UNIVERSITY OF TWENTE.

Assessing Fairness in Machine Learning: The Use of Soft Labels to Address Annotator Bias in NLP

A Hate Speech Application

by

Abel van Raalte

A thesis submitted to the
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
in partial fulfilment of the requirements for the degree of

MSc in Business Information Technology

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

University of Twente

Enschede, Overijssel, The Netherlands

August 2024

ABSTRACT

The integration of soft labels in the learning process of Machine Learning (ML) models represents a relatively novel approach in the field of Natural Language Processing (NLP). Unlike traditional hard labels, which are typically derived through majority voting, soft labels represent probability distributions that capture the range of annotators' opinions during the annotation process. This approach allows for a more nuanced representation of annotator disagreement. This relatively new approach to aggregate and use labels has been researched in terms of performance and usability in practice. However, to the best of our knowledge the use of soft labels have not yet been explored in terms of fairness in AI. Algorithmic fairness can be defined as the practice of developing an algorithm which is not discriminatory or subjective to bias. Developing fair and unbiased models is a challenge in the field of subjective classification tasks. The use of hard labels often result in the loss of different opinions of annotators. This way a bias can form against annotator groups which are underrepresented in the data. Therefore, This thesis aims to assess the effects of using soft labels on fairness in ML models. The models are trained and evaluated on hate speech classification tasks. In line with this goal, the following main research is addressed during this thesis:

> *How can a soft label modelling approach, combined with bias detection methods enhance fairness in hate speech detection models?*

The datasets used in the thesis are sampled from tweets focused on hate speech. The tweets are annotated by multiple annotators. From each annotator various demographic attributes, such as age, gender and race are stored. Based on these attributes potential biases in the data are found. This is done by analyzing annotator disagreement between different groups of annotators, in combination with their level of representation in the dataset. These analysis showed that there are potential biases found in the dataset based on annotator subgroups. The subgroups are used to train and analyze models on both hard and soft labels. The resulting models are finally evaluated in terms of overall performance and the fairness metrics "disparate impact" and "equal opportunity". The results show that based on these metrics the models which are trained on soft labels indicate less bias against demographic subgroups compared to the models trained on hard labels. However, overall predictive performance does not improve when transitioning to soft labels. This thesis adds to the literature on annotator disagreement and soft labels by demonstrating that potential biases in data can be identified through agreement measures and descriptive statistics of annotations. Additionally, the research assesses the impact of soft labels on algorithmic fairness with respect to bias.

Keywords: Hate Speech; Natural Language Processing; Soft Labels; Annotations; Bias; Fairness

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Twente to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Twente to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Abel van Raalte

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the following individuals for their invaluable support throughout the course of this project:

I want to express my most sincere appreciation to my thesis supervisors, Marcos Machado, Daniel Braun, and Andrea Papenmeier, for sharing their expertise and guiding me in the development of my thesis. In particular, I would like to thank Marcos and Daniel for introducing me to the project, providing timely, comprehensive and consistent feedback, and for the moral encouragement given during the past year.

CONTENTS

List of Figures

List of Tables

1

INTRODUCTION

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) which is concerned with giving computers the ability to understand text and spoken language the same way human beings can [2]. For certain Machine Learning (ML) models to learn from data items, they need to know to what class the data item belongs. Assigning data items to a class which can be used in the training process of the model, is also known as labelling the data [3]. In the field of NLP, such labels are often necessary for algorithms to interpret and learn from data. Machine learning algorithms which are trained using labeled data are referred to as supervised learning algorithms whereas machine learning algorithms which are trained with unlabeled data are referred to as unsupervised learning algorithms [4, 5]. In unsupervised learning algorithms an assumption is made that the data contains underlying structural properties on which it can be analyzed [4]. Research in NLP can generally be focused on unsupervised learning [6–11], supervised learning [12–16], or semi-supervised learning (a combination of unsupervised learning and supervised learning) [6, 8]. The labeling process of NLP datasets often consists of multiple annotators which assign annotations/labels to each of the data items in the dataset. Aggregation methods, such as the majority vote, are used to establish a final label which can be seen as the "ground truth" [17]. The use of multiple non-expert annotators in the annotation process often has better results in terms of accuracy and algorithmic bias than using one expert annotator [18]. It is hypothesized that using a single expert annotator can have a bias when annotating a dataset, thus resulting is a biased algorithm [18]. Therefore, the current status quo in NLP uses the votes of multiple annotators to determine the "ground truth" through a majority voting system instead of using one single annotator. In other words, this means that the label with which the most annotators agree will become the final label of the data.

Some well-known examples of NLP applications which are considered subjective regarding data labelling are sentiment analysis [10, 11], hate speech detection [19] and topic modelling [20]. The datasets on which these applications are often trained, such as twitter datasets, are la-

beled using the multi-annotator approach. Most often, despite the exact nature of the dataset, the labeling process is considered to be subjective. The subjective nature of the labeling process means that one annotator might choose a different label for a data item than another annotator. This does not mean that each dataset is equally subjective in the context of labeling. Sentiment analysis such as detecting offensive language in tweets could be considered more subjective than an objective tasks such as determining in what language a tweet is written.

The phenomenon of spreading online hate speech on social media platforms is rapidly growing over the past few years [21]. Studies on the detection of hate speech in different settings, such as social media [22] and in law enforcement [23] have been explored.
Hate speech causes emotional pain [21], polarization and in general can lead to less desirable experiences on social media platforms. From a business perspective, using ML algorithms to effectively and automatically target hate speech on a social media platform could lead to a safer and more inclusive platform for users. Next to these benefits of removing hate speech from social media platforms it also could protect companies from legal repercussions and lawsuits for not complying with regulations.
When automated hate speech detection by ML algorithms is used to remove certain posts from social media platforms, the algorithm needs to make sound and robust decisions. Biased algorithms form a threat to principles of fairness in the decision making process of hate speech detection. Human annotators are crucial in the training process of ML algorithms when it comes to a subjective problem as hate speech.

Annotators' sociodemographic environment, surroundings, moral values and experiences often influence their interpretations on such subjective data, causing disagreement among annotators [12]. Using the majority voting system is a quick solution for such disagreements, however it disregards the variety in opinions, and thus potentially important viewpoints. Therefore, using the majority voting system to find the 'ground truth' of data items might not be the most optimal practice in the field of NLP.

A promising approach towards an alternative multi-annotator modelling approach is the use of soft labels. Soft labels are probability distributions of annotations, which can be used in the training process as training targets (labels) or in the loss function of a model. In recent years, studies have explored the use of soft labels [12–14]. These studies focus mainly on the effects of soft label on performance, and how to implement soft labels in the NLP pipeline. While these studies have significantly advanced our understanding of using soft labels in the modeling of NLP applications, they mainly focus on performance metrics like accuracy, precision and F1-score. Aspects concerning fairness in AI and biases remain left to explore.

Fairness in AI is a broad concept. The paper by Memarian et al. finds several descriptive definitions of fairness in literature. In general, the concept is described as the development of algorithms that do not create discriminatory or unjust consequences [24]. Fairness in the context of this thesis is in line with the aforementioned description of fairness, focusing on algorithmic bias towards often underrepresented minority groups in datasets.

This thesis aims to bridge this gap by investigating the use of soft labels in terms of fairness and bias, thus contributing to a more comprehensive view of the effects of this multi-annotator modelling approach. By doing so, this research seeks to offer new insights into how soft labels can be leveraged to mitigate biases and promote fairness.

These goals will be achieved by analyzing label variations across different demographic groups of annotators and using these sub-groups to evaluate models based on a soft label modeling approach. Based on the described research objectives, the study aims at providing an answer to the following main research question:

> *How can a soft label modelling approach, combined with bias detection methods enhance fairness in hate speech detection models?*

This research question is divided into the following sub-research questions:

- Can biases be detected in annotators labelling behavior based on their demographic information (age, gender, politics, race)?
- How effective are soft label based algorithms in mitigating biases in hate speech detection compared to hard label approaches?

These questions will be addressed by following the CRISP-ML(Q) [25] research cycle. Two datasets containing tweets focused on hate speech are used in the process of analyzing annotator demographics and their effects on labeling behaviour. These datasets contain unaggregated annotations from a large amount of annotators. Various demographic attributes, such as race, gender and age, are collected in the annotation process, making these datasets useful for the research purposes of this thesis. Long Term Short Term (LSTM) models are used in this research, in order to analyze the effects of soft- and hard labels on fairness metrics such as Disparate Impact Ratio and Equal Opportunity of the models (see Section 3.2).

This thesis provides an understanding of how demographic attributes influence labeling behavior of annotators. It establishes that biases can be detected based on these demographic factors, revealing substantial differences in labeling patterns. This work extends the existing literature by analyzing the difference of labeling patterns in terms of annotator agreement, and their relation to the detection of annotator bias. Moreover, the research contributes to the development of soft label modeling approaches in ML. The implementation and evaluation of soft labels, compared to traditional hard labels, are used to highlight their impact on Fairness in AI. These findings suggest that hard labels are more susceptible to bias than soft labels.

From a practical standpoint this thesis provides insights into the effects of soft labels on hate speech detection, making it directly applicable to NLP applications in practice. Next to that the distinct patterns found in labeling behaviour between demographic attributes and hate speech could be used into practice by adjusting the manner in which annotations are collected, to make datasets containing annotations from different groups more balanced.

The rest of this paper is structured as follows: Chapter 2 contains the literature review and is used to provide a background into the field of study concerning annotator disagreement, the majority voting method, and alternative approaches compared to the majority vote. Chapter 3 describes the methods such as the research cycle, models, and other techniques used during the course of this thesis. Chapter 4 describes the experimental setup of the experiment used in this thesis together with the validation methods. The results of the experiments are described in Chapter 5. The results are discussed in Chapter 6, together with the limitations of the experiment and results. The conclusion of the thesis can be found in Chapter 7.

2

LITERATURE REVIEW

This section provides an review of the existing literature in the field of annotator disagreement in NLP. First the methodology of the literature review is given. Secondly the existing methods of handling annotator disagreement in NLP are discussed. Finally future research area's are suggested.

2.1. METHODOLOGY

The first phase of the literature review consisted of a systematic review of papers found in Scopus [1], while the second phase consisted of a snowballing technique based on the already analyzed papers, resulting in a hybrid search strategy [26]. By dissecting the research questions and extracting important elements from each, these keywords were identified and used as input for the Scopus search query: "text annotation", "majority voting", "predictive framework", "text classification", "annotation quality", "inter-annotator disagreement", "crowd sourcing", "natural language processing". These keywords in combination with Scopus operators resulted in the following used input queries:

- "text classification" AND "majority voting"
- "predictive frameworks" AND "annotation"
- "annotation quality" OR "inter-annotator disagreement"
- "natural language processing" AND "annotator disagreement"
- "text classification" AND "crowd sourcing"

The queries were combined with a filter which only included papers in the subject area's of Computer Science, Social Sciences and Business/Management. The subject area of Computer

[1] https://www.scopus.com/home.uri

Science was chosen since this subject area is most directly related to the technical aspects of the topic. The subject area's Social Sciences and Business/Management were included since papers in these area's could help highlight the causes behind annotator disagreements, and which practical implications such disagreements have in business or social settings. From each query the resulting papers were evaluated based on their titles and estimated relevance to the topic (details provided in Figure 2.1). This process resulted in a list with 86 papers. These 86 papers were reduced to 37 papers based on a brief overview of the papers (abstract/introduction/conclusion). Upon the first brief overview of the selected literature, it became apparent that the term "majority voting" in machine learning has two interpretations. The first interpretation, which was assumed in this literature review, refers to the "majority voting" principle/method as a way to assign a final label to a data item based on the label that got the most votes among annotators [18]. The second interpretation of the term "majority voting" refers to a practice in using machine algorithms which uses multiple models together in order to predict a label (ensemble learning). The label which is predicted most often by the selected models ('voted' on), is selected [17]. Due to the multiple interpretations of the term, and the use of the term in the input queries on Scopus, some of the papers were deemed irrelevant after further examination. Thus the systematic literature gathering resulted in 20 relevant papers. To collect enough papers for a comprehensive literature review, the rest of the papers were gathered by using the snowball method [26] on the found papers. Based on the resulting papers and the main research question of this literature review, we analyzed main themes in the annotator disagreement literature. We identified a clear distinction of groups of papers focusing on three main methods to explore problems related to disagreement:

1. Defining a single label as "ground truth" [18, 27–30].
2. The use of multiple labels to train algorithms [13–16, 29, 31].
3. The use of clustering techniques on annotations or annotators [6, 7, 9–11].

Based on these topics further papers were acquired through snowballing. This finally resulted in 30 papers. Figure 1 depicts a systematic overview of the paper selection process.

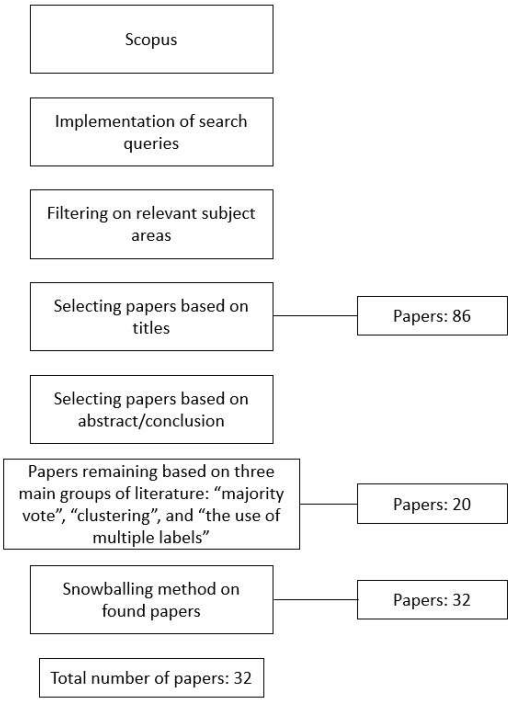


Figure 2.1: Paper selection process

The selected papers can be found in appendix A. This table summarizes the contents of the selected papers and contains five columns describing the papers: The author, main purpose, algorithms used, evaluation metrics used, setting of the research, and the general theme of the papers are described in that respective order. The last column, containing the general theme of the papers, is included to show the groupings of papers distinguished based on the predominant themes in literature. The abbreviation "DA" refers to papers relevant to causes of annotator disagreement and this group of papers is used to answer the first group of sub-research questions (6 papers). The abbreviation "MULTI" refers to papers focused on multi-label based architectures (10 papers). The term "CLUSTER" is used to refer to papers which use clustering techniques on NLP tasks as the main topic of research (6 papers). The term "OTHER" is used on the remaining papers, which are not solely classifiable to one of the other categories (10 papers).

2.2. GLOBAL ANALYSIS OF THE LITERATURE

This section aims to present a comprehensive overview of the reviewed literature, utilizing both visual representations and a narrative approach to highlight global trends and predominant themes in literature.

As mentioned in Section 2.1 of this paper we found 3 distinct groups of papers dealing with annotator disagreement (Single/gold label-, multi-label- and clustering based techniques). However, a more comprehensive layout of the literature can be made. In Figure 2.2, a global overview of the literature topics can be found.

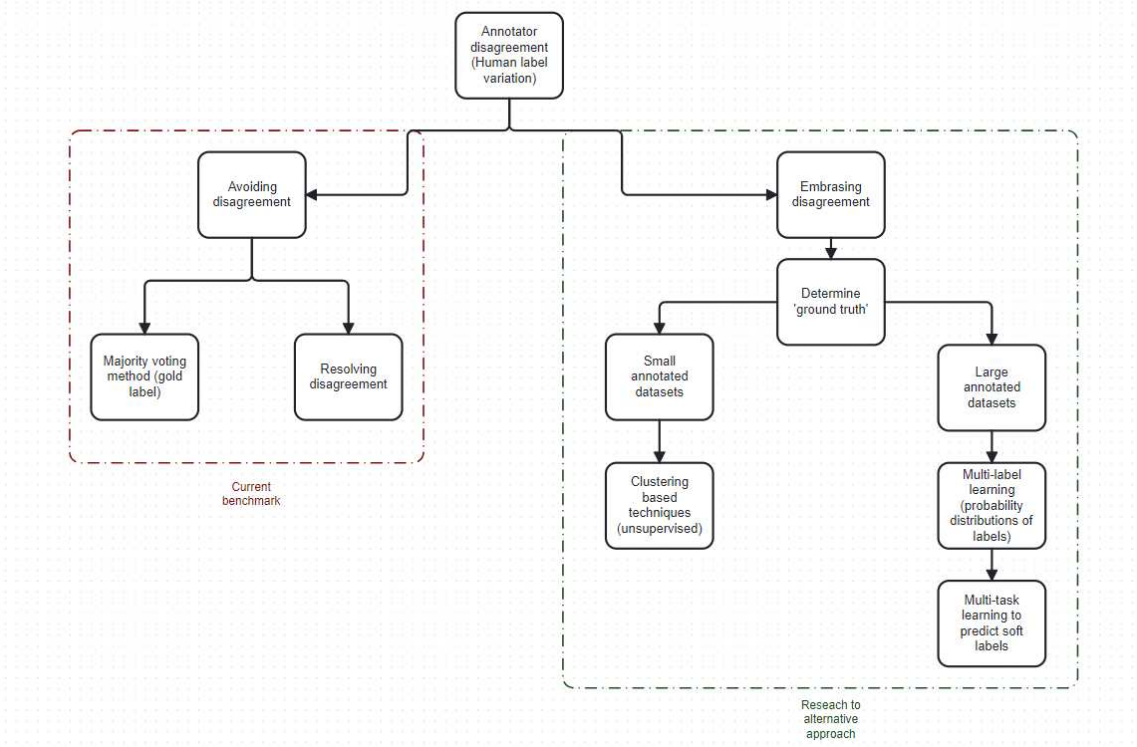


Figure 2.2: Global overview of literature themes

In the literature the first distinction on manually labeled data in NLP tasks can be made in the manner which annotator disagreement is addressed (see Figure 2.2). Papers focus either on **avoiding/resolving** disagreement [18, 32] or **embracing** disagreement [7, 13–15, 28, 29, 33]. Several papers advocate for furthering advancements in the former category [28, 30, 33].

The paper by Plank [28] refers to the concept of one data item having multiple different labels caused by annotator disagreement to "human label variation". This term captures the fact that inherent disagreement in annotation tasks can be due to genuine disagreements, subjectivity or simply because multiple views are plausible [28]. Plank provides an overview of aspects in the NLP pipeline which would benefit from taking human label variation into account instead of dismissing it [28]. Such aspects include the enrichment of the data collection process, modelling tasks and the evaluation of models with human label variation kept in mind.

Another study which argues for taking multiple labels into account in NLP tasks is presented by Cabitza et al. [33]. Cabitza et al. describes a concept called "data perspectivism". "Data perspectivism" revolves around moving away from the traditional gold standard datasets and towards the adoption of methods that integrate opinions and perspectives of annotators into the NLP pipeline [33]. This trend opposes the current benchmark methods in which disagreement among annotators/human label variation is either resolved/analyzed or dismissed.

Studies presented by [34–36] analyze disagreements among annotators and their causes. The paper by Ramirez et al. [37] proposes the technique of highlighting certain aspects of texts in

order to study its effect on human annotations. Since the causes and techniques these papers discuss can be reflected upon in order to resolve disagreements, they fall into the category of "resolving disagreement" in Figure 2.2. Another approach to resolve disagreement, is to assign a final label from an expert annotator when an disagreement can not be solved by majority vote [38].

Following a different approach, the paper by Snow et al. [18] provides empirical proof that using the majority vote among multiple annotators improves ML performance in terms of accuracy and reliability when compared to using a single expert annotator. Majority voting is easy to incorporate in labelling tasks since it results in one single label for each instance [12]. However, it does not aim to embrace perspective and annotator disagreement into NLP tasks.

Therefore, as mentioned in Section 2.1, there arise three predominant themes in literature when working with annotator disagreement:

1. Defining a single label as "ground truth" (majority vote/label aggregation) [18, 27–30].
2. The use of multiple labels to train algorithms (learning directly from the crowd) [13–16, 29, 31].
3. The use of clustering techniques on annotations or annotators [6, 7, 9–11].

Figure 2.3 displays a word cloud that highlights words that occur the most in the group of papers. The same is done for papers focused on incorporating multiple labels to train algorithms in Figure 2.4 and for papers focused on clustering based techniques in Figure 2.5. The word clouds provide an visual insight on terms often used in the corresponding groups of papers. As can be seen the terms which are highlighted in the word clouds correspond to the topics this literature review is interested in. When looking at words in the word cloud which are less highlighted and not part of the search queries used in the literature review, interesting terms arise. Examples of such terms are "distribution", "linguistics", "performance", "metrics", "quality" and "classification". These terms are valuable in the sense that they provide an idea of which aspects of the literature on annotator disagreement can be important to consider next to the obvious main aspects.

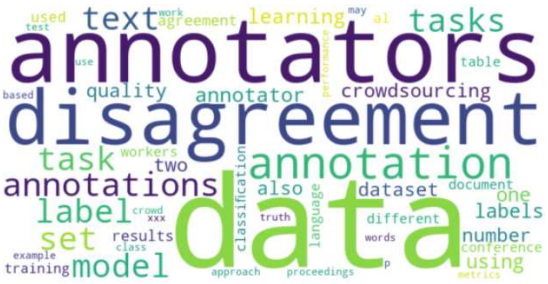


Figure 2.3: Word cloud of all papers combined

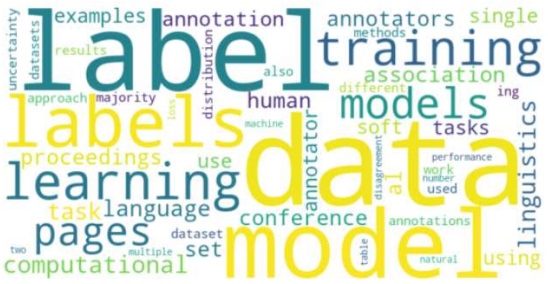


Figure 2.4: Word cloud of papers based on multi-labels

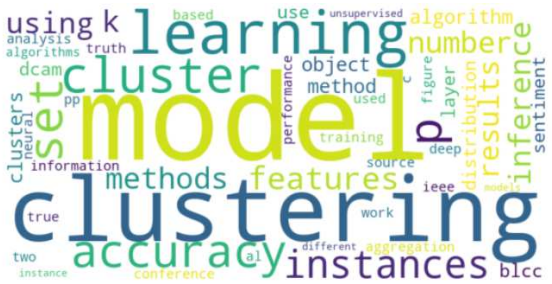


Figure 2.5: Word cloud of papers based on clusters

In Figure 2.6 a bar graph displays the distribution of papers and their year of publication. The publication years range from 2008 to 2023. In the Figure, a rise in publications can be seen since 2018, with a peak in 2021. This can indicate an increase of interest on annotator disagreement and how to handle it.

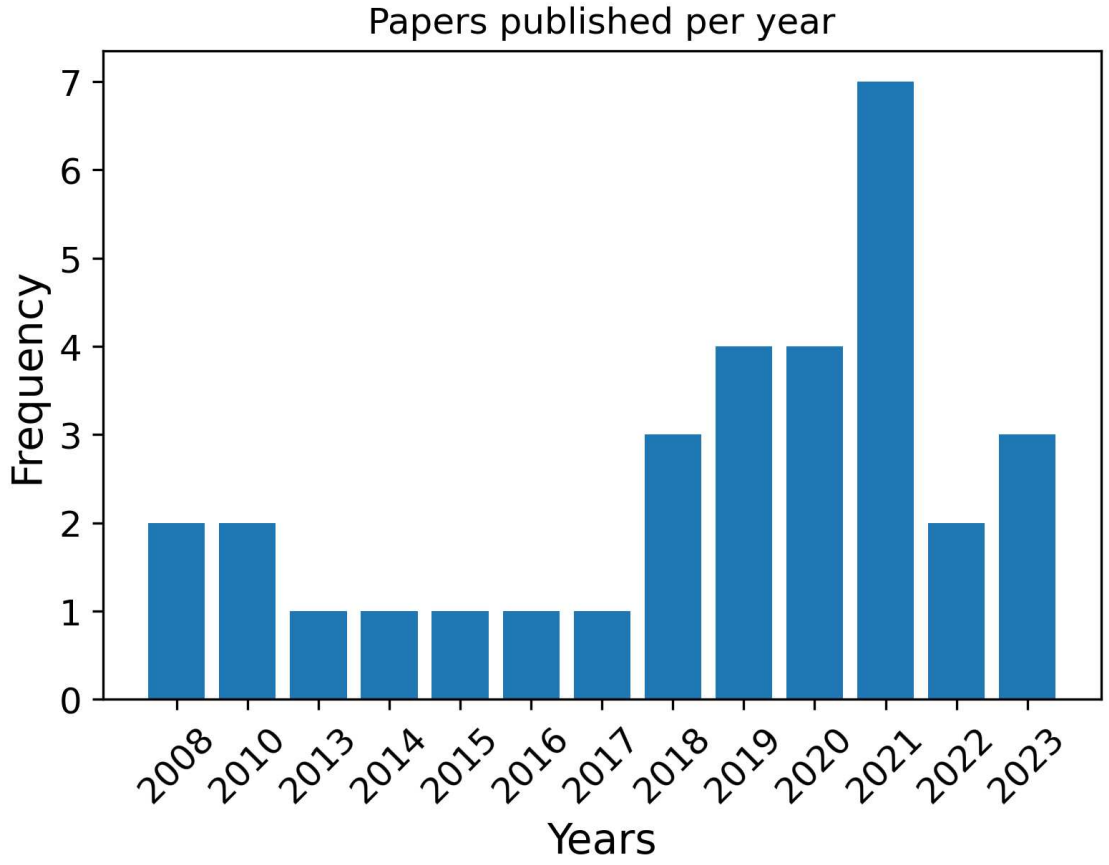


Figure 2.6: Years of article publications

2.3. NATURE OF DISAGREEMENT

To answer the first group of sub-research questions regarding how and why disagreements occur several papers are considered to be relevant [12, 32, 34, 35, 39]. In appendix A these papers can be recognized by the "DA" abbreviation in the "topic" column.

Disagreements between annotators arise on annotation tasks with a subjective nature [39] or are due to annotators lack of attention during annotation tasks [40]. Annotating data on which multiple annotators can form different opinions is what can be defined as a disagreement.

In literature several causes for such disagreements are given. It is argued that annotator disagreement mostly stems from a difference in annotators' knowledge, experiences and understanding of a subject [32]. This would mean that annotator disagreement could be solved by choosing annotators based on their knowledge and reasoning skills on a subject. Another cause of disagreement among annotators is the lack of clear task descriptions [34, 39]. Literature shows that vague annotation task description result in different interpretations of such tasks, leading in different labelling outcomes. [34] shows that in short sentence translations a more clear and detailed description of the annotation task results in less annotator disagreement. Finally it is argued that the annotators' sociodemographic factors, moral values, and lived ex-

periences often influence their interpretations [12]. This insight, combined with the subjective nature of annotation tasks in general, explains why disagreements among annotators would occur. These reasons for annotator disagreement are the cause for variation in the labels given to data items. While a part of these disagreements stem from logistical errors, a large part is caused by a genuine fluctuation in opinion. The latter becomes an important feature when taking fairness in AI into account since these labels can not be dismissed as noise, and are thus making label aggregation practices more difficult.

There are several ways to measure and keep track of annotator disagreement in a dataset (see "Evaluation" column in Appendix A). Inter-annotator agreement metrics like Cohen's Kappa and Fleiss' Kappa are used in [35, 39]. Other metrics like inter annotator agreement F-measures [18, 32] are also used to empirically determine the level of agreement/disagreement.

To summarize, it can be concluded from literature that here are various factors (sociodemogrhapics, task guidelines, subjective nature of a task, annotators experiences and knowledge) which can cause inter-annotator disagreement during the labeling process of NLP tasks. Such disagreements often occur in tasks with a high level of subjectivity. Examples of such tasks found in literature are sentiment analysis [10, 11], emotion recognition [35, 39] and NER [32]. These insights are relevant to this literature review in order to gain a better understanding in why disagreements occur. Looking at the causes of annotator disagreement it can be concluded that a substantial part of these disagreements can be of value since they represent different viewpoints, and can thus not be dismissed as label noise. Label noise can be described as the labels which are rendered useless due errors in the labelling process (e.g. unclear task description, lack of attention or knowledge from annotator).

2.4. EXISTING METHODS

In this section, a more in depth explanation of current methods for handling disagreement among annotators will be given. The predominant themes in literature, as mentioned in Section 2.1, will be discussed. Initially, the discussion focuses on the existing benchmark, the majority vote. Following that, two innovative approaches, involving the utilization of multiple labels and clustering techniques, will be addressed. Papers related to these groups of new approaches are marked by the tag "MULTI" and "CLUSTER" respectively in the "topic" column in Appendix A.

2.4.1. MAJORITY VOTE/LABEL AGGREGATION

The first, and most commonly used method in handling annotator disagreement is called the "majority vote" [18, 27]. The majority voting method is used to determine a ground truth based on the number of votes a label receives by multiple annotators. Using the majority voting when dealing with multiple annotators results in a single 'gold' label of each data item. Training algorithms on single labels is also referred to as Single-Task learning (STL) [13]. Handling disagreements in such a manner is easy since the outcome will always result in a single label, however

it does not take "human label variation" [28] into consideration. The majority voting method is proven to be effective even when dealing with only a small number of annotators and is thus easily applicable in practice when the 'ground truth' needs to be determined [18].

2.4.2. MULTIPLE LABELS

Another, more novel, approach is learning directly from the multiple labels acquired from the pool of annotators [12–16, 28, 29, 33]. This can be done using soft-labels [13, 14, 29] or hard-labels in either single task learning- (STL) or multi-task learning (MTL) algorithms [13].

First, the difference between hard- and soft labels will be explained. Next, STL and MTL are explained in the context of NLP. Finally an explanation will be given on how these concepts relate to each other and how it can be used to include annotator disagreement in the learning process of an NLP task.

SOFT-LABEL VS HARD-LABEL

To find a 'ground truth' in NLP tasks multiple annotations are often used and combined in order to create a single label for a data item. Such type of label is often referred to as a 'hard' label and classifies a data item in one definite category. This can result in an overconfident algorithm when the quality of annotations of is of low quality or subject to high levels of annotator disagreement [41]. Soft labels on the other hand allocates probabilities to each data item instead of one hard label, thus mitigating the overconfidence problem [41].

STL VS MTL

Multi-task learning (MTL) is a type of ML algorithm which is trained to perform multiple tasks simultaneously. Through MTL, information across multiple different tasks can be shared and leveraged to improve the general performance of a model. [42]

In NLP context, MTL is used to include multiple annotations in the learning process of a model. A MTL algorithm performs multiple tasks, whereas a single-task model only performs one task. When dealing with annotator disagreement, an MTL algorithm considers each annotator/annotation as an separate task. It learns from each annotator separately and combines this information to make predictions. Another application of MTL in NLP is the prediction of the soft label distribution after predicting the gold labels, as is done in [13].

COMBINING HARD LABELS WITH DISAGREEMENT METRICS

Combining hard labels (majority voting) with information about the disagreement among annotators, and including this information in the loss function during the training process is an another approach in NLP tasks utilizing multiple labels [29, 43–45]. This approach uses the unaggregated annotations from each annotator, however it does not necessarily mean that all the labels are used directly in the training process. The annotations are used in the form of disagreement metrics and included in the loss function of the algorithm. This approach is argued to increase the effectiveness of the loss function and thus improve the learning process of the models.

Metrics like inter-annotator F1-scores or confusion probability between annotators are used in the loss function in [29]. Cohen's kappa or Fleiss' kappa are other examples which can be used in the loss function to improve performance [39].

ALGORITHMS AND LIMITATIONS

Several studies uses a pre-trained BERT (Bi-directional encoder representations from transformer) for classification tasks [12, 15, 39, 46]. Other types of algorithms include Support Vector Machines (SVM) and Gaussian processes (GP) [31, 43], LSTM RNN and RNN's [13–16].

Amazon Mechanical Turk as an platform to collect non-expert annotations on data, in order to annotate datasets [18]. This platform allows for a cheap and relatively fast collection of crowd-sourced annotations.

For hard label based algorithms metrics like accuracy, F1-score, Recall, variance of annotations (model uncertainty) are used. For the soft labels, the produced probability distributions of the models are compared to that of the full distributions of the annotators using cross-entropy [16]. Furthermore in some literature classification accuracy is compared against different labeling techniques [15]. Next to that inter-annotator measures like the F1-scores between annotators on individual tasks and confusion probabilities are used [29].

There are several limitations found in literature regarding multi-annotator architectures. Firstly, it is mentioned in several papers that there are not many datasets available which provide not only the majority vote label, but also the unaggregated annotations on each data item [10, 28, 33]. Another limitation in current literature is the computational expensiveness of using a large number separate annotator heads [12]. Clustering annotators based on their annotating behavior is suggested as a possible solution. Next to these limitations there are only a few papers which assess the applications of different multi-label strategies in real-world scenario's. More research in such applications could improve the explainability and usability of algorithms based on multiple labels.

From the literature it can be concluded that the exploration of learning from multiple labels provided by annotators represents a dynamic and innovative approach in the field of NLP. This approach, as is explored by various studies [12–16, 28, 29, 33], involves the utilization of soft-labels or hard-labels in both STL and MTL algorithms. The discussion on the use of hard labels versus soft labels sheds light on how multiple annotations can be effectively integrated to enhance the learning process. Next to that, the incorporation of disagreement metrics in the loss function, such as inter-annotator F1-scores or confusion probabilities, highlights a method in which multiple annotations can be included in the learning process.

Despite these advancements, the literature highlights certain limitations, such as the scarcity of datasets providing unaggregated annotations and the computational challenges associated with employing a large number of separate annotator heads. The suggestion to address these challenges through clustering annotators based on their behavior reflects the ongoing efforts to overcome practical obstacles. Furthermore, the call for more research to assess the applica-

tions of different multi-label strategies in real-world scenarios emphasizes the need for further exploration to enhance the explainability and usability of algorithms relying on multiple labels. The exploration of learning from multiple labels not only contributes to the theoretical understanding of NLP tasks, but can also lead the refinement of practical applications. Finally it can lead to advancements in terms of robustness of machine learning models trained on multiple annotations.

2.4.3. CLUSTERING TECHNIQUES

The final alternative approach to handling disagreements among annotators revolves around clustering techniques. Clustering belong to the branch of ML algorithms called unsupervised learning algorithms [4]. Most NLP tasks uses algorithms trained on manually labeled data (supervised learning). Using labeled data often requires the use of human annotators to manually annotate data. As described in Section 2.3, this often leads to disagreement among annotators when dealing with subjective annotation tasks.

Clustering is often not used in NLP tasks due to a lower accuracy and less stable results [11]. However, unsupervised machine learning algorithms do not deal with labels, and therefore avoids disagreements which arise during the labeling process. Next to that, clustering can also be used as an approach to share labels on datasets with an small amount of annotators [6].

Another application of clustering in NLP tasks is to detect and reduce label noise in datasets. Clustering can be used to detect outliers (noise) in the labels based on the clustering of data items. When assumed that data items in a cluster belong to the same class/label, otherwise labeled data items can be seen as noise [9]. This does not include annotator disagreement directly in the training process, but it does in theory improve inter-annotator agreement metrics due to the removal of outliers.

In the literature concerning clustering techniques several algorithms are used and compared. The most occurring algorithm is K-means, which is used for the clustering itself [8, 10, 11, 47]. Other algorithms used for clustering include FMM, GMM and LDA [47]. The paper by Yin et al. [9] builds a deep clustering-based aggregation model (DCAM) based on several clustering models. Next to clustering algorithms, text pre-processing techniques like TF-IDF are used in order to represent textual data into vectors [11]. These vectors are then used in the training process of the algorithm.

Validation is mostly done through assessing the clustering quality, analyzing sentiment scores for each tweet, and other metrics like accuracy [10, 11]. Accuracy is mostly measured against other state-of-the art algorithms [8]. Clustering quality is analyzed through well-known metrics like KL divergence, Euclidean distance and Chebyshev distance [6].

Clustering could be used in the training process of ML algorithms in NLP tasks either in combination with supervised ML algorithms (semi-supervised) or stand-alone (unsupervised) [6–11].

In the literature concerning disagreement among annotators, models purely based on unsuper-

vised learning techniques are rare. It is mentioned that the reason for this is that unsupervised techniques in general result in poorer performance in terms of prediction accuracy [10, 11]. Further research could be conducted to improve performance of unsupervised learning algorithms.

Further evaluation on how clustering techniques can be used to increase the size of annotated datasets [47] or distinguish noise from a variation in opinion [28] could be important. Advancement in this field would result in clustering techniques as a supportive tool to already existing label-based approaches (semi-supervised learning).

In conclusion, clustering can be integrated into the training process of NLP algorithms either in conjunction with supervised ML algorithms (semi-supervised) or independently (unsupervised). Further research is needed to enhance the effectiveness of unsupervised learning algorithms in NLP tasks. From literature it becomes evident that future research directions include the evaluation on how clustering techniques contribute to increasing annotated dataset sizes, distinguishing label noise from variations in opinion, and can be used as stand alone algorithms in NLP tasks.

2.5. FINAL CONSIDERATIONS

Inter-annotator disagreement during the labeling process of NLP tasks is influenced by various factors, including sociodemographics, task guidelines, the subjective nature of tasks, annotators' experiences and knowledge. Such disagreements are particularly prevalent in tasks characterized by a high level of subjectivity, such as sentiment analysis, emotion recognition, and Named Entity Recognition (NER), as reported in the literature. Inter-annotator agreement is mostly measured with measures such as Cohen's kappa and inter-annotator agreement F-measure.

The majority voting method in determining a ground truth on subjective data items is often used and can be seen as the current benchmark for NLP tasks. However, it fails to entail perspective in the training process of an algorithm due to its dismissal of human label variation.

existing frameworks for handling annotator disagreement in NLP tasks encompass the majority vote, multiple label learning (utilizing soft or hard labels in STL or MTL), combining hard labels with disagreement metrics, and clustering techniques. While the majority vote is widely used and effective, newer approaches such as learning from multiple labels directly or incorporating disagreement metrics into the loss function are gaining attention. These new methods mostly perform equally or even better than the majority vote in terms of predictive accuracy, but more importantly are considered to provide a less biased view on data. Clustering techniques, although less common, offer an alternative perspective, particularly in reducing label noise by using the option to exclude human annotators. The choice among these methods depends on the nature of the NLP task and the specific challenges posed by annotator disagreement.

Although distinct groups of papers are available for each of the predominant themes in litera-

ture, there is some overlap in the methodologies. It is suggested in literature focused on multi-label based architectures that clustering techniques can be used to increase dataset size and overcome issues concerning computational power.

2.5.1. LIMITATIONS

While the literature review provided interesting insights concerning annotator disagreement in NLP tasks, there are some limitations which are to be considered when analyzing the results:

1. **Scope of Search:** While efforts were made to ensure a comprehensive review, the findings are contingent on the databases searched and the search terms used. Despite utilizing Scopus and employing a snowballing technique, it's possible that some relevant studies may have been omitted, introducing a potential bias in the selection of literature.
2. **Publication Language Bias:** The literature search was conducted in English, and studies published in other languages were not considered. This language restriction may have led to the exclusion of relevant research, particularly in regions where English is not the primary language of scientific communication.
3. **Overemphasis on NLP:** The focus of this review is mostly limited to annotator disagreement in NLP tasks. While this specificity allows for an in-depth exploration of a particular domain, it may overlook insights and perspectives from related disciplines that could contribute to a more holistic understanding of annotator disagreement.

2.5.2. GAPS IN LITERATURE & VENUES FOR FUTURE RESEARCH

Based on the review of the current literature, we have defined the following focus area's for future research:

1. **Developing an approach to understand on what data items disagreement occurs, and using this information to train models to predict items on which disagreement is likely to occur.** This could be combined with an evaluation of multi-label based methods on metrics that take inter-annotator agreement into account in order to increase an algorithms robustness.
2. **Developing a model which assesses the applicability of multi-label based models on real-world applications, in order to improve and understand the usefulness of such models in practice.** Current models show solutions utilizing the unaggregated labels from multiple annotators to train their algorithm with. However, these models lack in explainability/usability towards practice. Future research could consider these algorithms and how they can be applied in practical domains or industries.
3. **Exploring transferability of multi-label based models across different domains or industries.** This research gap fits the previous item, but is more focused on the transferability of a multi-label based approach onto specific NLP tasks.
4. **Comparing and optimizing different multi-label based algorithms.** A comprehensive

comparison of multi-based algorithms can provide valuable insights and lead to improved performance.

5. **Exploring possibilities in enlarging NLP datasets by combining and clustering unlabeled and labeled data.** This can be combined with techniques which cluster annotators based on their characteristics (NBP) [6]. Current literature mentions the lack of usable data for NLP tasks. Mainly datasets in which the multiple labels in unaggregated form are still available are rare. Next to that a limitation in current literature is the computational expensiveness of using a large number separate annotator heads.

6. **Developing a technique that distinguishes true difference in opinion from label noise.** This could be done by expanding the previous item on the use of clustering techniques to detect outliers, and combine these techniques with other inter-annotator agreement metrics to detect noise. This would increase the quality and thus the reliability of datasets.

7. **Developing an unsupervised (or semi-supervised) methodology for NLP tasks.** Developing such a methodology could evade the use of manually labeled data, and thus does not use human annotators. This would automatically avoid disagreements among annotators.

8. **Assessing the application of multi-label based algorithms in terms of explainable AI (XAI) and FAIR AI**. Taking these principles into account in the modeling process could help detect biases stemming from annotator groups [33]. This could also be combined with clustering techniques when enough data points per annotator are available to cluster them based on their characteristics.

3

METHODOLOGY

The primary objective of this research is to explore the impact of soft labels on fairness in AI, compared to the traditional use of hard labels. The experimental design is set up to identify potential biases and evaluate the performance of models trained on both hard and soft labels. Potential biases are found by analyzing disagreement between annotator groups, alongside imbalances in data representation. The trained models are evaluated on the annotator groups in relation to selected fairness metrics. Since the models are not only evaluated on overall predictive performance, but mainly on their bias against certain subgroups different evaluation metrics have been considered. For the disagreement analysis, Cohen's kappa has been used to measure inter-rater reliability between subgroups. This measure is supported by the Pearson's correlation to show that the disagreement and difference in labels between subgroups is significant. Finally the fairness metrics "equal opportunity" and "disparate impact" are used to evaluate the extent in which the different models shows bias when trained on both hard and soft labels. These metrics are used to draw conclusions towards the effects of soft labels on bias in hate speech detection models. Since this research revolves around the production of a ML application, the CRISP-ML(Q) research cycle has been used.

3.1. CROSS-INDUSTRY STANDARD PROCESS FOR MACHINE LEARNING

CRISP-ML(Q) is a process model for the development of ML applications. The process model is derived from the CRISP-DM process model, which is the cross industry standard for data mining applications. The main difference between the DM and ML model is that the ML model covers a monitoring and maintenance phase to address changes in the application environment. Next to that the CRISP-ML(Q) model introduces a quality assurance methodology in each phase to ensure confidence in the ML model [25]. As shown in Figure 3.1, the CRISP-ML(Q) methodology is illustrated. There are studies that have applied the CRISP-ML(Q) research cycle in practice [48–50].

The CRISP-ML(Q) model consists of six phases, which consists of a waterfall life cycle with backtracking [25]. A visual display of the process model and its phases can be found in Figure 3.1, a more detailed description of each phase is given below:

1. **Business and Data Understanding:**
   The first phase is focused on defining business objectives, how to translate these objectives to ML objectives, to collect data and assessing the feasibility of the project.

2. **Data Preparation:**
   This phase consists of producing the dataset for the modeling phase. In Figure 3.1 it can be seen that the data preparation phase is not a static phase, but can be circled back to from the modeling phase. This is designed so that when erroneous data is revealed in the modeling phase, the data processing can be adjusted accordingly. Sub tasks of this phase consists of selecting and constructing data for both train and test sets (feature selection/engineering, handling class imbalances etc.), cleaning data and standardizing data.

3. **Modeling:**
   The modeling phase consists of selecting and designing models which are appropriate for reaching the objectives defined in phase 1. A literature review often shows which models are used in similar research, which can be used in deciding on the type of model and its construction. Sub tasks of this phase consists of model selection, defining of quality measures for the model and training the model.

4. **Evaluation:**
   The goal of the evaluation phase is to evaluate and validate the performance and robustness of the trained models. Next to that the evaluation phase can be focused on increasing explainability of the models, so that the model is more clear and easy to use for ML practitioners or end users of the application. Finally the results of the models shall be compared in terms of success criteria, as defined in phase 1.

5. **Deployment:**
   This phase deals with the deployment of the constructed and evaluated models of the previous phases in the designated field of application. Sub tasks include the selection of appropriate hardware, model evaluation under production conditions, assuring user acceptance and usability, minimizing risks of unforeseen errors and defining a deployment strategy.

6. **Monitoring and Maintenance:**
   When a ML application is used for a longer period of time, it is vital to monitor the performance over time to ensure that it does not deteriorate. The sixth and last cycle of the CRISP-ML(Q) process model deals with the monitoring and updating of the applied models in order to uphold the required level of performance.

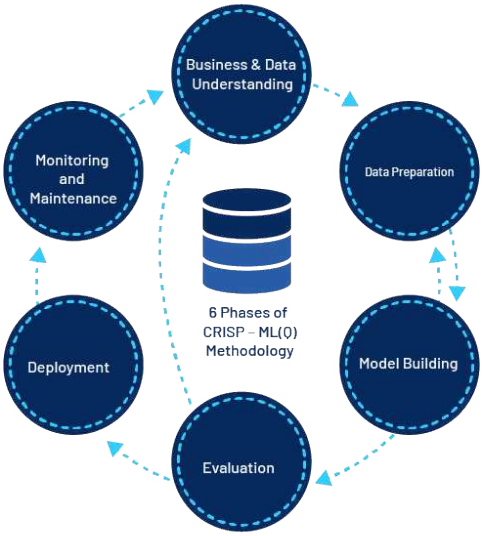


Figure 3.1: CRISP-ML(Q) process model [1]

3.2. ANALYTICAL METHODS

This section provides an explanation of the analytical methods used in the thesis. The concepts which are addressed are mostly applied in the "evaluation" phase of the CRISP-ML(Q) process model.

The goal of this thesis is to show the effects of using soft labels instead of hard labels in terms of fairness and bias. Towards this goal, the evaluation of the experiment will be done by comparing the performance of models trained with the use of different type of labels (hard labels and soft labels). In order to study the effects of an alternative approach to using hard labels hate speech classification, first an appropriate research group should be found within the available data. More concretely this means that certain sub-groups based on demographic features of annotators have to be found in which inter-annotator disagreement is presented. Finding such groups is supported by utilizing the following metrics concerning annotator disagreement:

- Cohen's Kappa is a measure for inter-rater reliability proposed by Cohen in 1960 [51]. The metric is essentially a chance-adjusted measurement of agreement between annotators or groups of annotators. Cohen's Kappa is used to evaluate different annotator groups in terms of agreement. The Cohen's Kappa is defined as follows [51]:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{3.1}$$

p_o = the proportion of units in which the annotators agreed

p_c = the proportion of units for which agreement is expected by chance.

- Pearson's correlation (r) is used to show correlation between the amount of positive hate speech labels given by annotators from groups based on demographic attributes. The correlation is interpreted as follows [52]:

Observed Correlation Coefficient	Interpretation
0.00–0.10	Negligible correlation
0.10–0.39	Weak correlation
0.40–0.69	Moderate correlation
0.70–0.89	Strong correlation
0.90–1.00	Very strong correlation

Table 3.1: Interpretation intervals of Pearson Correlation

These metrics, combined with a global exploration of the dataset, form the basis in the decision making process of which sub-groups in the data will be included in the experiment.

The first comparison in the experiment will be made by overall performance on the testing data in terms of accuracy, precision, recall and f1-score.

The second comparison consists of the evaluation of model performance on different demographic sub-groups in the data. The metrics used for this comparison are the following fairness metrics:

- **Disparate Impact:**
  This metric is used to determine whether a model or application has a form of unintended discrimination. It is measured by comparing the percentage of "favorable" or positive outcomes for a protected group against the percentage of positive outcomes for the reference group [53].
  The metric can be interpreted by using the 80% rule, which states that there is a disparate impact against a certain group when the disparate impact ratio is below 80% [53].

  The definition of the metric:

  $$\frac{\Pr(Outcome = \text{HATE SPEECH} \mid protected\_attribute = X)}{\Pr(Outcome = \text{HATE SPEECH} \mid protected\_attribute = Y)} \leq \tau = 0.8$$

  for positive outcome class HATE SPEECH and majority protected attribute (e.g. race, gender etc.) X where $\Pr(Outcome = outcome \mid X = x)$ denotes the conditional probability that the class outcome is $c \in C$ given protected attribute $x \in X$ of a protected group. Y where $\Pr(Outcome = outcome \mid Y = y)$ denotes the conditional probability that the class outcome is $c \in C$ given protected attribute of a non-protected group $y \in Y$.

  This metric is not directly applicable in the experiment of this thesis. In order for it to become applicable, a few modifications have to be made. In the original definition of the metric it is assumed that there is an favourable outcome a model can predict. Then the ratio with which the model classifies a protected group to the favourable outcome is compared to the same ratio of the unprotected group. Such a set up is not feasible in the

scope of this experiment since the tweets in the test sets of each demographic group are not identical. Therefore, the ratio of hate speech labels of one annotator groups can not be directly compared to that of another group without some form of normalization first. To solve this issue the disparate impact ratio is adjusted as follows:

$$\frac{\Pr(\text{Outcome} = \text{HATE SPEECH} \mid \text{protected\_attribute} = X)}{R_X}$$

where R_X is defined as the true ratio of hate speech label for the corresponding protected group of that attribute:

$$R_X = \text{Actual hate speech ratio of protected group X}$$

- **Equal Opportunity:**
  Equal Opportunity, also known as the recall of a model, is a fairness metric that can be used to determine whether a model is predicting a desired outcome equally well across different groups. In other words, it measures the true positive rate of a desired outcome. In this case it is used to compare the models' ability to detect hate speech across different demographic groups.

4

EXPERIMENTAL SET-UP

This section will describe the experimental set-up of this thesis. Subsections include the data collection process, pre-processing steps, set-up of the experiment and the method of validation.

4.1. DATA COLLECTION AND UNDERSTANDING

Hate speech is often targeted against peoples religion, ethnicity, nationality, race, gender or other demographic attribute a person or group has [54]. Such groups often form minorities, which may be underrepresented in the annotation processes of hate speech data. This may lead to an algorithm trained on said data to be biased against such a group.
The current benchmark method of aggregating all labels from the annotation process on a social media post to form a single hard label (gold label) may intensify the bias resulting from underrepresented groups in the data. Therefore, including all labels in the learning process of the ML algorithms might increase fairness in the model by reducing potential biases.
The dataset should therefore include the original labels from all annotators. Next to that, the dataset should include demographic information on the annotators. This is important so that they can be split into different groups in order to evaluate potential biases in the dataset.

Given that the research of this thesis focuses on bias and annotator disagreement in hate speech detection, the data needed to meet two requirements. First of all the dataset needed to include the unaggregated (raw) annotations of multiple annotators. This requirement arises from the need to analyze the disagreement between annotators and annotator groups. Secondly, the dataset needed to include demographic information (age, gender, race etc.) of the annotators. Including these annotator demographics in the annotation process allows an analysis of classification model performance on different subsets of the dataset based on these demographics.

The paper by Plank [28] provides a GitHub repository [1] dedicated on creating and publishing datasets with the aforementioned requirements. From this repository two interesting datasets suited for this thesis are selected:

- The first dataset used in the experiments is introduced in the paper by Maarten et al. [55] [2]. The dataset contains tweets, hatespeech labels of multiple annotators (varying per tweet) and annotator demographic information of annotators. The annotations are collected using the MTurk platform. The dataset contains 1,323 tweets which are annotated by 334 different annotators, which combined leads to a total of 11,706 rows. In table 4.1 the names of columns included in the dataset can be found. These columns include a selection of demographic attributes, which are important for the purposes of this thesis. Columns 0 up to 4 contain the demographic attributes such as age, gender, politics and race. These attributes are considered to find potential biases. The column "offensive2anyoneYN" contains the annotation (hate speech label). There are two columns containing an annotation of a tweet: "offensive2youYN" and "offensive2anyoneYN". The difference between the two types of annotations is that the latter is aimed at whether something is hate speech in general whereas the former is aimed at answering whether something is offensive to a that specific individual. For this research the more general label is used in order to establish among other things the ground truth on hate speech labels for the tweets.

annotatorAge	annotatorGender	annotatorMinority	annotatorPolitics
annotatorRace	intentYN	offensive2anyoneYN	offensive2youYN
dialectIsWrong	raceIsWrong	WorkerIdHashed	tweet
dialect	condition	username	davidson_label
founta_label	preprocessed_text		

Table 4.1: Columns dataset 1

- The second dataset dataset [3], among other things, collects the same type of information from annotators as the first dataset. The dataset is created by the Berkley University of California, and is introduced in the paper by Kennedy et al [56]. It contains 39,565 tweets annotated by 7,912 annotators, which combined leads to a total of 135,556 rows. This makes the Berkley dataset significantly larger in volume than the first dataset. This dataset contains 131 columns in total. This high number is partly due to the manner in which variables are stored. Where the column named "annotatorRace" in the first dataset can contain multiple values, the columns in the second dataset only contain boolean values. This design choice results in one demographic attribute with multiple values to

[1] https://github.com/mainlp/awesome-human-label-variation?tab=readme-ov-file
[2] http://maartensap.com/racial-bias-hatespeech/
[3] https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech/blob/main/README.md

cause an increase in column size. Next to that there are several more collected attributes of annotators stored in the second dataset, mainly related to income and social status.

These two datasets were chosen because they collect unaggregated annotations of annotators on a large selection of tweets focussed on hate speech. Next to that they both collect a broad selection of demographic attributes of annotators. This is vital to evaluate the annotations in terms of annotator disagreement, biases and the effect of soft labels on fairness in AI.

4.2. EXPERIMENT

The goal of this thesis is to increase fairness in hate speech detection models by developing a soft-label based method. For this the following experiment is defined:

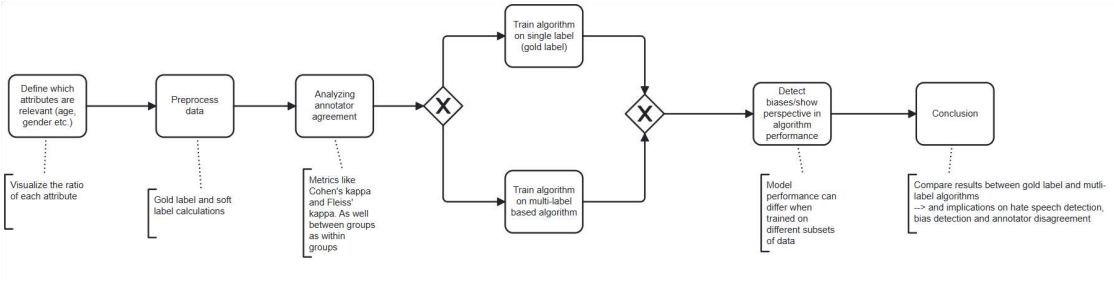


Figure 4.1: Experimental setup

In Figure 4.1 a global overview of the experiment can be seen. The first phase of the experiment is about finding relevant attributes to use for showing perspective and possibly bias in the data and algorithms. This will be done by analyzing disagreement metrics for several attributes and showing that there is a difference in annotating behaviour between annotators and annotator groups based on demographic features. Secondly the data has to be subjected to a series of pre-processing steps, after which the models can be trained. One model will be trained on the tweets and their aggregated labels (based on majority-vote). Another model will be trained on the tweets and their corresponding class probabilities (soft-labels). This will result in two trained models with different labeling based techniques. Finally the resulting model will be evaluated on different subgroups of data based on the in phase one defined demographic features. The results will be used to draw conclusions with regards to the performance of the novel approach of using soft-labels against the current benchmark of aggregating labels to a hard-label based on majority vote.

4.3. DATA PRE-PROCESSING

The processing of the data before it could be used can be divided in two sections: processing of the labels and processing of Twitter texts.

4.3.1. LABELS

A record of the original data format consists of an tweet, annotator id, several columns containing demographic attributes (age, gender etc.) and the hate speech label of the annotator. The hate speech label can be one of three options: "0", "0.5", or "1.0" for the first dataset, and "0", "1" or "2" for the second dataset. With 0 being a tweet classified as "not hate speech", 2 (or 1.0) being a tweet classified as hate speech. The option in between hate speech and not hate speech was "0.5" or "1". Since there were only a few instances where a tweet was actually classified as such an "in-between" label, the decision was made to remove the instances for more clarity in the training process of the models in later stages. The exact distribution of the labels in both datasets can found in Appendix B. To explore the problem of this thesis in terms of disagreement analysis and the training of the models the labels need to be collected per tweet rather than per annotator. This required some modifications to the manner in which the data is originally stored. Performing these modifications on different demographic groups of the data allows for the calculations of Inter-annotator agreement metrics within a group as well as the agreement between the groups.
The modifications to the dataset also allows for the aggregation of the labels per tweet to establish two different forms of labels: hard labels and soft labels. The hard labels, also considered to be the ground truth, are calculated by majority vote among all the annotations of the tweets. The soft labels are calculated by taking the total of votes per label (hate-speech or no hate-speech) of each tweet and normalizing these by dividing them against the total number of annotations for the corresponding tweet. This results in a label in the form of a probability distribution for each tweet, referred to as a soft-label.
These labels are used to execute and analyse the experiments, in which models are trained on both hard and soft-labels.

4.3.2. TEXT PROCESSING

The text of the tweets stored in the dataset consists of the "raw" tweet. This means that the texts have not yet been subjected to any form of pre-processing.

Regex are used in combination with the English stop word removal corpus from the NLTK package to clean the body of the text in python. The following steps are applied in the cleaning process:

- Making sure all characters are lowercase, in order to optimize further cleaning and processing steps the texts undergo. By making sure all letters are lowercase it will be made sure that identical words will be interpreted as such in the processing of the texts by ML models.
- Removing hyperlinks, usernames (e.g. "@user"), removing numbers, special characters and stopwords. These aspects of tweets usually do not contain important information re-

garding the sentiment of a text, such as hate speech. Removing them leads to less clutter and noise when transforming the texts into numerical sequences.

LEMMATIZATION

The next step in processing the texts of the tweets is lemmatization. Lemmatization is a data cleaning technique commonly used in NLP which transforms words to their vocabulary form, also known as the lemma of a word [57, 58]. For example, the verb "run" may appear as "running" "runs" or "ran" in a text. By applying lemmatization to these words all of them are transformed to "run". Using this technique as part of the text processing pipeline leads to a single representation of a word which can have different inflected forms. This can help identifying the underlying concepts and topics in the texts [57]. The "WordNetLemmatizer" from the "NLTK" package in python is used for this purpose.

VECTORIZATION AND PADDING SEQUENCES

For textual data to be used by ML models it first has to be transformed into vectors. Vectorization is a technique which transforms textual data into numerical data (vectors). The vectors created consists of only numerical values, and are padded with zero's to ensure uniform length. The first step to vectorize the data is called tokenization, where the text is separated into individual units called tokens[59]. In general those tokens can be words or characters. In the case of this experiment those tokens are the words of which a tweet consists. The tokens are assigned a numerical value, and a dictionary is created where each word is associated with a integer.
Secondly the tokenized texts are transformed into sequences of these integer indices. This step ensures that each tweet is represented as a sequence of numbers according to the dictionary created in the tokenization process. This step ensures that the text data is in such a format that it can be used for further processing by ML models.
Finally, in order to make the data compatible for neural network architectures like LSTMs, the data is padded. This process adds zeroes to sequences to ensure a unified length across all tweets.

For the ML models to interpret the textual data of the tweets the texts are first vectorized by using tokenization and padding techniques. These techniques are applied to transform the textual data to numerical data. This numerical data is then used by the models to learn from. The tokenizer (from Tensorflow Keras) transforms the texts into sequences. Since the texts, and thus the sequences, can vary in length they need to be padded. It is common to ensure that the sequences which are used as input for the models are of uniform length [60]. This is where padding is used. Padding the sequences reduces longer sequences to the agreed upon length, and increases shorter sequences by "padding" them with zero's to the agreed upon length.

4.4. MODELLING

4.4.1. LONG SHORT-TERM MEMORY

A Long Short-Term Memory (LSTM), designed by Hochreiter and Schmidhuber in 1997 [61] is a type of Recurrent Neural Network (RNN) architecture that can be used to model textual data.

As opposed to a normal RNN the LSTM can capture long-term dependencies, making it useful in text-classification.

The layout of a LSTM consists of three "gates": The forget gate, the input gate and the output gate. The forget gate decides which part of the previous cell state to discard, the input gate adds new information to the cell state and the output gate decides which part of the cell state to output [61].

The specifications of an LSTM model can be customized to suit specific applications. Examples of such variables are the number of hidden layers, their size and the choice of activation function in hidden and output layers.

A Bi-directional LSTM (Bi-LSTM) is a variation of an LSTM in which the model can learn in both directions. Where a LSTM can only learn from past to future, a bi-directional variant can learn both ways. Both LSTMs and Bi-LSTMs are used in sequential modelling tasks, such as classification tasks and sentiment analysis [14–16].

Bi-LSTM models are used for the modelling phase of this thesis, since these type of RNNs are used commonly in existing literature on similar tasks [14–16]. For this experiment two different type of models needed to be constructed. One model which is suitable for the classification of hate speech using the hard labels, and one which is suitable for the soft labels. Both models contain the following layers:

- **Embedding Layer (input layer)**
  This layer handles input data by transforming the padded sequences to dense vectors.
- **Two Bi-directional LSTM Layers**
  Each layer contains 64 hidden units.
- **Dense Layer**
  This layer has 24 hidden units, each applying a ReLU activation function. A L2 regularization of 0.01 is applied in this layer.
- **Output layer**
  The output layer consists of another Dense layer, with 1 unit for binary classification. A Sigmoid function is applied to output a value between 0 and 1.

The difference between the model for soft labels compared to the one suitable for soft labels can be found in the output layer of the model. For the hard label classification one unit is required, since it only deals with binary outcomes (either 0 or 1). Binary classification can be seen as a variant of multi-class classification, where there are only two classes. For a model to handle soft labels, it automatically needs to be treated as a multi-class classification problem.

For the use of soft labels the model is adjusted so that it handles the labels as a probability distribution of the votes per label. In practice this means that the model handles the modelling problem as a multi-class classification. To achieve such a model the final Dense layer (output

layer) needs to be adjusted in the following manner:

- The number of units needs to be changed from 1 to 2, in order for the model to output a multi-class label in the form of [prob_label0, prob_label1].
- The activation function is changed from a sigmoid function to a softmax function, which is suited for multi-class classification.

5

RESULTS

This Chapter will provide the results of the performed experiment. The experiment is performed to gather insights on the the use of soft labels and its effects towards bias and other fairness metrics as opposed to the current benchmark method of using hard labels. The results section has the same layout as the experiment described in Chapter 4: First the results of the global exploration of both datasets is done to identify which demographic sub-groups will be used during further steps in the experiment. Secondly the results of the annotator disagreement metrics are displayed to show how the extent of disagreement across different annotator subgroups. Finally the results of the models are provided. This part of the results is divided in the overall performance of the models, and the results of chosen Fairness metrics to show the effect on bias.

5.1. IDENTIFYING POTENTIAL BIASES ACROSS DEMOGRAPHIC GROUPS

The first step in the experiment was to find appropriate sub-groups of annotators and annotations to use. Since the experiment revolves around the effects of using soft labels on the models bias and fairness the selection of demographic groups has to fulfill certain characteristics. For a demographic group to be considered there should be:

- A difference in number of annotations and annotators from a certain subgroup. An imbalance of a certain sub-group can be an indication towards a bias in the dataset. Thus a group which fits this criterion might lead to useful results.
- Inter-annotator agreement between sub-groups, measured in Cohen's kappa, should show low levels of agreement (<0.40). If different sub-groups of a demographic attribute disagree with each other this may lead to a detectable bias in the data which can be used for

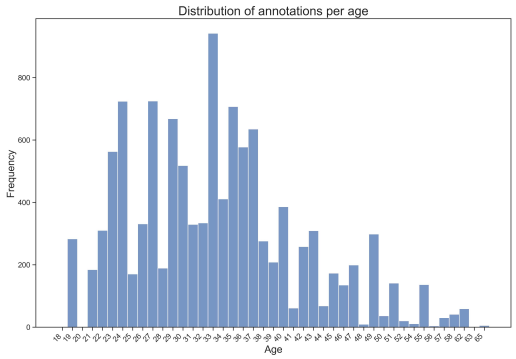
the experiment.

- A correlation between a positive hate speech label and a change in value for the demographic attribute. If there is a disagreement between sub-groups for a demographic feature, a correlation coefficient should be calculated in order to presume a bias.

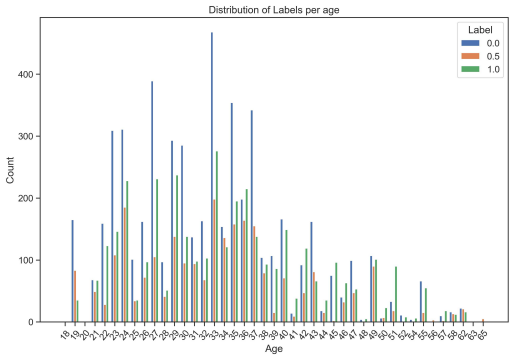
As mentioned in Chapter 4 there are two datasets used during the experiment. The first dataset contains less unique tweets, which makes it more difficult to use when training models. Training a model with a small amount of training samples means that the model has less instances to learn from. This can lead to poor overall predictive performance and biases on unseen data [62]. Due to a label imbalance, combined with the limited size of this dataset another dataset was considered. The second dataset contains just as many demographic attributes, and contains more annotated tweets than the first dataset. Therefore, this dataset is used to train and evaluate the models during the experiment. The first dataset is still used to show merit in using certain sub-groups for the evaluation of the experiment. So both datasets are used to find useful demographic attributes, while only the second (larger) dataset is used for modelling processes.

5.1.1. AGE AND GENDER

The age of annotators is the first demographic attributes which has been explored. This attribute is filled in by most annotators, making it a promising attribute for evaluation.
In Figure 5.1a and 5.1b the distribution of annotations per age and the labels per age are displayed.
It can be seen that younger ages (20-40 years) are more represented in the data than older ages (40+ years). This difference in representation is an important characteristic of the data which can lead to biases in the model.



(a) Distribution of ages



(b) Labels per age

Figure 5.1: Comparison of age distributions and labels per age (dataset 1)

Next, the difference in voting behaviour has to be considered. For this purpose the ratio of offensive hate speech labels per age group and the correlation between this ratio and age is used. In Figure 5.2 an increase in the ratio of annotated hate speech labels can be seen in older age groups, with a slight drop in the last age group. In Figure 5.3 the ratio be age can be seen

together with the calculated Pearson correlation and corresponding p-value.

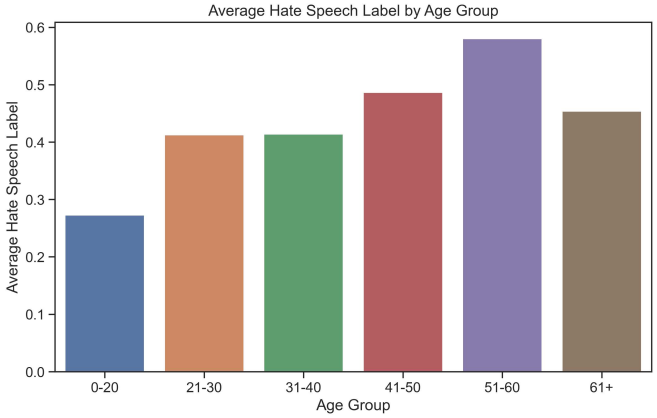


Figure 5.2: Hate speech ratio for age groups (dataset 1)

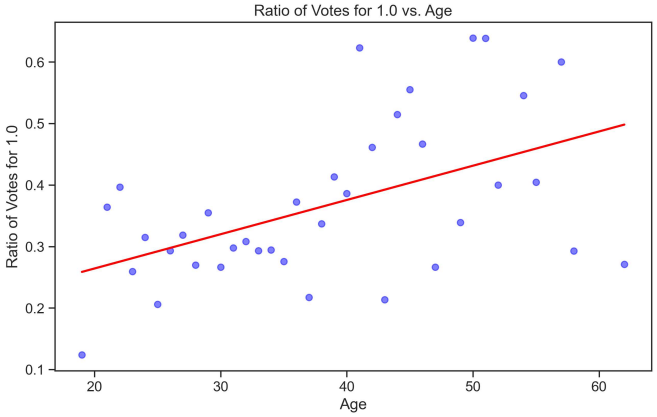


Figure 5.3: Hate speech ratio for each age – Pearson correlation: **0.502**, p-value: **.0016** (dataset 1)

The Pearson correlation (r) of 0.502 suggests a moderate positive linear relationship between a positive hate speech label and an increase in age. The p-value (p) of 0.0016 indicates that the correlation is statistically significant at the 0.05 significance level, with a confidence level of 95%. The red regression line in the figure shows a visualization of this linear relationship. Based on these insights in the data, the age attribute suggests to have a sound basis to be considered in the experiment.

Since the second dataset is used for training purposes the same process has been followed as with the first dataset to ensure a that the chosen sub-groups are suitable for evaluation in terms of bias. Figures 5.4, 5.5 and 5.6 provide visual insights into the distribution of ages, the ratio of hate speech per age group and its correlation of dataset 2.

Similar trends as with the first dataset can be seen. Younger age groups (20-40 years) are more represented in the data than older age groups (40+ years), as can be seen in Figure 5.4.
The increase in positive hate speech labels is less clear compared to the first dataset, but still visible in Figure 5.5. When looking at the scatter plot in Figure 5.6 this increase becomes more

clear. The Pearson correlation (r) of 0.61, and a P-value (p) of less than .00001 indicates a positive linear relationship with a moderate correlation slightly higher compared to the first dataset.

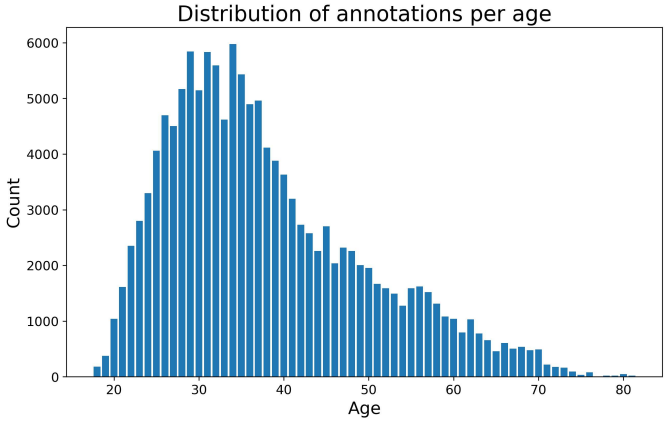


Figure 5.4: Age distribution (dataset 2)

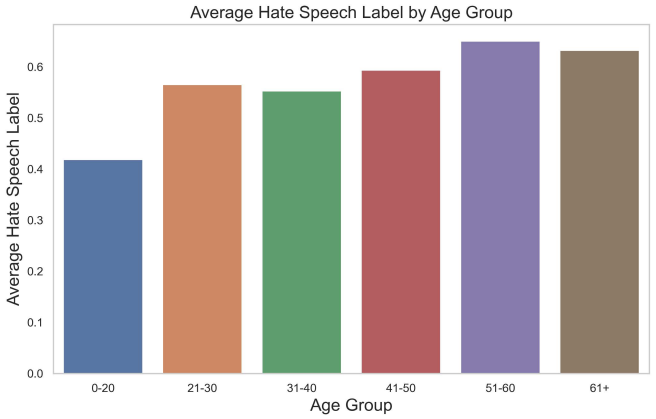


Figure 5.5: Hate speech ratio for age groups (dataset 2)

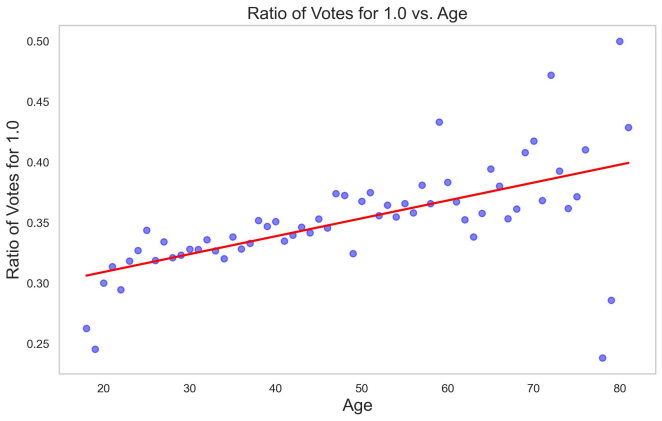


Figure 5.6: Hate speech ratio for each age – Pearson correlation: **0.610**, p-value: $\mathbf{1.14 \cdot 10^{-7}}$ (dataset 2)

The correlation analysis on the relation between an annotators age and hate speech across two datasets show that it is a suitable demographic attribute to be used in the experiment of this thesis. The annotators gender is used in combination with age to define the first demographic groups used in the experiment. The gender attribute was included to narrow down the different sub-groups and create more specific groups. The final groups were defined as follows:

- Young men (age groups ranging from 18 to 40)
- Young women (age groups ranging from 18 to 40)
- old men (age groups ranging from 40 to 80)
- old women (age groups ranging from 40 to 80)

The age ranges were chosen based on the age distribution of the dataset combined with the increase in positive hate speech labels around the age of 40. This way the four comparable sub-groups were defined. In Figure 5.7 the distribution of the number of annotations from these subgroups can be seen.

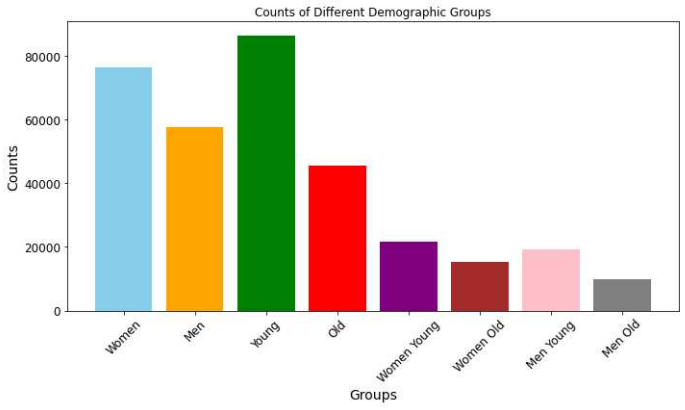


Figure 5.7: Annotations distribution for age-gender sub-groups (dataset 2)

5.1.2. RACE

Next to biases based on age and gender, racial biases have also been shown in literature [63]. Therefore, the second demographic attribute which is taken into consideration is the race of annotators.

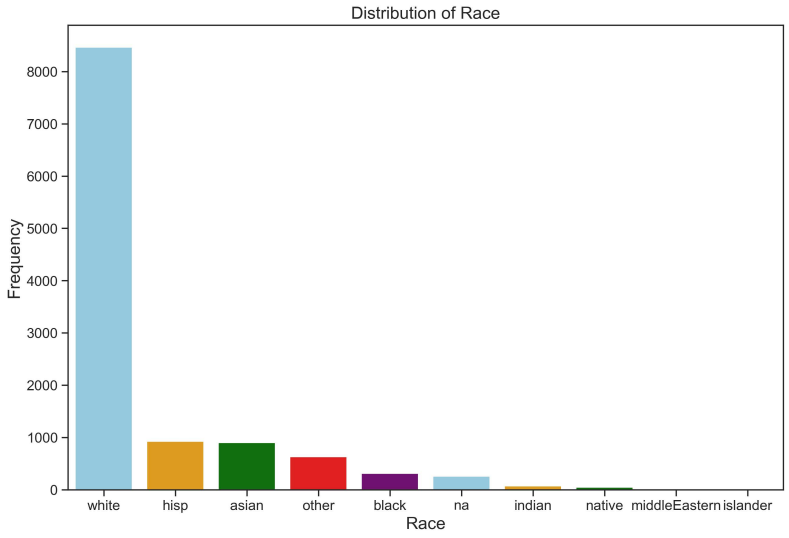


Figure 5.8: Distribution of races in data (dataset 1)

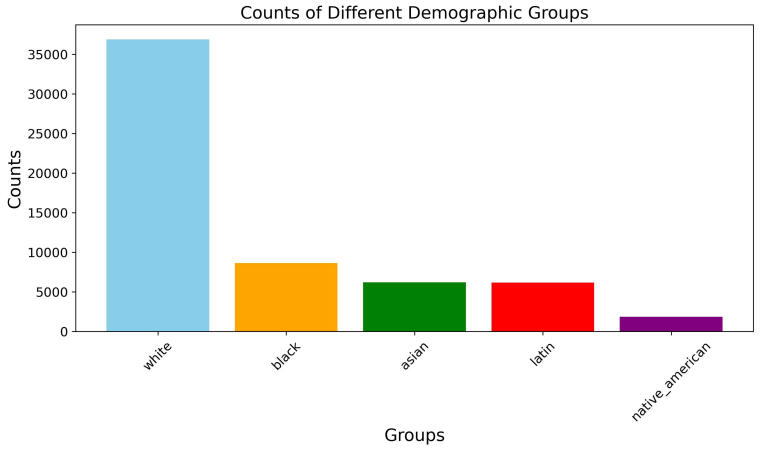
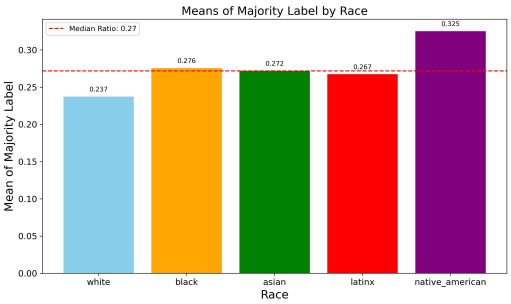


Figure 5.9: Distribution of races in data (dataset 2)

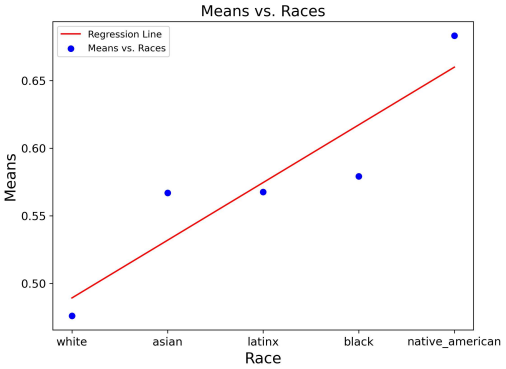
From Figures 5.8 and 5.9 it becomes clear that there is an imbalance in the data concerning the race of annotators. White annotators are more represented in the data than other races in both datasets. This is a first indication that models trained on this data might share this imbalance and thus be biases against certain minorities.

In Figures 5.10a and 5.10b the difference in voting behaviour between groups is shown. It can be seen that the average hate speech label is higher in most groups compared to the white annotator labels. The Pearson correlation (r) of 0.92 and p-value (p) of 0.028 suggests a significantly strong correlation between race and positive hate speech labels.

Based on these results and the in general sensitive nature of racial biases this demographic attribute is included in the evaluation of the experiment of this thesis.



(a) Average hate speech label per racial group



(b) Hate speech ratio for each race – Pearson correlation: **0.92**, p-value: **0.028** (dataset 2)

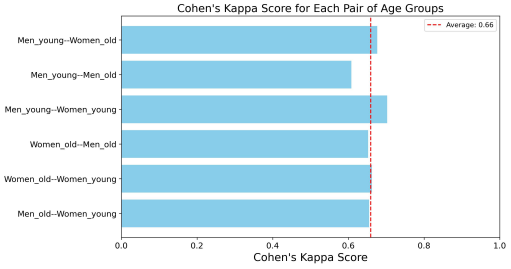
Figure 5.10: Comparison of hate speech metrics across racial groups (dataset 2)

5.2. ANNOTATOR DISAGREEMENT

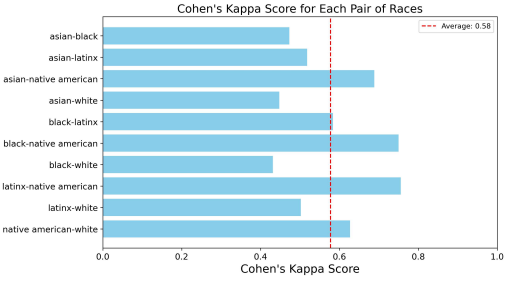
To show the level of disagreement among annotator groups, Cohen's Kappa has been used.

In Figure 5.11 the calculated Cohen's Kappa for each combination of demographic sub-group can be found.
As can be seen in Figure 5.11a, the Cohen's Kappa for the different age/gender groups is relatively equal, with an exception for young men compared to old men. The average agreement is 0.66, which shows moderate agreement for most group. However, since the kappa value is not close to 1.0 there is still some disagreement between groups.

In Figure 5.11b the Cohen's kappa for different racial groups can be found. Compared to the age/gender groups the racial groups show more fluctuation in agreement between different groups. The average lies at 0.58, also showing moderate agreement. Several annotator group combinations show more disagreement, such as black-white, Asian-white and Asian-black. These combinations show only fair/slight agreement.



(a) Cohen's Kappa age groups



(b) Cohen's Kappa racial groups

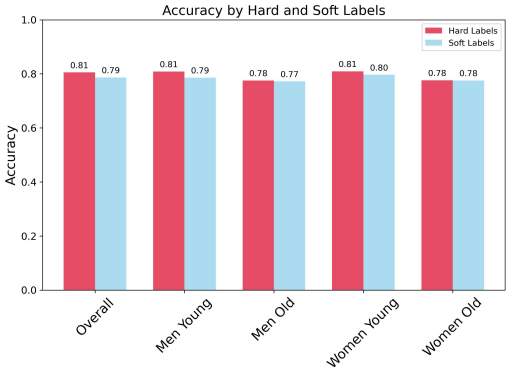
Figure 5.11: Cohen's Kappa for both demographic attributes

5.3. MODEL EVALUATION

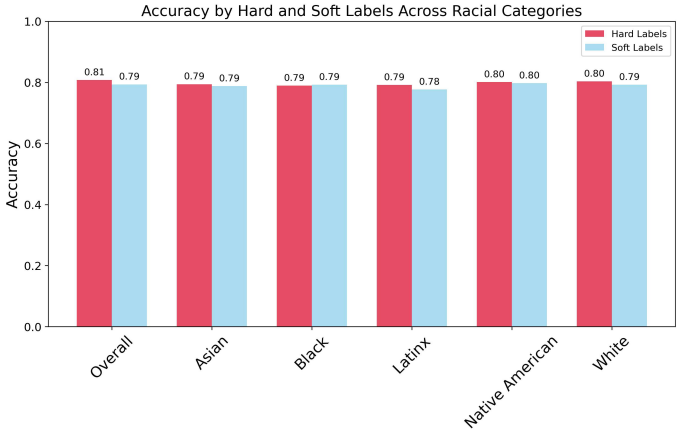
This section provides the results on the performance of the trained models. The results of the data analysis, which provided several sub-groups of the data based on demographic attributes of annotators, are used as subgroups for evaluation of the experiment. First the overall performance of the models is considered, in order to show that the models perform well in terms of accuracy. Next two fairness metrics are evaluated: The Disparate Impact of the models, and the Equal Opportunity. These two metrics provide information on the Fairness of the model, mainly in terms of bias.

5.3.1. OVERALL PERFORMANCE

In Figures 5.12a and 5.12b, the model performance in terms of accuracy on the test set are displayed. The red bars represent the models trained with hard labels, whereas the blue bars represent the models trained with soft labels. As can be seen for both sub-groups (age-gender and racial groups), the hard label models perform similar to slightly better in overall performance for most groups.



(a) Overall performance of models on age-gender groups



(b) Overall performance of models on racial groups

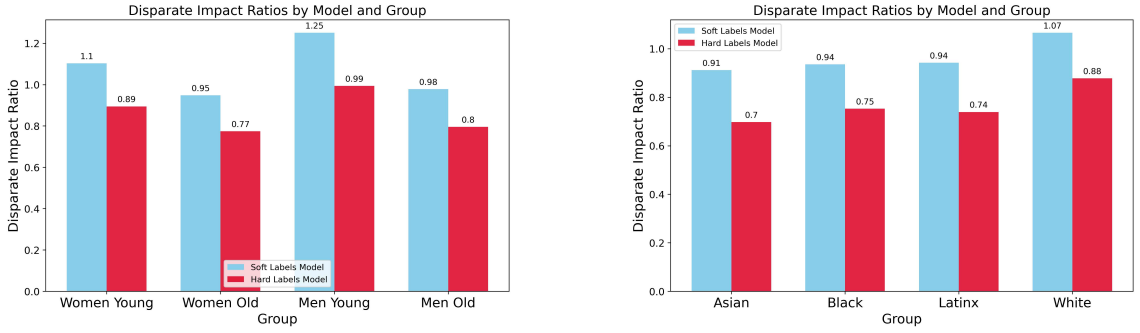
Figure 5.12: Overall performance (accuracy)

5.3.2. FAIRNESS METRICS

The first fairness metric that is used to evaluate the models is the Disparate impact as described in Chapter 4. This metric, as interpreted in this thesis, describes the unwanted favouritism against or towards certain sub-groups.

In the Figure 5.13 The disparate impact ratios are displayed for both hard and soft-labels. The red bars again display the models trained with hard labels and the blue bars the models trained on soft labels. A common rule concerning disparate impact is that for a group to score under 0.8 means that a model is negatively biased towards that group.

For the different age groups displayed in Figure 5.13a it can be seen that the disparate impact ratio's are closer to 1.0 for most of the soft label models compared to hard label models. For the underrepresented groups (Women-Old and Men-Old) the disparate impacts ratio's are well within the accepted range of 0.8. It can also be seen that for the group 'Men-Young' the disparate impact ratio increases to 1.25.
The disparate impact ratios for the different race groups show the same trend in Figure 5.13b. For the hard labels the underrepresented minorities (black, Asian and Latin annotators) have a lower disparate impact (under 0.8), which is lower compared to white annotators (0.9). The soft labels show increases for all these minority groups in the data, meaning that the disparate impact ratio approaches 1.0.

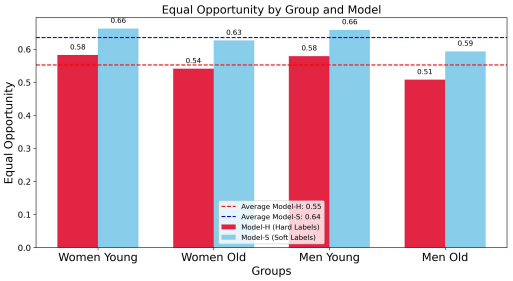


(a) Disparate Impact age groups

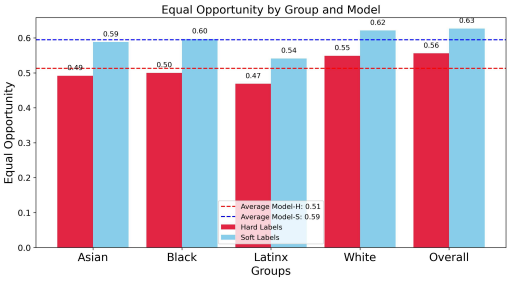
(b) Disparate Impact racial groups

Figure 5.13: Disparate Impact for models trained on both demographic attributes

In Figure 5.14 the Equal Opportunity metric for both sub-groups are displayed. The equal opportunity describes the true positive rate of the models. It can be seen that across sub-groups the Equal Opportunity score is higher for all sub-groups.



(a) Equal Opportunity age groups



(b) Equal Opportunity racial groups

Figure 5.14: Equal Opportunity for models trained on both demographic attributes

6

DISCUSSION

In Chapter 5 the results of the experiments can be found. In this section these results will be discussed and compared with previous studies.

When looking at the overall performance of the models trained with both type of labels, there is little to no difference in predictive accuracy with a slight advantage to hard label models. This results is shown for both demographic groups (age/gender and race). The models demonstrate accuracy's of approximately 80% for both demographic groups, indicating that they perform reasonably well and thus can be used for further evaluation.

However, since accuracy alone is not a good indicator when evaluating a models fairness, other metrics like disparate impact and equal opportunity were obtained. Hard labels are established by majority vote, therefore such labels might be more influenced by an imbalance of represented groups in the data. Soft labels use the probability distribution of all labels, so they keep all votes into account. Based on the bias analysis of the data itself it is expected that the models were biased against the groups which are underrepresented in the data. Based on the definition and nature of the different labels, the use of hard labels could aggravate biases while the use of soft labels could introduce more nuance in the models and thus reduce the impact of biases in the data.

The results on the disparate impact ratio, as defined in Chapter 4, can be found in Section 5.3.2 of Chapter 5. The disparate impacts of the racial groups show results which are in line with the hypothesis. Asian, Black and Latin annotators are underrepresented compared to white annotators. The hard label models show a slight bias against these races. The disparate impact ratio's of these groups lie between 0.7 and 0.75, while the disparate impact ratio of white annotators reaches up to 0.88. The disparate impact ratio's of the same groups are higher when using models trained with soft labels. These values range from 0.91 to 0.94 for the underrepresented racial groups, and is 1.07 for white annotators. The disparate impact ratios are closer

to 1.0 for the soft label models. This is an indication that models which use soft labels in hate speech detection are less susceptible to bias when there are imbalances concerning annotators demographic groups.

The disparate impact ratio's for the different age groups also seems to improve for underrepresented groups in the data ("women old" and "men old"). For the hard label based models the groups "women old" and "men old" perform poorly in terms of disparate impact, with scores of 0.77 and 0.8 respectively. For models based on soft labels these values increase to 0.95 and 0.98. This implicates that the models are more biased against these groups when hard labels are used in the training process compared to the use of soft labels.

While the results regarding disparate impacts on racial groups were clear, the findings related to age groups are less definitive. This is due to the effect of disparate impacts on the more largely represented groups. The change from hard labels to soft labels increases the disparate impact ratio of "women young" from 0.89 to 1.1, and that of "men young" from 0.99 to 1.25. Where a disparate impact ratio drops below 0.8 it can be said that a model is biased against the corresponding group. The opposite can be concluded when a disparate impact ratio becomes larger than 1.2. A model groups which scores higher on disparate impact indicates a bias in favour of such a group. As this is the case with the group containing young men and women, the use of soft labels in this case seem to increase the favour the models have towards these groups. Based on the amount of annotators belonging to these groups in the data, combined with the higher agreement between the groups "men young" and "women young" this increase was not expected. The ratio's for these groups were expected to remain the same or closer to 1.0. For the racial groups this effect is less visible. The disparate impact ratio increases to 1.07 for the group containing white annotators, which is no drastic increase compared to its hard label counterpart. There is no clear explanation for this unwanted increase in the groups containing young men, further research of the data will be necessary to analyse the cause of this increase.

The equal opportunity metric extracted for both sets of sub-groups can be found in Section 5.3.2 of Chapter 5. This metric is designed to show the true positive rate of the models, and is used to evaluate bias in models. An increase can be seen in both set of sub-groups when switching from models based on hard labels to soft labels. In this instance a higher equal opportunity score is positive as it indicates more correctly positive predicted instances. Next to a higher average equal opportunity (0.51 for hard label models, 0.59 for soft label models), the variance for the soft labels is also lower. These insights combined indicate that in general that using soft labels result in a model where different groups have a higher equal opportunity.

In addition to the technical implications of the results, there are also ethical considerations to be addressed. As mentioned in Chapter 1 a definition of algorithmic fairness can be described as the development of an algorithm which has no discriminatory tendencies and is unbiased. The representation of fairness in practice is subjective in itself, and is hard to narrow down to exact measures. This makes an fully unbiased and fair model hard to achieve in practice. Whether a fully unbiased and fair algorithm is viable in practice or not, the results of this thesis

show that using soft labels may aid in achieving an algorithm which works towards this goal. The disagreement metrics of different annotator groups are shown in Section 5.2 of Chapter 5. It can be seen that there is a moderate to high level of disagreement between the studied groups in general. Next to that there are imbalances between groups, mainly presented in the different groups based on gender. It can be argued that in such cases where there is disagreement between groups, and an imbalance in representation of these groups, the use of hard labels can be problematic. Using hard labels which are aggregated by a majority voting principle will result in the majority vote being won by the group with the most annotators in it. While this in principle is sound in a scenario where the distributions among groups are close to equal, it is not in this scenario where there is an imbalance. A soft label modelling approach leads to the minority groups still being included in the labelling process. The imbalance of annotator groups is not removed in its entirety, but its effects are made less severe. Therefore, it can be argued that a soft label modelling approach promotes a more ethically sound and fair way of using annotations from minority groups compared to using a hard label approach.

6.1. COMPARISON WITH PREVIOUS STUDIES

Papers by Basile et al. [33] and Plank et al. [28] provide open issues regarding perspective and ground truthing in NLP models. One of the themes discussed as open issues and potential field for future research is concerning applications of perspectivist ML in fair AI and bias detection. Combined with these suggestions and the outcomes of the literature review in Chapter 2 the direction for this thesis was chosen. The implications of this thesis mostly relate to whether biases can be detected based on annotators demographic attributes and disagreements between annotators, and how a multi-label approach relates to these biases. Previous studies on the use of soft labels in the training process of NLP models show that soft labels can introduce more ambiguity to labels [13], and also improve performance for certain NLI tasks [14]. Other papers which advocate for more “perspectivism” [33] in ML models, or another approach to finding a ground truth [28] also suggest the use of soft labels as alternative to fusing labels based on majority vote. Existing literature primarily assesses the impact of soft labels on performance metrics such as accuracy and precision [12–14]. This thesis contributes by facilitating the detection of biases in data resulting from annotator disagreements and by introducing a bias evaluation of the more nuanced soft labels in the training process, as opposed to the use of hard labels.

The contributions from empirical results of this thesis can be summarized in two-fold:

First of all, it is shown that it is possible to find potential biases within the data induced by groups of raters, it being caused by an imbalance of this groups representation and a disagreement between the annotators from these groups.
Secondly, these groups are used to evaluate the effects of using soft labels on model fairness, in terms of disparate impact and equal opportunity. The results indicate a positive effect when using soft labels instead of hard labels. These results, both corresponding to answers to the two

sub-research questions respectively, are used to answer the main research question. Analyzing labels from different groups of annotator as is done in this thesis provides help with detecting potential biases. Such insights can help detecting in which areas an ML model is susceptible to bias, which is an essential step to increasing Fairness in AI. Secondly, the multi-label approach based on soft labels as described in this research shows improvements in terms of fairness metrics (disparate impact ratio and equal opportunity). This approach indicates promising results in this thesis for increasing model fairness, and could potentially be applied in NLP applications focused on hate speech or other type of NLP applications with a labeling process which is prone to subjectivity. Therefore, this thesis provides an answer to its research question by considering methods to detect biases in data based on demographic attributes, and assess the effects of soft labels on fairness in AI.

7

CONCLUSION

7.1. SUMMARY OF MAIN FINDINGS

This thesis aimed at enhancing fairness in NLP models focused on hate speech by using soft labels in the learning process of ML models. The main research question and the sub-research questions were defined as follows:

Main research question:

How can a soft label modelling approach, combined with bias detection methods enhance fairness in hate speech detection models?

Sub-research question:

- Can biases be detected in annotators behavior based on their demographic information (age, gender, politics, race)?
- How effective are soft label based algorithms in mitigating biases in hate speech detection compared to hard label approaches?

This main research goal stems from challenges in the field of NLP related to annotator disagreement and how to handle these challenges. The first part of the main research goal was to evaluate whether a bias could be found based on annotators demographic information. The execution of this part of the thesis resulted in two sets of subgroups defined on several demographic attributes of the annotators. Age, gender, and race have been examined as demographic attributes of annotators, along with their influence on the labeling behavior of these groups. Defined subgroups based on these attributes showed significant difference in labelling behaviour towards hate speech, and were thus used for the bias evaluation section of this thesis. For the second part of the thesis a soft label approach was implemented in the training process of the models. This soft label approach was compared to the current benchmark method of using the hard labels (majority voting) in the training process of the ML models. The resulting models on both type of labels were evaluated in terms of overall performance, and fairness metrics such

as the disparate impact ratio and equal opportunity. The results show that although in overall predictive performance the models trained on hard labels are slightly better, improvements can be seen when using soft labels compared to hard labels in terms of disparate impact and equal opportunity.

The main research question was answered by addressing the sub-research questions. These results address questions surrounding the detection of biases based on annotator demographics and the effectiveness of soft labels in mitigating these biases compared to traditional hard label approaches.

> *Can biases be detected in annotators behavior based on their demographic information (age, gender, politics, race)*

The findings of this study reveal significant insights into annotator labelling behaviour. By examining annotator demographics, such as age, gender and race, we identified substantial differences in labels across different annotator groups that contribute to biases in hate speech detection models. Analysis on sub groups based on gender, age and race showed that these groups exhibit distinct labeling patters. Increases in hate speech labels were found in annotator groups containing older annotators compared to younger annotators. Differences in the ratio of hate speech labels were also found across different racial groups of annotators in the data. The results indicate that demographic attributes of annotators can be used to detect distinct patterns in annotations and indicate potential annotator biases in the data.

> *How effective are soft label based algorithms in mitigating biases in hate speech detection compared to hard label approaches?*

The implementation of soft labels in the training process revealed notable improvements in terms of fairness. Such improvements were shown in both the disparate impact ratio and equal opportunity. The effects on fairness were most clearly shown for subgroups based on race, hypothetically caused by a notable bias and higher disagreement between racial groups compared to the age-gender subgroups. Although the use of soft labels did not show superior performance in terms of predictive accuracy, it is indicated that soft labels helped mitigate biases against underrepresented groups in the data better compared to the models trained with hard labels.

This thesis contributes to the growing body of literature on fairness in AI by highlighting the potential soft labels provide to enhance fairness in NLP models on hate speech. Next to that it contributes by analyzing annotator disagreement and its relation to potential biases based on annotators demographic attributes.

7.2. LIMITATIONS AND FUTURE RESEARCH

While this thesis provides valuable insights into enhancing fairness and bias evaluation in multi-label NLP applications, it is important to acknowledge limitations of the study. The limitations are inherent to the data and methods used. Addressing these limitations aid with defining directions for future research opportunities for this field of research.

Limitations regarding the data and the collection of the data include the the size of the different sub-groups defined, and their to which extent they accurately represent the broader population. The biases are analyzed by taking into account sub-groups which are not well represented in the data. Next to that these groups should have a difference in opinion compared to the larger groups in the data. Although this reasoning is logical in terms of showing bias in data, the small amount of annotators belonging to these minority groups can also be considered a limitation. The demographic sub-groups based on age and gender are more balanced in terms of group sizes compared to the sub-groups based on race. In these race groups white annotators clearly form majority of the dataset. It could be argued that drawing conclusions concerning labelling behaviour of these groups is affected by the sizes of these groups. Next to that it speaks to the reliability of such conclusions.

Another limitation, that stems from the just mentioned imbalance in the data is that not every tweet has the same amount of annotations. This results in tweets being annotated by different distributions of annotator demographics. Although this practice can in fact aggregate biases, it is a clear limitation in drawing conclusions concerning annotator agreement between groups since only a part of the tweets are annotated by annotators from all included sub-groups.

The CRISP-ML(Q) research cycle as described in Chapter 3 includes evaluation and deployment phases. For these phases to be successful there should be a sound understanding of the effects on using soft labels on an application in practice. Although the results in this thesis show that the use of soft labels compared to the majority vote indicates an increased level of fairness, the overall performance decreases in some instances. This is a limitation in which there seems to be a trade-off between overall performance and fairness in AI.

The final limitation which is important to address is related to the evaluation of the soft label based models to those based on hard labels. For these two type of models and labels to be compared the labels there had to be some common standard to which the predictions should be compared. For the hard labels this was already decided by majority vote, but for the soft labels this standard was more complicated. It was possible to compute the soft label predictions to the actual distribution of hate speech votes in terms of error margin. This however would make it harder to compare the results to the hard label models. Therefore it was decided to aggregate the soft labels in the same manner as the hard labels, but based on the actual soft label distributions. This resulted in comparable results, however it raises a limitation of wasting some of the nuance introduced by soft label predictions. A different way of evaluating soft

labels could perhaps provide interesting insights concerning its effects on fairness in AI.

Based on the limitations described above, several promising areas for future research have been defined:

- Future research could focus on expanding and diversifying annotator samples to ensure that sub-groups are well-represented. Such an expansion could enhance the quality of a dataset. This would ultimately lead to a higher level of reliability of the conclusions drawn about labeling behaviour and biases.
- Another domain which could be the focus area of future research is the refinement of soft label evaluation metrics. Alternative methods for evaluating soft label models could reveal new or more nuanced insights into their performance, and relation to fairness.
- Applying soft label based approaches to other NLP tasks beyond hate speech detection, to assess its effectiveness in different domains could be an interesting field for future research. The tasks which might have the most relevance are tasks of which the annotation process has a highly subjective nature, such as sentiment analysis or fake-news detection. Investigating the behaviour of different annotator sub-groups in these contexts might reveal new insights towards bias detection and fairness in AI for various types of NLP models.
- A direct expansion of this research could consider other demographic attributes to be included in the analysis. Attributes which have not been considered in these datasets, or the ones available which have not been elaborated on yet might provide valuable insights in terms of bias analysis. Further research could be conducted on exploring these attributes, and using them to evaluate NLP models on bias and other fairness concepts.
- Finally the practical employment of using soft labels and its impact on various performance indicators could be researched in future studies. As mentioned in this thesis there is an indication that there might be a trade of between an increase in fairness and a decrease in accuracy. Conducting field studies to evaluate the actual effects of soft label models with different configurations in practice can reveal interesting insights. This could be combined with the development of guidelines for practical applications to effectively use soft labels and bias detection methods in their NLP pipelines.

In conclusion, this thesis demonstrates that potential biases in data can be identified based on annotators' demographic information, and that a multi-annotator approach involving soft labels can mitigate these biases, thereby indicating to enhance fairness in hate speech detection models.

REFERENCES

[1] S. R. Chinta, Crisp-ml(q): Methodology in machine learning, 2024. URL: https://medium.com/@chintasandhyarani973/crisp-ml-q-methodology-in-machine-learning-889bf799b739.

[2] IBM, What is natural language processing?, n.d. URL: https://www.ibm.com/topics/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can.

[3] F. Nakano, R. Cerri, C. Vens. Active learning for hierarchical multi-label classification. Data Mining and Knowledge Discovery 34 (2020) 1496–1530. doi:10.1007/s10618-020-00704-w.

[4] M. I. Jordan, T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science 349 (2015) 255–260. doi:doi:10.1126/science.aaa8415.

[5] IBM, What is supervised learning?, n.d. URL: https://www.ibm.com/topics/supervised-learning.

[6] T. C. Weerasooriya, T. Liu, C. M. Homan. Neighborhood-based pooling for population-level label distribution learning. CoRR abs/2003.07406 (2020). URL: https://arxiv.org/abs/2003.07406. arXiv:2003.07406.

[7] S. Wang, Z. Wang, W. Che, T. Liu, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1813–1822. URL: https://aclanthology.org/2020.emnlp-main.142. doi:10.18653/v1/2020.emnlp-main.142.

[8] J. Zhang, V. S. Sheng, J. Wu. Crowdsourced label aggregation using bilayer collaborative clustering. IEEE Transactions on Neural Networks and Learning Systems 30 (2019) 3172–3185. doi:10.1109/TNNLS.2018.2890148.

[9] L. Yin, Y. Liu, W. Zhang, Y. Yu. Truth inference with a deep clustering-based aggregation model. IEEE Access 8 (2020) 16662–16675. doi:10.1109/ACCESS.2020.2964484.

[10] S. Ahuja, G. Dubey, in: 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), pp. 1–5. doi:10.1109/TEL-NET.2017.8343568.

[11] G. Li, F. Liu, in: 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering, pp. 331–337. doi:10.1109/ISKE.2010.5680859.

[12] A. Mostafazadeh Davani, M. Díaz, V. Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics 10 (2022) 92–110. URL: https://aclanthology.org/2022.tacl-1.6. doi:10.1162/tacl_a_00449.

[13] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, M. Poesio, 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, Association for Computational Linguistics, Online, 2021, pp. 2591–2597. URL: https://aclanthology.org/2021.naacl-main.204. doi:10.18653/v1/2021.naacl-main.204.

[14] J. P. Lalor, H. Wu, H. Yu. Soft label memorization-generalization for natural language inference (2019). arXiv:1702.08563.

[15] S. Zhang, C. Gong, E. Choi. Learning with different amounts of annotation: From zero to many labels (2021). arXiv:2109.04408.

[16] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. URL: https://aclanthology.org/2021.semeval-1.41. doi:10.18653/v1/2021.semeval-1.41.

[17] K. R. Remya, J. Ramya, in: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). doi:10.1109/iccicct.2014.6993144.

[18] R. Snow, B. O'Connor, D. Jurafsky, A. Ng, in: M. Lapata, H. T. Ng (Eds.), Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 254–263. URL: https://aclanthology.org/D08-1027.

[19] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, S. H. Malik. Detecting twitter hate speech in covid-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights 2 (2022) 100120. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000635. doi:https://doi.org/10.1016/j.jjimei.2022.100120.

[20] L. Hong, B. D. Davison, in: Proceedings of the First Workshop on Social Media Analytics, SOMA '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 80–88. URL: https://doi.org/10.1145/1964858.1964870. doi:10.1145/1964858.1964870.

[21] Nazmine, K. Manan, H. K. Tareen, S. Noreen, M. Tariq. Hate speech and social media: A systematic review. Turkish Online Journal of Qualitative Inquiry 12 (2021) 5285–5294.

[22] R. T. Mutanga, O. O. Olugbara, N. Naicker. Bibliometric analysis of deep learning for social media hate speech detection. Journal of Information Systems and Informatics (2023). URL: https://api.semanticscholar.org/CorpusID:261984627.

[23] P. A. Atmajaya, F. I. Amorokhman, M. D. Prasetya, A. F. Ihsan, D. Junaedi. Ite law enforcement support through detection tools of fake news, hate speech, and insults in digital media. 2022 5th International Conference on Information and Communications Technology (ICOIACT) (2022) 452–456. URL: https://api.semanticscholar.org/CorpusID:254615994.

[24] B. Memarian, T. Doleck. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review. Computers and Education: Artificial Intelligence 5 (2023) 100152. URL: https://www.sciencedirect.com/science/article/pii/S2666920X23000310. doi:https://doi.org/10.1016/j.caeai.2023.100152.

[25] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, K.-R. Mueller, Towards crisp-ml(q): A machine learning process model with quality assurance methodology, 2021. arXiv:2003.05155.

[26] C. Wohlin, M. Kalinowski, K. Romero Felizardo, E. Mendes. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. Information and Software Technology 147 (2022) 106908. URL: https://www.sciencedirect.com/science/article/pii/S0950584922000659. doi:https://doi.org/10.1016/j.infsof.2022.106908.

[27] X. S. Lu, M. Zhou, H. Liu, L. Qi. A comparative study on two ground truth inference algorithms based on manually labeled social media data (2019) 436–441. doi:10.1109/ICNSC.2019.8743287.

[28] B. Plank. The 'problem' of human label variation: On ground truth in data, modeling and evaluation (2022). arXiv:2211.02570.

[29] B. Plank, D. Hovy, A. Søgaard, in: Conference of the European Chapter of the Association for Computational Linguistics. URL: https://api.semanticscholar.org/CorpusID:9198407.

[30] G. Rizos, B. Schuller. Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. Information Processing and Management of Uncertainty in Knowledge-Based Systems 1237 (2020) 42 – 55. URL: https://api.semanticscholar.org/CorpusID:219323752.

[31] T. Cohn, L. Specia, in: H. Schuetze, P. Fung, M. Poesio (Eds.), Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 32–42. URL: https://aclanthology.org/P13-1004.

[32] Z. Zhang, S. Chapman, F. Ciravegna. A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality (2010) 301–315.

[33] F. Cabitza, A. Campagner, V. Basile. Toward a perspectivist turn in ground truthing for predictive computing. Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 6860–6868. URL: http://dx.doi.org/10.1609/aaai.v37i6.25840. doi:10.1609/aaai.v37i6.25840.

[34] M. Popović. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output (2021) 234–243. URL: https://aclanthology.org/2021.conll-1.18. doi:10.18653/v1/2021.conll-1.18.

[35] W. Mieleszczenko-Kowszewicz, K. Kanclerz, J. Bielaniewicz, M. Oleksy, M. Gruza, S. Wozniak, E. Dziecіol, P. Kazienko, J. Kocoń, in: NLPerspectives@ECAI. URL: https://api.semanticscholar.org/CorpusID:265467579.

[36] S. Larimore, I. Kennedy, B. Haskett, A. Arseniev-Koehler, in: L.-W. Ku, C.-T. Li (Eds.), Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Online, 2021, pp. 81–90. URL: https://aclanthology.org/2021.socialnlp-1.7. doi:10.18653/v1/2021.socialnlp-1.7.

[37] J. Ramírez, M. Baez, F. Casati, B. Benatallah. Understanding the impact of text highlighting in crowdsourcing tasks (2019). arXiv:1909.02780.

[38] D. Braun. I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. Artif Intell Law (2023). URL: https://doi.org/10.1007/s10506-023-09369-4. doi:10.1007/s10506-023-09369-4, accepted 12 June 2023, Published 27 June 2023.

[39] E. Troiano, S. Padó, R. Klinger. Emotion ratings: How intensity, annotation confidence and agreements are entangled. CoRR abs/2103.01667 (2021). URL: https://arxiv.org/abs/2103.01667. arXiv:2103.01667.

[40] B. Beigman Klebanov, E. Beigman, D. Diermeier, in: R. Artstein, G. Boleda, F. Keller, S. Schulte im Walde (Eds.), Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, Coling 2008 Organizing Committee, Manchester, UK, 2008, pp. 2–7. URL: https://aclanthology.org/W08-1202.

[41] W. Wang, Z. Wang, M. Wang, H. Li, Z. Wang. Importance filtered soft label-based deep adaptation network. Knowledge-Based Systems 265 (2023) 110397. URL: https://www.sciencedirect.com/science/article/pii/S0950705123001478. doi:https://doi.org/10.1016/j.knosys.2023.110397.

[42] T. Marquet, E. Oswald, in: J. Zhou, L. Batina, Z. Li, J. Lin, E. Losiouk, S. Majumdar, D. Mashima, W. Meng, S. Picek, M. A. Rahman, J. Shao, M. Shimaoka, E. Soremekun, C. Su, J. S. Teh, A. Udovenko, C. Wang, L. Zhang, Y. Zhauniarovich (Eds.), Applied Cryptography and Network Security Workshops, Springer Nature Switzerland, Cham, 2023, pp. 121–138.

[43] V. Sharmanska, D. Hernández-Lobato, J. M. Hernández-Lobato, N. Quadrianto, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2194–2202. doi:10.1109/CVPR.2016.241.

[44] A. Dumitrache, L. Aroyo, C. Welty. Crowdsourcing ground truth for medical relation extraction. ACM Trans. Interact. Intell. Syst. 8 (2018). URL: https://doi.org/10.1145/3152889. doi:10.1145/3152889.

[45] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Alianne-jadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 316–342. URL: https://doi.org/10.1007/978-3-031-42448-9_23. doi:10.1007/978-3-031-42448-9_23.

[46] R. Wan, J. Kim, D. Kang. Everyone's voice matters: Quantifying annotation disagreement using demographic information (2023). arXiv:2301.05036.

[47] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, M. S. Bernstein, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445423. doi:10.1145/3411764.3445423.

[48] I. Kolyshkina, S. Simoff. Interpretability of machine learning solutions in public healthcare: The crisp-ml approach. Frontiers in Big Data 4 (2021). URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.660206. doi:10.3389/fdata.2021.660206.

[49] E. Ohata, C. Mattos, P. Rêgo, pp. 133–144. doi:10.5753/semish.2024.2438.

[50] J. J. Magdaong, A. Culaba, A. Ubando, N. Lopez. Generating synthetic building electrical load profiles using machine learning based on the crisp-ml(q) framework. IOP Conference Series: Earth and Environmental Science 1372 (2024) 012082. doi:10.1088/1755-1315/1372/1/012082.

[51] J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1960) 37–46. URL: https://doi.org/10.1177/001316446002000104. doi:10.1177/001316446002000104. arXiv:https://doi.org/10.1177/001316446002000104.

[52] P. Schober, C. Boer, L. A. Schwarte. Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia 126 (2018) 1763–1768. doi:10.1213/ANE.0000000000002864.

[53] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, KDD '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 259–268. URL: https://doi.org/10.1145/2783258.2783311. doi:10.1145/2783258.2783311.

[54] A. Guterres, United nations strategy and plan of action on hate speech, United Nations, 2020. URL: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf.

[55] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678. URL: https://aclanthology.org/P19-1163. doi:10.18653/v1/P19-1163.

[56] C. J. Kennedy, G. Bacon, A. Sahn, C. von Vacano, Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application, 2020. URL: https://arxiv.org/abs/2009.10277. arXiv:2009.10277.

[57] I. Zeroual, A. Lakhouaja, pp. 1–6. doi:10.1109/ISACV.2017.8054932.

[58] D. Khyani, S. B S. An interpretation of lemmatization and stemming in natural language processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology 22 (2021) 350–357.

[59] R. Friedman. Tokenization in the theory of knowledge. Encyclopedia (2023). doi:10.3390/encyclopedia3010024.

[60] M. M. Krell, M. Kosec, S. P. Perez, A. Fitzgibbon, Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance, 2022. arXiv:2107.02027.

[61] S. Hochreiter, J. Schmidhuber. Long short-term memory. Neural computation 9 (1997) 1735–80. doi:10.1162/neco.1997.9.8.1735.

[62] A. Vabalas, E. Gowen, E. Poliakoff, A. Casson. Machine learning algorithm validation with a limited sample size. PLoS ONE 14 (2019). doi:10.1371/journal.pone.0224365.

[63] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678. URL: https://aclanthology.org/P19-1163. doi:10.18653/v1/P19-1163.

[64] A. Dumitrache, O. Inel, B. Timmermans, C. Martinez-Ortiz, R. Sips, L. Aroyo, C. Welty. Empirical methodology for crowdsourcing ground truth. CoRR abs/1809.08888 (2018). URL: http://arxiv.org/abs/1809.08888. arXiv:1809.08888.

[65] A. Dumitrache, O. Inel, L. Aroyo, B. Timmermans, C. Welty. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. CoRR abs/1808.06080 (2018). URL: http://arxiv.org/abs/1808.06080. arXiv:1808.06080.

[66] A. Dumitrache, in: F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux, A. Zimmermann (Eds.), The Semantic Web. Latest Advances and New Domains, Springer International Publishing, Cham, 2015, pp. 701–710.

A

APPENDIX A: SUMMARY OF SYSTEMATIC LITERATURE REVIEW

Table A.1: Table reporting all the articles that have been examined to conduct the literature review and highlighting their main features

Author	Main purpose	Algorithms	Evaluation	Setting	Theme
Lu et al., 2019 [27]	Comparing two algorithms for determining the ground truth on manually labeled social media data (PLAT and Majority voting)	Majority voting and Positive label frequency threshold (PLAT)	Accuracy, z-test, McNemar test, Average execution time, p-value	Social media data (short texts)	OTHER
Zhang et al., 2010 [32]	Proposing an alternative approach to document annotation by firstly studying annotators' suitability based on the types of information to be annotated, then identifying and isolating the most inconsistent annotators who tend to cause the majority of discrepancies in a task. Finally distributing annotation workload among the most suitable annotator	Matching best-fit-annotators to best-fit classes	Inter annotator agreement F-measure	NER tasks in the domain of archaeology	DA
Dumitrache et al., 2018c [64]	Proposing an empirically derived methodology for efficiently gathering of ground truth data in a diverse set of use cases covering a variety of domains and annotation tasks. This would increase the stability of crowd results, the measuring quality in open-ended tasks. Finally the paper improves the semantics of ambiguity	performing experiments on four crowdsourcing tasks (2 closed tasks and 2 open-ended tasks) to compare Majority vote, Single annotator and Expert annotator.	CrowdTruth metrics (inter-annotator agreement metrics): wwa, wma, na, UAS. Precision, Recall, F1 score, Accuracy	Crowdsourcing tasks on Medical relation extraction, twitter event identification, news event extraction and sound interpretation	OTHER

Troiano et al., 2021 [39]	Investigation of the relationship between three human judgements:presence of emotions, their intensity and the confidence of the annotation decision. This aims at understanidng in what cases annotators differ regarding the judgement that an emotion is expressed	pre-trained BERT.	Cohen's k (1960) and Fleiss' k (1971). Counts of emotion/neutral items.	Corpus of Contemporary American English (neutral vs emotional sentences). Datasets like self-reports, tweets and newspapers.	DA
Dumitrache et al., 2018b [65]	Presenting ongoing work into the CrowdTruth metrics, that capture annotator diagreement in crowdsourcing. The goal of the metrics is to capture the degreee of ambiguity in each of these three components	CrowdTruth method	Media unit quality, worker quality and annotation quality	-	OTHER
Popović, 2021 [34]	Analyzing inter-annotator disagreement in human evaluation of machine translation output.	-	Nature of disagreement, overlap of words perceived as errors by two annotators	Translations of English user reviews to Croatian and serbian (Qrev). TED talks translated into German.	DA
Mielenszczenko-Kowszewicz et al., 2023 [35]	Analyzing how personalized perception of texts is affected by individual human profile and bias.	-	Human Bias measure, Big five personality traits, inter-annotator agreement measures, Cohen's kappa, Cohen's kappa on binarized annotations, Kendall Tau rank correlation coefficient	Internet forums	DA
Wan et al., 2023 [46]	Examining whether the text of the task and annotators' demographic background information can be used to estimate the level of disagreement among annotators	pre-trained BERT, (RoBERTa)	MSE, F1	SBIC, SChem101, scruples-dilemmas, dyna-sentiment, sikipedia politeness	DA
Larimore et al., 2021 [36]	Investigating how annotator perceptions of racism in tweets vay by annotator racial identity and two text features of the tweets: relevant keywords and latent topics	291 human annotators applied to 4188 unique tweets.	descriptive summaries of data + statistical evaluation of three distinct models	twitter data	OTHER
Dumitrache et al., 2018a [44]	Proposing the CrowdTruth method for crowdsourcing training data for machine learning in medical relation extraction	CrowdTruth method	Precision, Recall, F-measure	semantic data in medical domain	OTHER
Dumitrache, 2015 [66]	Investigate whether disagreement-aware crowdsourcing is a scalable approach to gather semantic annotation across various tasks and domains	Defining the crowdsourcing setup, experimental data collection, and evaluating both the setup and results, building on CrowdTruth framework	Evaluation plan: cross-validation and comparison of accuracy and F1	Medical domain, open domain (twitter)	OTHER
Mostafazadeh Davani et al., 2022 [12]	Introducing a multi-annotator architecture to preserve and model the internal consistency in each annotators' labels as well as their systematic disagreements with other annotators. Obtain an interpretable way to estimate model uncertainty that better correlates with annotator disagreements than traditional uncertainty estimates	BERT on three different multi-annotator architectures: ensemble, multi-label and multi-task	Precision, Recall, F1, model uncertainty (variance of annotations), error analysis	Gab Hate Corpus, Go emotions dataset (subjective annotated datasets, with per annotator labels)	MULTI

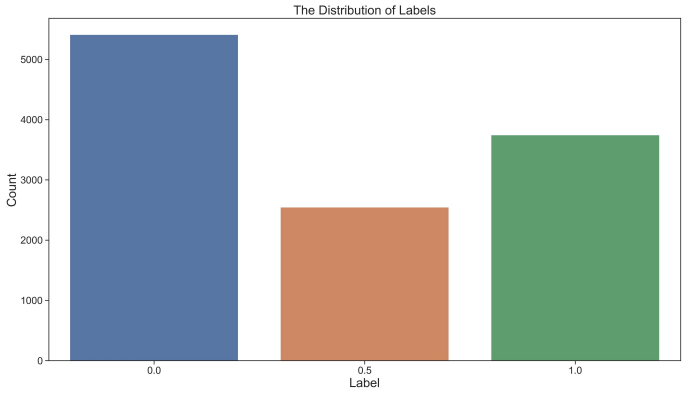
Weerasooriya et al., 2020 [6]	Proposing an algorithmic framework and new statistical tests for PLDL that account for sampling size. Propose a new approache for label sharing, called neighborhood-based pooling (NBP)	Comparing models: FMM, GMM, K-Means, LDA	KL divergence, Chebyshev distance, Euclidean distance, Canberra metric	Jobs dataset, Natural scenes dataset, Facial expression dataset	CLUSTER
Gordon et al., 2021 [47]	Introducing a transformation that more closely aligns ML classification metrics with the values and methods of user-facing performance measures	Applying ML tasks to the 'disagreement deconvolution'	ROC AUC, Precision, Recall, Accuracy, Pflip estimator, mean Pflip	Classic ML task dataset (such as Jigsaw dataset, CIFAR-10h)	OTHER
Beigman Klebanov et al., 2008 [40]	An application of Beigman Klebanov and Sahmir's (2006) methodology for analyzing annoation data to metaphor identification annotations. Finding the difference between attention slips and genuine disagreements	-	Accuracy, reliability	Brithish press articles (151 articles)	DA
Sharmanska et all, 2016 [43]	Defining disagreeements among annotators as priviliged information about the data instance. Proposing a framework to incorporate annotation disagreements into the classifiers (LUPI)	SVM-based methods, GP-based methods. Comparing models	Classification accuracy as performance measure	Images (computer vision dataset)	OTHER
Ramírez et al., 2019 [37]	Investigate if and under what conditions highlighting selected parts of the text can improve classification cost and/or accuracy, and in how it affects the process and outcome of the human intelligence tasks	Crowdsourcing experiments running over different datasets, BERT-Base	Decision time, Worker accuracy	systematic literature reviews, amazon reviews	OTHER
Plank, 2022 [28]	Arguing the necessity of taking human label variation into account in the three core aspects of an NLP pipeline: data, modeling and evaluation. Presenting and working out future research for these core aspects concerning human label variation	Comparing current literature	-	-	MULTI
Snow et al., 2008 [18]	Comparing non-expert annotations (mechanical Turk) to the existing gold standard labels provided by expert labelers on five tasks: Affect recognition, word similarity, recognizing tesxtual entailment, event temportal ordering and word sense disambiguation	Amazon Mechanical Turk (non-expert annotator platform)	Inter-annotator agreement (ITA)	Affect recognition, word similarity, recognizing tesxtual entailment, event temportal ordering and word sense disambiguation.	OTHER
Cabitza et al., 2023 [33]	Describing and advocating for a 'data perspectivism' paradigm, which moves away from traditional gold standard datasets, towards the adoption of methods that integrate the opinions and perspectives of human subjects involved in the knowledge representation step of ML processes	Comparing current literature	comparing ML training methods based on perspectivists ground truths	NLP	MULTI
Plank et al., 2014 [29]	Using small samples of doubly-annotated part-of-speech (POS) data for trwitter to estimate annotation reliability and show how those metrics of likely inter-annotator agreement can be impleemented in the loss function of POS taggers. (incorporating uncertainty exhibited by annotators)	Online structured perceptron with drop-out, with inter-annotator F1-scores as loss function or the confusion probability between annotators.	F1-scores between annotators on individual POS, and tag confucsion probabilities. Further focus on downstream evaluation	POS, Twitter (500 sampled tweets).	MULTI

Cohn & Specia., 2013 [31]	Modelling the task of predicting the quality of sentence transloations using datasets that have been annotated by several judges with different levels of expertise and reliability	Multi-task Gaussian Processes, baseline model (SVM)	Predivctive accuracy, using two measures: mean absolute error (MAE) and Root mean square error (RMSE)	Quality estimation datatsets (QE)	OTHER
Fornaciari et al,. 2021 [13]	Showing that Multi-task learning (MTL) models trained with soft labels consistently outperform Single-Task Learning (STL) networks. Evaluating the use of different loss functions for soft labels	Different types of neural networks, for each a STL model and an MTL model. Context bi-RNN, sequence bi-RNN (KL/Cross-entropy).	Accuracy, F1 (cross-validation 5-fold)	POS tagging, morphological stemming	MULTI
Lalor et al., 2019 [14]	Proposing the soft label memorization-generalization (SLMG) framework for incorporating soft labels in machine learning training, estimating soft label distributions for NLI and demonstrating that soft labels can encode ambiguity in training data that can improve model generalization in terms of test set accuracy	three deep learning models: LSTM RNN, memory-augmented LSTM network and hierarchical network	Accuracy, compared to baseline model, KL-divergence	SNLI dataset (Natural language inference)	MULTI
Zhang et al., 2021 [15]	Presenting an in-depth study comparing models trained with multi label data and models with single lable data	RoBERTa, bidirectional LSTM	classification accuracy, computed twice, once against aggregated gold labels in the original dataset and once against the newly labled dataset. Macro-avaraged precision, recall, F1, MMR	NLI, fine-grained entity typing	MULTI
Uma et al., 2021 [16]	SemEval-2021 provides a unified testing framework for methods for learning from data containing multiple annotations	bi-LSTM (POS tagging), Gaussian Process preference learning (GPPL, Humour dataset), fine-tuned BERT	F1 and comparing soft labels (produced by models probability distribution) to that of the full distribution produced by annotators using cross-entropy (Hard score vs Soft score)	Datasets preserving disagreements (Gimpel POS, PD IS, Humour datasets, LabelMe, Cifar-10h corpus	MULTI
Wang et al., 2020 [7]	Exploring unspervised learning paradigm which can potentially work with unlabeled text corpora that are cheaper and easier to obtain	Self-supervised learning method to train a weak disfluency detection model. ELECTRA-Base model with 110M hidden units, 12 heads, 12 hidden layers	Token-based precision, recall, F1	Automatic speech recognition, English Switchboard (SEBD) for disfluency detection. CallHome, SCOTUS ,FCIC	CLUSTER
Zhang et al., 2019 [8]	Proposing a bilayer collaborative clustering (BLCC) method for the label aggregation in crowdsourcing	K-means clustering, BLCC	Accuracy against eight state-of-the-art algorithms	Computer vision (CV)	CLUSTER

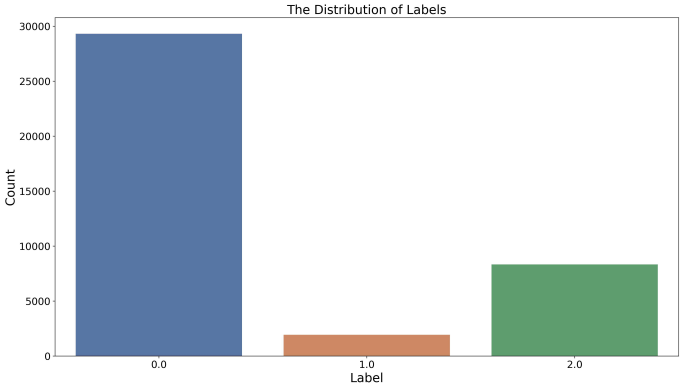
Yin et al., 2020 [9]	Proposing a deep clustering-based aggregation model (DCAM) to overcome the shortcomings of noiy real-world data labels. DCAM introduces clustering for object features to form fine-grained clusters, where objects in the same cluster are supposed to have similar labels	DCAM	Accuracy	Computer vision (CV)	CLUSTER
Rizos & Schuller, 2020 [30]	presenting an overview of approaches utilised to aaddress the issue of ground-truth uncertainty due to subjectivity, a discussion of current state-of-the-art methods, challenges and an outline of promosing future directions	Comparing current literature	-	NLP	MULTI
Ahuja & Dubey, 2017 [10]	Conducting a study on the use of clustering techniques in efficiently distinguishing tweets based on their sentiment scores and characteristics	K-means	Sentiment scores for each tweet, clustering quality	NLP, sentiment analysis	CLUSTER
Li & Liu, 2010 [11]	Introducing a clustering-based sentiment analysis approach	TF-IDF, K-means	Accuracy, Efficiency, Human participation	NLP, sentiment analysis in movie reviews	CLUSTER

B

APPENDIX B: LABEL DISTRIBUTIONS DATASETS



(a) Distribution of labels in dataset 1



(b) Distribution of labels in dataset 2

Figure B.1: Distribution of labels in datasets