including variables such as gender, education level, income bracket, and political ideology. This demographic data opens avenues for analyzing the influence of annotator backgrounds on hate speech perception and annotation, which is critical for understanding potential biases in hate speech detection. In the context of hate speech detection, the utility of annotator metadata for clustering purposes is a key component, which motivated the choice to focus on this specific dataset. Indeed, using these data allows to identify and distinguish patterns regarding how different groups perceive and categorize hate speech, thereby shedding light on subjective biases in annotation. This insight is particularly valuable for enhancing the accuracy and fairness of automated hate speech detection systems. Some examples of annotator's metadata are the following columns visualized in Table 1

It is important to acknowledge that while the dataset contains extensive metadata regarding

Table 1. Annotator Demographic Breakdown

Category	**Values**
Education	Some High School, High School Grad, Some College, College Grad (AA), College Grad (BA), Professional Degree, Masters, PhD
Gender	Men, Women, Non Binary, Prefer Not to Say, Self Describe
Income	<10k, 10k-50k, 50k-100k, 100k-200k, >200k
Ideology	Extremely Conservative, Conservative, Slightly Conservative, Neutral, Slightly Liberal, Liberal, Extremely Liberal, No Opinion
Race	Asian, Black, Latinx, Middle Eastern, Native American, Pacific Islander, White, Other
Religion	Atheist, Buddhist, Christian, Hindu, Jewish, Mormon, Muslim, Nothing, Other
Sexuality	Bisexual, Gay, Straight, Other

the annotator, it also includes various metadata columns about the targets of the comments subjects to hate speech detection. This study will include only the annotator's metadata, leaving the target's information for further analysis.

In essence, this dataset provides a comprehensive framework for the detailed analysis of hate speech on social media, offering broad quantitative measures that lend themselves for a variety of applications. By leveraging annotator demographics for clustering, it is possible to discover significant patterns in hate speech annotation, contributing to the ongoing refinement of detection methodologies and the broader discourse on online hate speech.

3.2 Dataset: preliminary observations

The first notable obaservation is that the dataset presents high levels of variance, which can be attributed to its crowdsourced nature. This characteristic often leads to a wide range of annotations per comment and per annotator due to the diverse backgrounds and perspectives of the contributors. The provided tables illustrate these core numbers: Table 2 details the distribution of annotations per comment, highlighting variations in engagement levels across different content. Table 3 describes the numbers of contributions per annotator, reflecting the varying degrees of participation. The observed concentration of comments with a lower number of annotators, predominantly within the 1 to 5 range, juxtaposed against the rarer instances of comments with a higher annotator count, indicates a skewed annotative effort. This skewness suggests that the majority of comments are subjected to a relatively limited diversity of perspectives, which could impact the robustness and generalizability of hate speech detection models trained on this dataset.

Annotations Count	**Number of Comments**
1 or 2	22,213
3 to 5	17,279
More than 5	73

Table 2. count of annotations per comment

Annotations Counts	**Number of Annotators**
Less than 5	0
5 to 10	251
10 to 15	1,706
15 to 25	5,869
More than 25	86

Table 3. count of annotations per annotator

From an NLP perspective, the variance in annotator engagement could introduce biases or inconsistencies in the labeling process, thus potentially affecting the model's ability to accurately identify and classify hate speech across a broad range of instances. The presence of comments with a significantly higher number of annotators might indicate ambiguous cases, which, while valuable for understanding complex hate speech nuances, may also introduce variability in the training data that could challenge the model's learning process.

4. Methodologies

4.1 Data pre-processing

The first step of the dataset preparation process involved a strategic filtration criterion, namely to exclude all comments with fewer than three annotators and all comments annotated by annotators with less than 15 annotations. This threshold was established to ensure a minimum level of comparability among annotators, thus enhancing the quality of the data used for training hate speech detection models. However, this filtering process presents its own inherent complexities which require consideration. Indeed, the reciprocal dependency between tweets and annotators is a critical aspect that complicates the refinement of the dataset. For example, the removal of an annotator with a relatively low number of annotations may inadvertently decrease the annotation count of certain tweets that initially surpassed the annotation threshold. Similarly, maintaining tweets that marginally meet the annotation threshold could disproportionately favor annotators with minimal contributions, thus impacting the overall dataset quality.

This is a recursive dilemma that lacks a flawless resolution due to the intertwined nature of the annotations between tweets and annotators. To address it, an additional filtering methodology was applied, in the form of a *min-max* optimization algorithm. The algorithm's objective was to maximize the annotations per tweet and per annotator simultaneously, ensuring a more balanced representation and utilization of both annotators and comments within the dataset. This approach allows for a strategic allocation of resources that aims to enhance the integrity of the annotation process, while acknowledging the constraints imposed by the underlying dataset structure. The rationale behind this decision is that a greater number of annotators per instance may lead to more accurate and dependable annotations, as it mitigates the impact of individual biases and errors. Following the filtering phase, the shape of the resulting dataset is the following: from 7912 to 1603 unique annotators and from 39565 to 1020 unique comments. The counts of annotations per tweet and per annotator for the filtered dataset are reported in table 4 and table 5

Annotations Count	Number of Comments
1 or 2	0
3 to 5	950
More than 5	70

Table 4. count of annotations per comment

Annotations Counts	Number of Annotators
Less than 5	6
5 to 10	1242
10 to 15	250
15 to 25	105
More than 25	0

Table 5. count of annotations per annotator

Three essential preprocessing steps remain, which are pivotal to preparing the data for subsequent anal- ysis via a machine learning model. The necessity for these steps arises from the structural discrepancy between the existing format of the raw dataset and the format required by the learning model. The filtering process had also an important role on the limitation of the variance from the annotators, since picking only the "high quality" data reduced the variance present on the number of annotations per annotator and per comment. Table 6 summarizes the datasets before and after the filtering process.

	Comments	Annotator
Original dataset	39565	7912
Filtered dataset	1020	1603

Table 6. Dataset Annotations Overview

Step 1: Pivoting the Dataset: The original format of the dataset presents a challenge for data analysis and model training as each line represents a unique annotation, including different fields such as the tweet ID, annotator ID, the specific annotation, and detailed annotator metadata. This structure is inherently complex for models that are designed to process each tweet as a single observational unit with features derived from multiple annotations. To rectify this, a pivoting transformation was necessary. By reorganizing the data, where each row now corresponds to a unique tweet and each column to a different annotator's annotations, we facilitate a more structured and model-friendly dataset. This transformation aligns with the analytical needs of typical learning algorithms (which expect a row-wise representation of data points), and it also enhances the efficiency of data handling and model computation, since it reduces the complexity of the data structure.

Step 2: Creating a Support Database: Given the richness of the metadata associated with each annotation, it is important to retain this information for a comprehensive analysis. However, incorporating such metadata directly into the pivoted dataset could lead to redundancy and an unwieldy number of columns, potentially diminishing the model's performance due to high dimensionality. To address this, a secondary support database was developed. This database, keyed on a composite primary key of tweet ID and annotator ID, meticulously preserves the relationship between tweets, annotators, and their respective metadata. This design allows for an efficient connection back to the original detailed data without overbur- dening the primary analysis dataset. It enables detailed queries to be performed without compromising the structure of the training dataset.

Step 3: Metadata Integration and Feature Engineering: this last apreprocessing step in- volves the integration of selected metadata elements back into the primary dataset as new features. This step is guided by the initial exploratory data analysis, which identifies metadata attributes that could potentially enhance model predictions, such as annotator expertise or the timestamp of annotations. By systematically incorporating these features, the model's input space is enriched, potentially improving its predictive accuracy and providing deeper insights into the factors influencing the sentiment or content of comments.

4.2 Clustering

In the development of machine learning models reliant on annotated data, understanding the variability and biases among annotators is critical. Research has shown that annotators' decisions can be influenced by a myriad of personal factors, including demographic backgrounds and socio-political beliefs (Hovy & Spruit, 2016) [17]. In this section, we explore the application of clustering techniques to segment annotators based on their metadata and voting behavior, aiming to optimize model training and enhance the reliability of annotations.

4.2.1 Metadata Based Clustering

The rationale behind using metadata for clustering is based on the premise that an annotator's background can significantly influence their annotations. By identi- fying and understanding these clusters, we can tailor model training processes, potentially compensating for these inherent biases.

Data Preparation

To understand the influence of annotator characteristics on annotation behavior, we first compiled a comprehensive dataset capturing a wide array of metadata, extracted from the original filtered dataset, with each row uniquely identifying an annotator and each column representing associated metadata. This dataset includes 56 columns that capture demographic information, educational background, political ideology, and other personal attributes, structured as follows:

- Demographic and Socio-economic Features: Includes gender, income levels, and educational attainment. These features are crucial as they often influence an individual's perspective and, potentially, their annotations.
- Ideological Orientation: Captured to assess if political or ethical beliefs correlate with annotation tendencies.
- Ethnic and Religious Backgrounds: Included to explore the diversity of cultural contexts among annotators.
- Sexual Preferences

Each annotator is represented uniquely within the dataset, with duplicated entries removed to maintain the integrity of the clustering analysis.

Methodology

Following dataset construction, a distance matrix was generated using both Euclidean and cosine distance metrics to assess the dissimilarity among annotators. For clustering, we employed the Ward method for a hierarchical clustering [26], which is known to reduce variance within clusters (Murtagh, Fionn & Legendre, 2014) [27], thereby ensuring a more uniform distribution of items per cluster.

Extensive testing was conducted to identify the optimal number of clusters. This involved calculating the Silhouette score [43] and Davies-Bouldin Index [49] for potential cluster sizes ranging from 5 to 25. Based on these evaluations, clusters of sizes 10, 15, and 25 were selected for further model training.

Three distinct datasets corresponding to these cluster sizes were subsequently created using hierarchical clustering. The accompanying visualization in figure 1 illustrates the structure of the 25 clusters derived using the Euclidean distance for metadata-based clusters, showcasing the detailed grouping and separations achieved through this analytical approach.

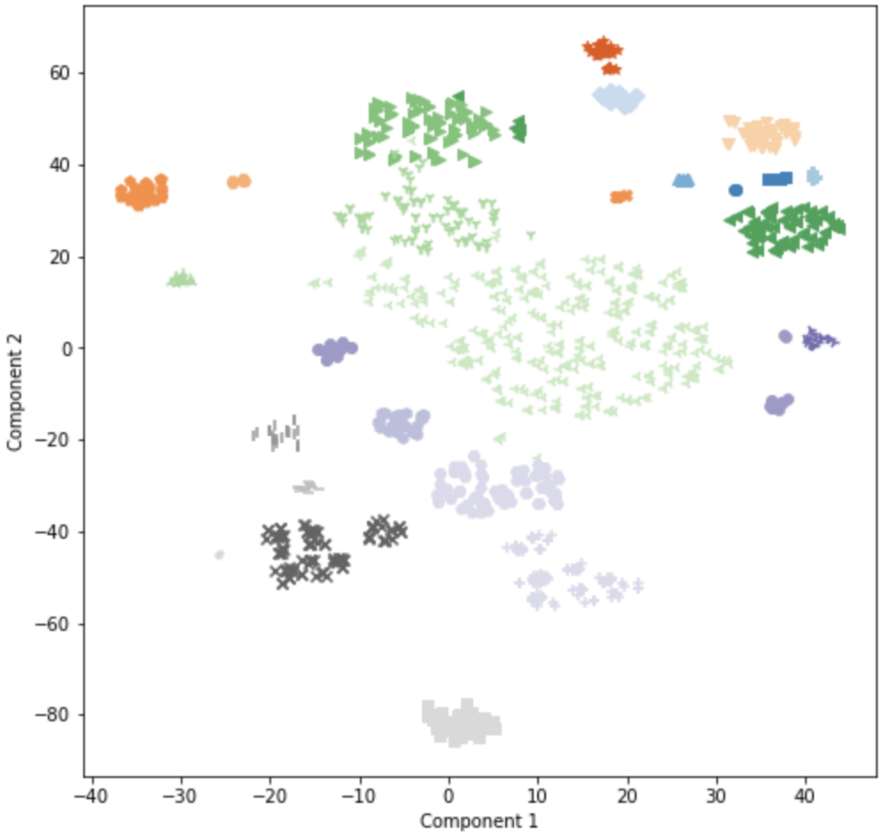

Figure 1. t-SNE Visualization of 25 clusters for metadata-based clustering. Distance: cosine distance

After the clustering process was completed, the votes within each cluster were aggregated using a majority voting system. Although we acknowledge that majority voting may not be the most effective method for examining these phenomena, we believe that this aggregation approach reflects a sufficient commonality based on the socio-economic backgrounds of the annotators. In other words, this method aims to capture the predominant trends and opinions within each distinct group, providing a generalized perspective based on collective attributes.

In addition, some analysis have been carried out for the metadata clustering. The hierarchical clustering of annotators revealed distinct profiles based on various demographic, ideolog- ical, and educational characteristics. Each cluster was defined by the top five features that demonstrated the highest variance and provided the most significant insight into the clustering logic. The analysis used a range of binary and categorical features, with emphasis placed on ensuring each cluster was uniquely characterized by its top features. To distinguish the unique characteristics that define each cluster within a complex dataset, such as those involving human annotators, it is effective to employ a feature importance analysis using Random Forest classifiers. This approach entails transforming the cluster analysis into a set of binary classification problems, where each cluster is contrasted against the others. By applying Random Forest, which is a method that is well-known for its ability to handle high-dimensional data and to provide interpretable results regarding the importance of features, we can accurately identify which features most significantly influence the assignment of data points to specific clusters. The model quantifies the importance of each feature based on its impact on model accuracy, specifically through the reduction in node impurity across the decision trees that make up the forest. The results are shown in table 9. This methodology ensures that the derived importance of features directly correlates to their capability to define the cluster, offering detailed insights into the underlying patterns and behaviors that differentiate data groups.

4.2.2 Clustering Based on Voting Behavior

Clustering based on voting behavior allows for an examination of annotator reliability and the identification of outlier behaviors, which are crucial for ensuring the quality of training datasets (Artstein and Poesio, 2008) [5]. This approach is functional to uncovering potential biases and areas of ambiguity among annotators, directly impacting model accuracy.

Data preparation

In parallel to the metadata-based approach, we also clustered annotators based on their actual annotation behavior — specifically, their voting on a binary scale across various texts. This approach directly quantifies the agreement and consistency among annotators, offering a dynamic insight into their decision-making patterns.

Methodology

Again using hierarchical clustering, we adjusted our approach to focus on the patterns of agreement and disagreement among annotators. To analyze these patterns, we constructed another dataset where each row represented a unique annotator and each column a unique comment. We explored various distance metrics—such as cosine similarity, Hamming distance, and Euclidean distance—to construct the distance matrix. Different linkage methods were tested, including the "average" and "Ward" methods. The Ward method proved superior in terms of the distribution of elements within the clusters.

Although the clusters were well-balanced and the silhouette scores were comparable to those calculated using the metadata-based method, the clustering results were ultimately unconvincing. For visualization using the t-SNE method, we needed to address the presence of NaN values in the dataset, which represent comments not annotated by some annotators. These NaN values were substituted with a placeholder value of -1, as the t-SNE algorithm does not accommodate null values. However, the t-SNE plots did not reveal a clear partition among the clusters as observed in the metadata case. This discrepancy may stem from several factors: notably, the high dimensionality of the data, with more than 1000 dimensions compared to 50 in the metadata case, which potentially overwhelmed the dimensionality reduction capabilities of the t-SNE algorithm. Additionally, the placeholder substitution for missing annotations may have distorted the true distances among annotators, further complicating the clustering outcome.

The results of this behavioral clustering procedure are reported in Figure 2

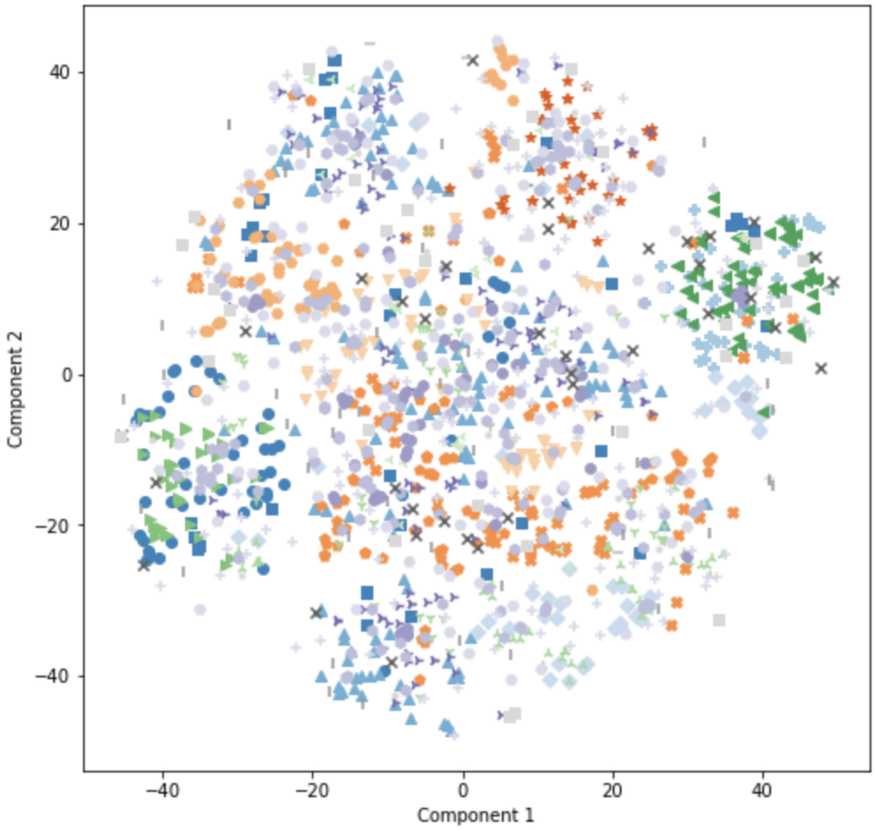

Figure 2. t-SNE Visualization of 25 clusters for behavioural based clustering. Distance: cosine distance

As we did with the metadata-based clusters reported above, we applied the majority voting system to the behavioural clusters as well, meaning that we aggregated the votes within each cluster into a single array. However, it is important to note that the boundaries of the behavioral clusters are less defined and show more overlapping tendencies compared to the metadata-based clusters. Therefore, the vote aggregation procedure inevitably yields a weaker degree of accuracy and representative potential in this case.

4.3 Comparison

Numerous tests have been carried out to determine the number of clusters to adopt, especially on the metadata-based cluster that was giving more promising results. A first attempt has been carried out calculating the silhouette score for a range 2-500. Silhouette score gave the best value to 292 clusters, a first attempt has been carried out but the model performed poorly, in the author's opinion that is because the data was still too sparse, we would see the same behavior later in the results for the model trained on the whole dataset. A more conservative second attempt has been carried out for a range between 5 and 27 included, both the Silhouette Scores and the Davies-Bouldin Index agreed on 25 as the optimal number of clusters.

Table 7 reports the scores for the range 5-27.

Table 7. Cluster Evaluation Metrics

Rank	Silhouette Score	Davies-Bouldin Index
1: 25 clusters	0.1517	1.5888
2: 26 clusters	–	1.5994
2: 22 clusters	0.1505	–
3: 21 clusters	0.1442	1.6300

Upon applying these methodologies, we determined that 25 clusters provided the optimal balance of within-cluster homogeneity and between-cluster heterogeneity, as indicated by silhouette scores (Rousseeuw, 1987) [37] calculated on a range from 7 to 30.

The comparison between metadata-based and voting-based clustering revealed that metadata clustering grouped annotators more clearly and without overlapping, while the voting behavior approach lacked consistency. Both clustering methods could provide valuable insights but they relate to different aspects of annotator analysis. By incorporating both approaches, when possible, we can achieve a more comprehensive understanding of annotator behavior, which is essential for developing robust and accurate machine learning models.

4.4 The Model

In the realm of text classification, modeling the variance among annotators presents a complex challenge. Traditional approaches often aggregate annotations to form a single label per instance, potentially obscuring valuable insights into the diversity of human judgment. The proposed BERT-based ensemble model addresses this issue by implementing a multi-task learning architecture that treats each annotator's decision as a separate but related task.

4.4.1 Model Architecture

The architecture in Davani and colleagues' model (2022), which this thesis draws from and builds on, utilizes BERT, a pre-trained transformer model, known for its effectiveness in generating contextual embeddings (Devlin et al., 2018 BERT) [11]. The model extends the BERT architecture by adding multiple output layers (one for each annotator) following the shared transformer layers.

The model employs the BERT base model to extract deep contextual representations from input texts. This choice is motivated by BERT's ability to understand subtle nuances in language due to its bidirectional training, which captures the context of each word from both left and right sides of the text. For each annotator, a dedicated linear layer is attached to the top of the BERT model. These layers are trained to predict the labeling behavior of individual annotators based on the embeddings provided by BERT.
The linear nature of these layers allows for maintaining computational efficiency while being flexible enough to model the complex decision boundaries that might characterize each annotator's judgments. A significant innovation in the training process of Davani and colleagues' model is the dynamic computation of class weights. This strategy addresses the issue of class imbalance that is often seen in datasets involving subjective judgments, such as hate speech, where one class may significantly outnumber others. By adjusting the weights during training, the model maintains sensitivity to less frequent but crucial classes, enhancing its detection capabilities.

4.4.2 Advantages of the Multi-Annotator Approach

Unlike methods that treat individual differences by aggregating annotations, this model preserves the granularity of each annotator's judgments. By sharing the lower layers of BERT across all tasks (annotators), the model reduces the number of parameters needed compared to completely separate models for each annotator. This shared learning mechanism improves learning efficiency, and it but also helps in regularizing the model, reducing the risk of overfitting individual annotator patterns. The model is inherently adaptable to different numbers of annotators and can handle varying levels of annotation complexity, from binary to multi-class scenarios. Furthermore, the model is trained using a task-specific cross-entropy loss function, which is well-suited for classification tasks. Each annotator's model component computes its loss independently, allowing the process to assess the model weights in a way that best predicts each annotator's labeling patterns. This choice of loss function reflects its effectiveness in handling classification problems where the output can be treated as a probability distribution over classes. The loss penalizes the divergence between the predicted probability distribution and the actual distribution, effectively guiding the model towards better accuracy.

In addition, by optimizing each annotator's output layer independently, the model can adjust to the specificities of each annotator's decision-making process without compromising the integrity of the shared features learned from the BERT layers. For what concerns evaluation metrics, the model's performance is assessed not just on traditional metrics like accuracy, precision, recall and F1-score, but also on measures that reflect its ability to model annotator disagreement, such as inter-annotator agreement indices. This twofold focus ensures that the model is robust not only in terms of predicting general trends, but also in capturing the variability among annotators. In addition to Davani and colleagues' model, for this study we have implemented additional score metrics, namely cross-entropy. More details on **Training and implementation** section.

4.4.3 Challenges and Limitations

Among other challenges, the major ones that the model faces are:

Scalability: : the model's scalability could be challenged as the number of annotators increases, due to the linear increase in the number of output layers and associated parameters. (i.e. the more annotator the more time and complexity)

Complexity in Aggregation: determining the best way to aggregate individual annotator predictions into a final decision remains important, requiring further research into effective ensemble methods. (i.e. as the output for all models is aggregated using majority vote. Some other method could be applied)

Dependency on Annotator Representation: the effectiveness of the model assumes a relatively balanced representation of each annotator in the training data, which might not always be available. (i.e. the model suffer from the sparsity of data, especially for annotations per annotator).

Notwithstanding these limitations, the BERT-based ensemble model for annotator-specific classification still represents a significant advancement in handling subjectivity in annotated datasets. By respecting individual annotator biases while leveraging the power of shared learning, the model enhances the accuracy of predictions, while also providing deeper insights into the annotation process itself.

In this study, we will try to address the *Scalability* and the *Dependency on Annotator Representation* limitations, with the addition of the clustering techniques, while the Complexity in Aggregation is suggested as the basis for future research endeavors. This thesis adopts the assumption that clustering might address both limitations taken into account. The Scalability issue is resolved by reducing the number of annotators, since the clusters will become the new annotators. This process will also address the representation of the annotator in the training data since the data will result more "dense" after the clusterization process.

4.4.4 Training and implementation

The architecture of the BERT-based ensemble model used in this thesis has been refined to include the calculation of cross-entropy as a performance metric. Specific adjustments were made to the loss function and the forward function to facilitate the computation of logits in a format suitable for cross-entropy evaluation. This adjustment allows for a more refined assessment of model performance by directly comparing the predicted probability distributions with the true labels.

Additionally, modifications were implemented in the results function to not only compute traditional metrics such as Accuracy, Recall, Precision, and F1-Score but also to integrate the calculation of the cross-entropy loss. This comprehensive metric assessment enables a more detailed evaluation of the model's classification effectiveness.

The model employed a set of selected training parameters optimized for performance:

- **Batch Size:** Set to 4 to balance memory usage and gradient estimation.
- **Learning Rate:** A low rate of 1×10^{-7} to facilitate gradual convergence.
- **Maximum Token Length:** Fixed at 128 to manage context retention and computational demands.
- **Number of Epochs:** Limited to 10 to prevent overfitting while allowing adequate learning.
- **Random State:** Set at 9999 for reproducibility.
- **Number of Folds:** Employed 5 folds in cross-validation to evaluate model stability.
- **Ensemble Mode:** Enabled to utilize the strengths of multiple model instances.
- **Stratification:** Applied in data splitting to maintain class proportionality.

- **Prediction Mode:** Focused on label prediction to emphasize classification accuracy.
- **Monte Carlo Passes:** Conducted 10 passes to enhance the robustness of uncertainty estimates.

These parameters were dynamically adjustable to respond to different training scenarios, ensuring flexibility and adaptability of the training process to various data characteristics. This configuration supports the enhanced performance metrics but aligns with the model's goal of providing reliable and nuanced insights into annotator-specific classifications.

The model was trained on three distinct datasets to evaluate its performance across various scenarios:

Complete Dataset: The first training scenario utilized the entire dataset without any application of clustering algorithms. This approach provides a baseline for performance assessment, representing the model's capability to handle diverse annotator perspectives without prior segmentation.
Metadata-Based Clustering: The second dataset was derived from applying hierarchical clustering based on metadata to group annotators into 25 clusters. This dataset helps assess how well the model performs when annotators are segmented according to their demographic and background characteristics.
Voting Behavior-Based Clustering: The third dataset resulted from clustering annotators into 9 groups based on their voting behaviors using hierarchical clustering. This method aims to explore the model's efficiency in contexts where annotator judgments are aligned more by decision patterns than by personal metadata.

The results section below will detail the outcomes of the analyses conducted on these datasets. Each dataset's results sheds light on the model's adaptability and accuracy in varying clustering contexts.

5. Results

This section presents the performance results of the model trained on three distinct datasets: the full dataset, the metadata-based clustered dataset, and the voting behavior-based clustered dataset. The evaluation metrics used to assess the model's performance include Accuracy, Precision, Recall, F1-Score, and cross-entropy. These metrics provide a comprehensive view of the model's effectiveness across different data segmentation strategies.

Table 8 below summarizes the results for various datasets on ensemble mode:

Dataset	**Accuracy**	**Precision**	**Recall**	**F1**	**Cross Entropy**
Full Dataset Single	0.7799	0.6973	0.7463	0.721	0.4233
Full Dataset Ensemble	0.5706	0.4204	0.4771	0.447	0.6804
Metadata Cluster	**0.8251**	0.7039	**0.7816**	**0.7407**	**0.3797**
Behavior Cluster	0.8158	**0.7156**	0.7621	0.7382	0.4081

Table 8. Model performance across different datasets

All models were trained using a V100 GPU within a high RAM cloud computing environment, reflecting robust computational resources aimed at efficiently handling large datasets.

The results for different number and type of clusterization can be visualized on figure 3 where lower score means better performance.

The low performance of the whole dataset training with Ensemble mode is probably caused by the fact that there are not enough annotation per annotator, a situation that is fairly common in crowd-sourced dataset, for the model to generalize the pattern, thus resulting in a poor performance. More interesting is the fact that the model trained with the Ensemble mode, but clustered performed better in both cases (metadata and behaviour) than the baseline model that captures the most common approach: a single-task classifier trained to predict the aggregated label for each instance (i.e. majority vote).

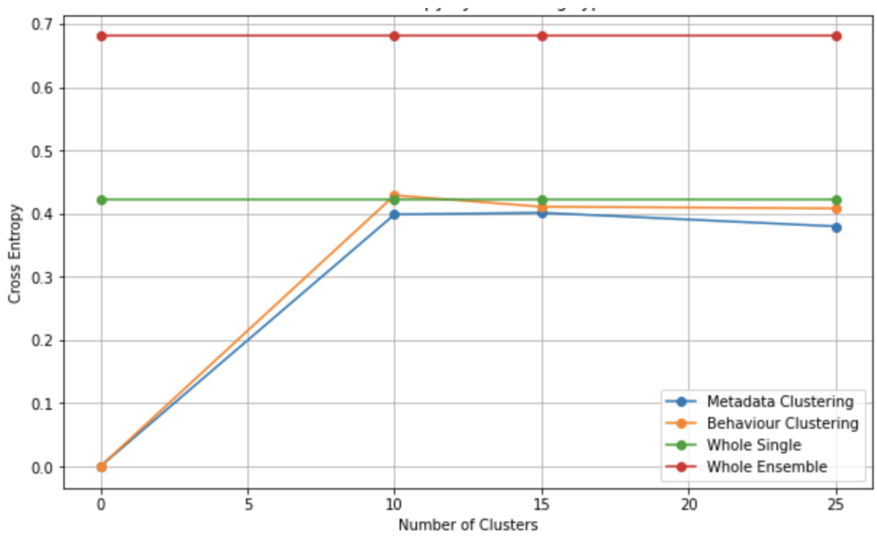

Figure 3. Cross Entropy metric for different clusters

5.0.1 Training Time Breakdown

Full Dataset: For the comprehensive dataset comprising 1,603 annotators and 1,000 unique comments, the training duration for each annotator averaged approximately 1:30 minutes. Consequently, the total training time for the full dataset was calculated to be 2,404.5 minutes. Clustered Datasets: In contrast, each clustered dataset, which included 25 clusters of annotators, exhibited a longer average training time of about 2:30 minutes per cluster. This increase to 2.5 minutes per cluster is attributed to the less sparse data in these segments. The total training time for all clustered datasets thus amounted to 62.5 minutes.

The increased training time for clustered datasets is primarily due to the higher data density within each cluster, which requires more comprehensive computational processing. Comparing annotations between average individual annotators and clustered groups revealed significant discrepancies, thus highlighting the impact of data sparsity on training dynamics.

Overall, these results provide critical insights into the scalability and efficiency of training annotator-specific models, highlighting the trade-off between training duration and data comprehensiveness.

6. Conclusions

The exploration of clustering methodologies to enhance resource efficiency and improve model performance in hate speech detection has provided interesting findings. The central research question addressed whether clustering annotators could be an effective strategy to reduce resource consumption while maintaining or enhancing evaluation score metrics.

The results indeed indicate a notable increase in performance when annotators are clustered. This improvement is likely attributed to the densification of the data within clusters, which seemingly facilitates model generalization. Notably, both metadata-based and voting behavior- based clustering strategies outperformed the baseline model, which relies on a single-task classifier trained to predict aggregated labels for each instance (i.e., majority vote).
This suggests that clustering, by consolidating similar annotator characteristics or response patterns, enhances the contextual density of the training data, thus enabling more effective learning and prediction by the models. However, it is important to note that, when compared individually, the metadata-based clustering method has proven to be a more effective method than the voting-behavior clustering method. This outcome is aligned with the hypotheses presented in the theoretical framework, where it was noted that the success of a voting-behavior clustering strategy is highly dependent on the characteristics of the dataset used in a specific study.
Overall, these indings seem to support that annotator alustering in the context of a multi-annotator model, represents a promising approach to reducing computational time, since it effectively addresses the variability in annotator responses. By grouping annotators with similar traits or behaviors, models can potentially consider shared informational contexts that would otherwise be diluted in a more heterogeneous dataset. However, this approach does not come without trade-offs. The aggregation of annotators, whether by metadata or voting behavior, still inevitably leads to a loss of individual annotator perspectives, at least to a certain degree.

6.0.1 Limitations and Future Research

Several limitations affect the current study. The effectiveness of individualized model training is dependent on having a densely annotated dataset, however obtaining the ideal dataset was not feasible, as it often is in crowd-sourced environments due to the sparse nature of annotations per annotator. Furthermore, the reliance on metadata for clustering necessitates comprehensive demographic data, which is not commonly available in many crowd-sourced datasets.

Future research should build and subsequently analyze datasets where each annotator contributes more extensively, in order to investigate if behavior-based clustering does outperform metadata-based clustering in these scenarios. Moreover, integrating both clustering approaches could potentially enhance the strengths of each method, eventually providing a more robust framework for handling annotator diversity in machine learning applications.

In conclusion, while clustering by annotator metadata and response patterns presents a viable method to enhance computational efficiency and model accuracy, it is crucial to balance these benefits against the loss of individual annotator insights. Based on the preliminary results observed in this thesis, further studies are necessary to refine these clustering techniques and fully realize their potential in practical applications.

Appendices

A. Appendix A

Table 9. Top 5 features per cluster

Clusters	Features	
cl 1	_gender_self_describe_True	_gender_self-describe
	_gender_self_describe_False	_sexuality_other_False
	_ideology_extremely_liberal	
cl 2	_gender_non-binary	_gender_non_binary_False
	_gender_non_binary_True	_sexuality_other_False
	_sexuality_other_True	
cl 3	_transgender_prefer_not_to_say_False	_transgender_prefer_not_to_say_True
	_trans_prefer_not_to_say	_cisgender_True
	_cisgender_False	
cl 4	_transgender_True	_transgender_False
	_trans_yes	_trans_no
	_cisgender_True	
cl 5	_income_>200k_False	_income_>200k
	income>200k_True	_educ_phd
	_age	
cl 6	_education_phd_False	_education_phd_True
	_educ_phd	_age
	income>200k	

Table 9. Top 5 features per cluster (continued)

Clusters	Features	
cl 7	_gender_prefer_not_to_say_True	_gender_prefer_not_to_say_False
	_gender_prefer_not_to_say	_gender_male
	_id	
cl 8	_education_professional_degree_False	_educ_professional_degree
	_education_professional_degree_True	_id
	_age	
cl 9	_religion_hindu_False	_religion_hindu_True
	_age	_id
	_race_asian_True	
cl 10	_sexuality_gay_True	_sexuality_gay_False
	_sexuality_straight_False	_sexuality_straight_True
	_age	
cl 11	_sexuality_bisexual_True	_sexuality_bisexual_False
	_sexuality_straight_True	_sexuality_straight_False
	_id	
cl 12	_educ_masters	_education_masters_True
	_education_masters_False	_education_some_college_False
	_age	
cl 13	_income_100k-200k	_income_100k-200k_True
	_income_100k-200k_False	_income_50k-100k_False
	_income_50k-100k	
cl 14	_religion_mormon_True	_religion_mormon_False
	_age	_id

Table 9. Top 5 features per cluster (continued)

Clusters	Features	
	_religion_christian_True	
cl 15	_sexuality_straight_False	_sexuality_straight_True
	_education_college_grad_ba_False	_education_college_grad_ba_True
	_education_some_college_True	
cl 16	_race_other_False	_race_other_True
	_id	_age
	_race_white_True	
cl 17	_religion_jewish_False	_religion_jewish_True
	_sexuality_other_True	_sexuality_other_False
	_id	
cl 18	_race_native_american_False	_race_native_american_True
	_id	_age
	_race_black_True	
cl 19	_education_college_grad_aa_True	_education_college_grad_aa_False
	_educ_college_grad_aa	_id
	_sexuality_straight_True	
cl 20	_ideology_conservative	_ideology_conservative_True
	_ideology_conservative_False	_education_high_school_grad_False
	_age	
cl 21	_ideology_slightly_conservative	_ideology_slightly_conservative_True
	_ideology_slightly_conservative_False	_race_white_True
	_race_white_False	
cl 22	_religion_buddhist_False	_religion_buddhist_True

Table 9. Top 5 features per cluster (continued)

Clusters	Features	
	_id	_age
	_religion_christian_False	
cl 23	_religion_muslim_True	_religion_muslim_False
	_id	_age
	_race_asian_True	
cl 24	_race_middle_eastern_False	_race_middle_eastern_True
	_id	_age
	_race_white_False	
cl 24	_ideology_extrem_conservative	_ideology_extrem_conservative_True
	_ideology_extrem_conservative_False	_id
	_age	

Bibliography

[1] Sindhu Abro et al. “Automatic hate speech detection using machine learning: A comparative study”. In: *International Journal of Advanced Computer Science and Applications* 11.8 (2020).

[2] Lora Aroyo and Chris Welty. “Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard”. In: *WebSci2013. ACM* 2013.2013 (2013).

[3] Lora Aroyo and Chris Welty. “The three sides of crowdtruth”. In: *Human Computation* 1.1 (2014).

[4] Lora Aroyo and Chris Welty. “Truth is a lie: Crowd truth and the seven myths of human annotation”. In: *AI Magazine* 36.1 (2015), pp. 15–24.

[5] Ron Artstein and Massimo Poesio. “Inter-coder agreement for computational linguistics”. In: *Computational linguistics* 34.4 (2008), pp. 555–596.

[6] Valerio Basile et al. “We need to consider disagreement in evaluation”. In: *Proceedings of the 1st workshop on benchmarking: past, present and future*. Association for Computational Linguistics. 2021, pp. 15–21.

[7] Federico Cabitza, Andrea Campagner, and Valerio Basile. “Toward a perspectivist turn in ground truthing for predictive computing”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 6. 2023, pp. 6860–6868.

[8] Tommaso Caselli et al. “Hatebert: Retraining bert for abusive language detection in english”. In: *arXiv preprint arXiv:2010.12472* (2020).

[9] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. “Dealing with disagreements: Looking beyond the majority vote in subjective annotations”. In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 92–110.

[10] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial bias in hate speech and abusive language detection datasets”. In: *arXiv preprint arXiv:1905.12516* (2019).

[11] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[12] Lucas Dixon et al. “Measuring and mitigating unintended bias in text classification”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 67–73.

[13] Eve Fleisig, Rediet Abebe, and Dan Klein. “When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 6715–6726.

[14] Tommaso Fornaciari et al. “Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 2021.

[15] Mor Geva, Yoav Goldberg, and Jonathan Berant. “Are we modeling the task or the annotator”. In: *An Investigation of Annotator Bias in Natural Language Understanding Datasets. ArXiv abs/1908.07898* (2019).

[16] Dirk Hovy and Anders Søgaard. “Tagging performance correlates with author age”. In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*. 2015, pp. 483–488.

[17] Dirk Hovy and Shannon L Spruit. “The social impact of natural language processing”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016, pp. 591–598.

[18] Dirk Hovy and Diyi Yang. “The importance of modeling social factors of language: Theory and practice”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 588–602.

[19] *huggingface measuring-hate-speech*. `https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech`. Accessed: 20-05-2024.

[20] Like Hui and Mikhail Belkin. “Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks”. In: *arXiv preprint arXiv:2006.07322* (2020).

[21] Chris J Kennedy et al. “Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application”. In: *arXiv preprint arXiv:2009.10277* (2020).

[22] Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. “Analyzing disagreements”. In: *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*. 2008, pp. 2–7.

[23] Elisa Leonardelli et al. “Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators’ Disagreement”. In: *arXiv preprint arXiv:2109.13563* (2021).

[24] Elisa Leonardelli et al. “SemEval-2023 Task 11: Learning With Disagreements (LeWiDi)”. In: *arXiv preprint arXiv:2304.14803* (2023).

[25] Sean MacAvaney et al. “Hate speech detection: Challenges and solutions”. In: *PloS one* 14.8 (2019), e0221152.

[26] Daniel Müllner. “Modern hierarchical, agglomerative clustering algorithms”. In: *arXiv preprint arXiv:1109.2378* (2011).

[27] Fionn Murtagh and Pierre Legendre. “Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion?” In: *Journal of classification* 31 (2014), pp. 274–295.

[28] Chikashi Nobata et al. “Abusive language detection in online user content”. In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 145–153.

[29] Matthias Orlikowski et al. “The Ecological Fallacy in Annotation: Modelling Human Label Variation goes beyond Sociodemographics”. In: *arXiv preprint arXiv:2306.11559* (2023).

[30] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. “Effective hate-speech detection in Twitter data using recurrent neural networks”. In: *Applied Intelligence* 48.12 (2018), pp. 4730–4742.

[31] Massimo Poesio and Ron Artstein. “The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account”. In: *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*. 2005, pp. 76–83.

[32] Fabio Poletto et al. “Resources and benchmark corpora for hate speech detection: a systematic review”. In: *Language Resources and Evaluation* 55 (2021), pp. 477–523.

[33] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[34] Vikas C Raykar et al. “Learning from crowds.” In: *Journal of machine learning research* 11.4 (2010).

[35] Giulia Rizzi et al. “MIND at SemEval-2023 Task 11: From Uncertain Predictions to Subjective Disagreement”. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. 2023, pp. 556–564.

[36] Filipe Rodrigues and Francisco Pereira. “Deep learning from crowds”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[37] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[38] Pratik Sachdeva et al. “The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism”. In: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*. 2022, pp. 83–94.

[39] Marta Sandri et al. “Why don’t you do it right? analysing annotators’ disagreement in subjective tasks”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023, pp. 2420–2433.

[40] Maarten Sap et al. “Annotators with attitudes: How annotator beliefs and identities bias toxic language detection”. In: *arXiv preprint arXiv:2111.07997* (2021).

[41] Maarten Sap et al. “The risk of racial bias in hate speech detection”. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 1668–1678.

[42] Maya Sen and Omar Wasow. “Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics”. In: *Annual Review of Political Science* 19 (2016), pp. 499–522.

[43] Ketan Rajshekhar Shahapure and Charles Nicholas. “Cluster quality analysis using silhouette score”. In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE. 2020, pp. 747–748.

[44] Arthur K Spears. “African-American language use: Ideology and so-called obscenity”. In: *African-American English*. Routledge, 2013, pp. 226–250.

[45] Alexandra Uma et al. “SemEval-2021 task 12: Learning with disagreements”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. 2021, pp. 338–347.

[46] Alexandra N Uma et al. “Learning from disagreement: A survey”. In: *Journal of Artificial Intelligence Research* 72 (2021), pp. 1385–1470.

[47] William Warner and Julia Hirschberg. “Detecting hate speech on the world wide web”. In: *Proceedings of the second workshop on language in social media*. 2012, pp. 19–26.

[48] Zeerak Waseem. “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter”. In: *Proceedings of the first workshop on NLP and computational social science*. 2016, pp. 138–142.

[49] Junwei Xiao, Jianfeng Lu, and Xiangyu Li. “Davies Bouldin Index based hierarchical initialization K-means”. In: *Intelligent Data Analysis* 21.6 (2017), pp. 1327–1338.

[50] Ziqi Zhang, David Robinson, and Jonathan Tepper. “Detecting hate speech on twitter using a convolution-gru based deep neural network”. In: *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer. 2018, pp. 745–760.