2023

Comparative Adjudication of Noisy and Subjective Data Annotation Disagreements for Deep Learning

Scott David Williams
Wright State University

COMPARATIVE ADJUDICATION OF NOISY AND SUBJECTIVE DATA ANNOTATION DISAGREEMENTS FOR DEEP LEARNING

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

By

SCOTT DAVID WILLIAMS
B.S., Southern Illinois University, 1991
M.B.A., Southern Illinois University, 1993
Ph.D., Texas A&M University, 1999

2023
Wright State University

WRIGHT STATE UNIVERSITY

COLLEGE OF GRADUATE PROGRAMS AND HONORS STUDIES

April 25, 2023

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Scott David Williams ENTITLED Comparative Adjudication of Noisy and Subjective Data Annotation Disagreements for Deep Learning BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

_____

Krishnaprasad Thirunarayan, Ph.D.
Thesis Director

_____

Thomas Wischgoll, Ph.D,
Chair, Department of Computer Science and Engineering

Committee on Final Examination:

_____

Michael Raymer, Ph.D.

_____

Shu Schiller, Ph.D.

_____

Shu Schiller, Ph.D.
Interim Dean, College of
Graduate Programs & Honors Studies

ABSTRACT

Williams, Scott David. M.S. Department of Computer Science and Engineering, Wright State University, 2023. Comparative Adjudication of Noisy and Subjective Data Annotation Disagreements for Deep Learning

Obtaining accurate inferences from deep neural networks is difficult when models are trained on instances with conflicting labels. Algorithmic recognition of online hate speech illustrates this. No human annotator is perfectly reliable, so multiple annotators evaluate and label online posts in a corpus. Labeling scheme limitations, differences in annotators' beliefs, and limits to annotators' honesty and carefulness cause some labels to disagree. Consequently, decisive and accurate inferences become less likely. Some practical applications such as social research can tolerate some indecisiveness. However, an online platform using an indecisive classifier for automated content moderation could create more problems than it solves. Disagreements can be addressed in training by using the label a majority of annotators assigned (majority vote), training only with unanimously annotated cases (clean filtering), and representing training labels as probabilities (soft labeling). This study shows clean filtering occasionally outperforming majority voting, and soft labeling outperforming both.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

Tables Page

1. Introduction

Machine learning models struggle to accurately classify texts when labels in the training datasets that the models have been provided have high levels of inter-rater disagreement. This challenge has been observed in research on recognizing parts of speech in tweet texts (Artstein & Poesio, 2008), recognizing anaphoric references in texts (Recasens, Hovy, & Martí, 2011), hate speech detection (Botelho, Vidgen, & Hale, 2021), etc. In the case of recognizing hate speech in social media posts, when annotations have been crowdsourced, disagreements should be expected. Crowdsourced annotations are prone to spamming as well as diverse subjective interpretation. The former is because the paid annotators try to maximize their compensation by selecting labels without regard to their accuracy (Hovy, Berg-Kirkpatrick, Vaswani, & Hovy, 2013). The latter is because opinions of what content constitutes hate speech will differ from person to person based on their personal perspective and cultural background (Almanea & Poesio, 2022). Despite the difficulties that annotator disagreements present for accurate classification, hate speech recognition is important for helping to promote decency and civility online, and to help protect the community from physical harm associated with inter-group hate (Blaya, 2019).

The majority vote approach to addressing inter-annotator disagreements is the simplest approach, and it treats annotations that agree as correct and excludes annotations that disagree from training. So, we consider all training instances but associate them with their majority vote. The clean label filtering technique eliminates training cases for which the annotators disagreed.

So, the size of the training set can be reduced. The probabilistic soft labeling approach retains all training cases and all annotations by representing each training class label as a probability distribution over the classes. In other words, all training instances and all the votes are preserved.

The purpose of this study is to compare majority voting, clean label filtering and probabilistic labeling on the task of classifying a hard dataset. The dataset, Multimodal Hate Speech 150K (MMHS150K), contains nearly 150,000 microblog posts from Twitter that were labeled by annotators sourced from the public at large (Gomez, Gibert, Gomez, & Karatzas, 2020). For 25% of the posts in the corpus, a majority of the annotators labeled them as containing hate speech, and 75% were labeled as not hate by a majority of annotators. Gomez et al. reported achieving only around 68% classification accuracy, and others have reported similar results (Cheung & Lam, 2022; Prasad, Saha, & Bhattacharyya, 2021; Sai, Srivastava, & Sharma, 2022). Among the potential causes of that low classification accuracy, this study primarily focuses on challenges presented by the diversity in annotations. An advantage of annotators sourced from the public at large is that their reactions to the tweets they evaluated should be similar to the reactions of the public at large, and we observed that their rate of agreement was low when annotating MMHS150K, which makes it an appropriate dataset for demonstrating differences among majority voting, clean label filtering, and probabilistic soft labeling.

The relevance of this study is grounded in practical uses for a classifier such as to recognize online hate speech. Hate speech is an attack against a group such as a race, gender, sexual preference, religion or other category. Internet and social media platforms are expected by many to control harmful content such as hate speech. Algorithmic recognition is essential given the volume of user posts on such platforms. Additionally, sociological research and analyses involving hate speech can benefit from algorithmic hate speech recognition. Agencies that

promote public safety can leverage algorithmic classification of hate speech to be alerted to rises in potentially dangerous sentiments as well. In general, annotator disagreement is a bigger problem for some applications than for others.

Our study is presented here as follows. Chapter 2 reviews prior research on classification with annotation disagreements, the occurrence of labeling disagreements in hate speech and similar corpora, and the use cases for classifiers that can recognize hate speech. In Chapter 3, this study's research methodology is described. For our focal analyses, we trained models six ways: binary and multi-class approaches using three different training label types for each. The remaining chapters report the results, discuss the implications, and provide a brief conclusion. While the techniques demonstrated here prove to be efficacious in exploring comparative adjudication of diverse annotations, the discussion section in Chapter 5 addresses a few additional measures that may be useful in future attempts.

2. Related Work

2.1 Annotation Disagreements in Machine Learning

Typically, for a classifier to be useful in a practical application, we need the classifier to be decisive and to not equivocate. In conventional machine learning with deep neural networks, we configure the final layer to return the most likely class for each instance of input. The model's decision about the class of the input can be the impetus for some real-world action. In the use case of online hate speech detection by a social media platform, an example of an action triggered by classification would be a decision to either block or allow a post that a user is attempting to make. Although we typically want a classifier to be decisive, humans may be indecisive when presented with the same input. That is, the judgments of different individuals can disagree, and the judgments of one individual over time can be inconsistent (Shewart, Wilks, Fleiss, Levin, & Paik, 2003). For this reason, when human judgment is used to annotate training cases, the norm is to obtain them from multiple annotators and observe the level of their agreement.

Among the first factors that can lead to annotator disagreements are researchers' choices about how class labels should be defined. Predefining target categories to be used by annotators is a difficult challenge and limitations of the categories provided to annotators, or their potential misinterpretation of the intended sense by the diverse hired annotators, can causes disagreements among them. Too few categories may make it less clear which class an instance belongs to (due to over abstraction), but too many classes can increase the likelihood of annotators making errors (due to over specialization) and can lead to seldom used classes that create class imbalances, which complicates model training. A related concern in defining classes is overlapping (Uma,

Almanea, & Poesio, 2022). Annotators are typically instructed to provide a single label for each instance without a way to indicate that multiple labels can fit. Disagreement can also result from annotators not being conscientious and meticulous while annotating. Spamming is the problem of annotators providing labels without regard to their accuracy (Hovy et al., 2013). Annotators sourced through services such as Amazon's Mechanical Turk are paid for their efforts and may have no incentive to be accurate, only an incentive for finishing. Similarly, annotators who intend to provide accurate annotations, but who work too fast or make recording errors for other reasons, increase the rate of annotator disagreements (Lommel, Popovic, & Burchardt, 2014).

For many annotation tasks, subjectivity can to contribute to disagreement among annotators. To demonstrate this, Almanea & Poesio (2022) intentionally developed a diverse set of annotators to label misogyny in Arabic tweets. Annotators included both men and women and individuals who identified themselves as liberal, moderate and conservative. While gender did not have an effect on annotations, the annotators' beliefs did. Classification results for the annotated tweets demonstrated the effects of disagreement. The researchers ran multiple trials with various loss functions, and with both majority vote and probabilistic soft label configurations. Results for F1 and accuracy metrics were all in the range of 0.73 and 0.78. The authors discussed the burgeoning movement away from the notion of gold standard labels in subjective research topics such as misogyny and toward the preservation of all annotations as potentially informative. They did not, however, emphasize practical applications for classifiers with lower sensitivity and specificity attributable to disagreements in annotations.

When annotations disagree, decisions must be made about how the classes of the instances in a training dataset should be determined for training. The majority vote approach to addressing inter-annotator disagreements is the simplest approach, and it assumes the training

labels of each instance in the corpus are equally certain even though they are not. For instance, when there is disagreement among three annotations, two annotations are treated as if they are right and one is assumed to be wrong, which ignores the possibility that whatever could have caused one annotator to be wrong could have caused two annotators to be wrong. This differs from an ambiguity resolution approach in which annotators who provided different labels discuss their perspectives to see if they can reach an agreement on a consensus label, or the approach of referring the disagreement to a third party with greater expertise to make a ruling (e.g., Botelho et al., 2021). Majority voting can reduce the impact of errors or spamming by annotators. On the other hand, it has the effect of treating instances where annotators might have had equally valid but differing judgments as equivalent to instances where all annotators agreed, which disguises disagreements in the dataset. Majority voting trains classification models on both the instances we have high confidence in and the instances we have lower confidence in as if there were no differences between them.

Clean label filtering is an alternative to majority voting for dealing with annotators' disagreements. In the context of clean label filtering, the term "clean" can be viewed as a type of instance for which annotators clearly received the "signal" from the input (i.e., the message of the social media post) and were able to correctly label it (Maity, Sen, Saha, & Bhattacharyya, 2022; Ravikiran, et al., 2020; Yao, Chen, Ye, Jin, & Ren, 2021). Instances where annotators provided conflicting annotations are viewed as "noisy" and able to corrupt training (Zhang, Wu, Chen, Zhao, & Lu, 2020). By filtering the training dataset such that only clean instances are provided for training, training results can be more accurate due to the removal of noise. Furthermore, this approach can be rationalized as expecting the machine algorithm to get an annotation right if all the annotators agree and ignoring cases with multiple annotated classes that

seem intrinsically difficult as observed. To maintain fidelity to real-world applications, clean label filtering is applied to only training data, and validation and testing datasets include all instances with class labels determined by majority vote. Cleaner, less noisy input can simplify machine learning. Clean filtering has limitations though. First, it reduces the number of input instances, which is particularly costly for minority classes and for difficult to recognize classes. Lower volume input hampers learning. Second, if there is disagreement in the real-world setting of the classifier's practical application, training the classifier only on the obvious cases might not satisfactorily prepare the classifier to recognize features of the less obvious cases. While clean filtering exempts "difficult" cases from being in the training set, we do encourage separate analysis of such "outliers" for additional insights about the problematic cases with the hope of seeking remedial measures.

Probabilistic labeling of training cases, which quantifies the disagreement among annotators and provides that as input for training, is an alternative to majority voting and clean label filtering. The input vector for training class labels is a probability distribution over the labels, and a Softmax function is used. Probabilistic labels provide richer input for model training than majority vote labeling and clean filtered labeling can provide. All the cases of a corpus that are randomly assigned to the training set can be used without treating them as cases for which there is equal certainty. In other words, probabilistic labels enable using all the available training data and can explore a larger classifier design space.

2.2 An Application to Online Hate Speech

Among the many domains in which inter-rater disagreement is a concern, online hate speech classification is prominent due to subjectivity influenced by cultural contexts. "Hate

speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristics such as race, colour, ethnicity, gender, sexual orientation, nationality, religion," (Tontodimamma, Nissi, Sarra, & Fontanella, 2021, p. 157). Schweppe and Perry (2022) placed hate speech on a continuum of hate. At the lower end, microaggressions are commonplace indignities or slights toward members of oppressed groups. Hate speech, in comparison, is a more intense manifestation of hate, and it is a spoken or written expression of hostility or prejudice toward members of an oppressed group. Hate crimes are even more severe, and they combine acts that would be criminal irrespective of the motivation behind them with an expression of hatred toward members of oppressed groups. Online hate speech, also called cyberhate, is hate speech on the Internet or on a social media platform (Castaño-Pulgarín, Suárez-Betancur, Vega, & López, 2021).

Many subtleties make online hate speech difficult for humans and algorithms to detect. Social media microblogs tend to be brief and, if viewed in isolation, might not provide sufficient context for understanding the posts' true meanings. Accurate recognition of online harassment, which is akin to hate speech, requires information about contextual factors such as features of the party that posted the message in question, characteristics of the target of that message, and the relationship between the poster and the target (Shekarpour, Alshargi, Thirunarayan, Shalin, Sheth, & Rezvan, 2020). Similarly, curse words can be used in an online attack, but it can be difficult to accurately detect whether an online post using one or more curse words constitutes an attack without knowing whether there is a dialog between parties and, if so, their relationship to each other (Wang, Chen, Thirunarayan, & Sheth, 2014). Furthermore, offensive expressions such as slurs or expressions of disapproval may not be strong enough to meet hate speech's criteria of being an "attack on a group." There is unavoidable ambiguity and subjectivity regarding the

threshold at which a negative expression is severe enough to constitute an attack. Finally, even when an attack against an individual is clearly expressed, it can be unclear whether they were targeted as an individual person or as a member of a group. Given these complications, disagreements among annotations pertaining to online hate speech should be expected.

Moreover, the concept of hate speech is socially constructed, which means that we develop our understanding of it through socialization processes (Laaksonen, Haapoja, Kinnunen, Nelimarkka, & Pöyhtäri, 2020), and the diversity of socialization experiences in the public at large add to the subjectivity of hate speech judgments. Although we offered commonly used descriptions of hate speech above, in reality there are multiple definitions of hate speech in use at any given time (Boromisza-Habashi, 2021). Definitions of hate speech continue to evolve (Schweppe & Perry, 2022).

As explained in the previous section, labeling schemes for annotating hate speech in a corpus can contribute to disagreements among annotators. In the MMHS150K corpus used in this study, and explained in more detail in Chapter 3, the predefined hate speech categories provided to annotators were Not Hate, Racist, Sexist, Homophobic, Religion-based, and Other Hate. The Other Hate label was frequently used by annotators, which indicates that providing categories for hate directed at targets such as nationalities, ethnicities or political groups, which is not present now, could have been useful. Such a "kitchen sink" category might have included too many varied examples in them for machine learning to be effective at gleaning reliable features for classification. On the other hand, increasing the number of predefined labels could increase the difficulty of annotating and thereby increase the rate of annotators' disagreements and errors. Furthermore, narrowly defined categories might not have included enough instances from the corpus to support deep learning. For instance, the Religion-based hate category in

MMHS150K includes only 163 of the nearly 150,000 cases. Classes that overlap and instances that overlap classes could be a complication in the corpus as well. The instructions given to MMHS150K annotators did not facilitate clear and accurate labeling of intersectional hate that is, for instance, both racist and sexist (Kim, Ortiz, Nam, Santiago, & Datta, 2020). We do not know whether an annotator perceiving such an instance in MMHS150K would be confounded and report it as Racist, Sexist or Other Hate.

Although hate speech classification is difficult, the literature indicates there are several practical applications for online hate speech recognition, and content moderation by online platforms is prominent among them. Content moderation has been described as the practice of screening user-generated content submitted for posting on Internet sites or social media (Myers West, 2018). Content moderation is viewed by many as an ethical responsibility of online platforms (Cohen-Almagor, 2012). Given the tremendous volume of social media posts every hour of every day, it is impractical to monitor online hate speech without algorithms (Gillespie, 2020; Llansó, 2020). For algorithmic content moderation to supplant other methods, a high accuracy rate is essential. False negative cases—situations where the content is incorrectly identified as not hate speech—would fail to regulate harmful content. False positive cases, on the other hand, would errantly apply regulation to cases that are benign. Regulation could take many forms; cautioning a user about a message they are trying to post, allowing the post but with a warning flag to those who would see it, entirely blocking the message from being posted, de-platforming the user temporarily or permanently, or some combination of these measures. Any of these measures applied to a false positive case is likely to irritate a user.

Algorithmic recognition of online hate speech also has implications for government agencies, nongovernmental organizations (NGOs), and scholars. In the U.S., for the sake of

public safety, the Federal Bureau of Investigations and local law enforcement agencies monitor hate speech as it can be an antecedent to hate crimes (Bilewicz & Soral, 2020; Blaya, 2019; Lupu et al., 2023; Paul, 2019; Tsesis, 2017). Accuracy for such an application is important, but perhaps less important than for automated content moderation of online platforms. In most jurisdictions, hate speech is not a crime in-and-of-itself (Paz, Montero-Díaz, & Moreno-Delgado, 2020; Guiora & Park, 2017). But hate speech in context and situations that trigger hateful actions and crimes is pernicious. False positives might be tolerable, but a high rate of false negatives would defeat the purpose of monitoring online hate speech that could foretell an increase in hate crimes. The same principle applies to NGOs such as the Anti-Defamation League and the Southern Poverty Law Center that monitor hate speech and periodically provide reports but that have no formal authority to take action against any of its purveyors (Henry, 2009; Pereira-Kohatsu, Quijano-Sánchez, Liberatore, & Camacho-Collados, 2019). Many academic researchers likely have the same priority for classification accuracy. When the goal is to show changes in the volume of online hate speech over time in association with various social phenomena (e.g., Williams, Burnap, Javed, Liu, & Ozalp, 2020), an elevated false positive rate would only undermine the analyses if the false positive rate systematically increased or decreased over time.

There are applications for both binary classification of hate speech and more granular classification by type of hate speech in social media posts. The binary classification is fundamental to applications such as content moderation. If a user submits content for a social media post, the service provider's decision to apply content moderation interventions to the post is a binary decision. Yet algorithmic recognition of the type of hate speech in terms of the target of the hate can also be useful. Classification by type supports various analytics and may reveal

important trends. While it would be useful to be alerted by algorithmic monitoring when hate speech is trending up, it would be more useful to know which groups are being targeted when such upticks occur.

2.3 Attempts to Classify MMHS150K

Prior research indicates MMHS150K is a hard dataset to classify. Gomez et al. (2020) created it as a multimodal corpus containing both tweeted images and texts. However, they reported that multimodal models did not provide higher accuracy than textual models. None of their models exceeded accuracy of 68.5%, and text alone yielded 68.3%. Subsequent analyses of MMHS150K's tweet texts have either reported low accuracy or altered the dataset.

Sai et al. (2022) report only a maximum accuracy across trials of 67.7% on MMHS150K. They collapsed the five hate speech classes for binary classification and did not discuss how training labels were represented or how annotation disagreements were addressed. Sai and colleagues acknowledged the class imbalance and addressed it with synthetic minority oversampling (SMOTE). The researchers attempted multimodal classification of the corpus, fusing textual and visual input. For textual features, they used a pre-trained BERT model. For visual feature input, they used Inception-v3, Inception ResNet, and ResNext. Accuracy was the only performance metric reported.

Prasad et al. (2021) demonstrated various classification approaches and achieved a highest accuracy of 75.2% with a 100-dimensional GloVe embedding (Pennington, Socher, & Manning, 2014) for binary classification. They collapsed the hate speech classes into one positive class and determined training class labels by majority voting. They used a combination of majority undersampling and minority oversampling in training to address the class imbalance.

As performance metrics, they reported mean accuracies (all between 0.68 to 0.75) and ROC AUC (between 0.69 and 0.77) for all models, including several multimodal models. There was no mention of precision, recall, or F1 for the hate speech class.

Cheung and Lam (2022) achieved F1 and accuracies of around 71% in multimodal learning with MMHS150K for binary classification. No class-specific performance metrics were reported. The researchers also did not discuss how they addressed annotation disagreements or the class imbalance.

Although Shome and Kar (2021) achieved 82.6% binary accuracy using a multimodal contrastive learning approach, they did not report class-specific classification performance metrics, so it is unclear whether the accuracy metric was inflated by overclassifying the majority class. In discussing their results, they mentioned a "large number of hard negatives" boosting model performance, which might allude to their model converging more effectively on the negative class that constitutes 75% of the corpus than on the five positive classes. They conducted multi-class analyses, but did not provide class-specific classification metrics or a confusion matrix. Instead, the authors provide t-distributed stochastic neighbor embedding (t-SNE) charts. With t-SNE charts, the more effective classification has been, the more defined the clusters of points for each class in the charts will be. The authors report seeing defined clusters for all classes except Religion-based hate. However, that is a subjective assessment, and one could argue that only the Homophobic class had a clearly defined cluster. No information was provided about annotation disagreements or measures to address the class imbalance.

Botelho et al. (2021) used MMHS150K tweets but created new labels with expert annotators that had higher inter-annotator reliability (*Kappa* of 0.40 rather than the original 0.15) and in doing so enabled classification to achieve higher accuracy. In other words, they predicted

a different target than other studies. Rather than sourcing annotators from the public at large, Botelho and colleagues used two annotators who had prior experience labeling hate speech and received four weeks of training. The researchers also gave the annotators a detailed codebook to guide their labeling decisions. The codebook included definitions of classes that differed from MMHS150K's original classes, prototypical examples of the classes, and examples of edge cases. The authors noted the expert annotators agreed more frequently on Not Hate instances than on cases from the positive classes. They agreed on 68.8% of the Not Hate instances. The instances for which the two trained annotators disagreed were referred to an expert annotator who was a Ph.D. student with previous experience working on multiple hate speech research projects. This was also a study that did not report class-wise classification performance metrics or present a confusion matrix.

The training and education of experts are the main reasons for the differences between their annotations and crowdsourced annotations. Training and education can create a common perspective, and annotators with a common perspective will agree at higher rates than the public at large will. However, every observer, including an expert, has a perspective (influenced by cultural context), and that perspective influences their judgments. Perspectives influencing judgments is a definition of bias (Blair, 2012). Again, training and education do not eliminate perspective or bias, but they can cause greater conformity of perspective. As training and education cause perspectives among members of a group to narrow and agreement among them to increase (Haski-Leventhal, Pournader, & Leigh, 2020), their agreement can feel validating and can convince members of the group of their rightness even in the presence of evidence that their shared judgment is wrong (Janis, 1972). It is important to ask whether the narrowed, shared perspective of such experts is more valid or more relevant than perspectives in the public at

large. Again, context matters in interpreting social media posts (Wang et al., 2014), and the extensive training and education of Botelho and colleagues' (2021) annotators gives them a shared context, which is not a context shared by all members of the public at large. Returning to the use case of a hate speech classifier for blocking content that would offend or harm people, would the content only be problematic if it were to offend or harm the people who have been trained to have the same perspective as the annotators used by Botelho et al., or would reactions of the public at large also be relevant? If the answer is the latter, then crowdsourced labels remain relevant.

Returning to the issue of class-specific results, papers on MMHS150K routinely omit classification report tables and confusion matrices that are needed to understand the true effectiveness of classification attempts. The accuracy metric can be misleading in situations of high class imbalances such as information retrieval, cancer screening and fraud detection (Han, Kamber, & Pei, 2011; Manning, Raghavan, & Schütze, 2008).

3. Methods

3.1 The MMHS150K Dataset

Gomez and colleagues (2020) created MMHS150K to be a public domain corpus of social media content with hate speech annotations.[1] Using the Twitter API, they obtained tweets containing at least one of 51 terms commonly appearing in hate speech (ElSherief, Nilizadeh, Nguyen, Vigna, & Belding, 2018). Annotations were provided by Amazon's Mechanical Turk workers who were given a definition of hate speech and some examples before being asked to provide their annotations. Given the minimal instruction they were provided, the MMHS150K labels can be considered the ground truth from the public at large. The researchers noted that the annotators provided feedback that their task involved subjective judgments. As a check against one indicator of annotator error, the researchers rejected annotations that annotators submitted within three seconds of viewing a tweet. While all the tweets contain terms that commonly appear in hate speech, 36,978 tweets were labelled as containing actual hate speech while 112,845 tweets were labeled as not containing hate speech. Gomez and colleagues report majority vote annotations for racist, sexist, homophobic, religion-based hate and other hate as yielding tweets totaling 11,925, 3,495, 3,870, 163, and 5,811, respectively. That sums to 149,823 across classes and matches the number of instances available in the GitHub repository for MMHS150K.

[1] https://gombru.github.io/2019/10/09/MMHS/

While Gomez and colleagues report that each tweet had three annotations, we found 74 instances that had a different number of annotations. Within that set of 74, the number of annotations per tweet ranged from 1 to 4. Without an explanation as to why the number of annotations differed from the reported number, their validity was unknown, and they were not included in the analyses reported here.[2]

The MMHS150K dataset is available as a JSON file, which we parsed as a Python dictionary and then wrote to a Pandas dataframe. While parsing, instances without exactly three annotations were eliminated. Tweet texts were tokenized with Python's Natural Language Tool Kit's (NLTK's) RegexpTokenizer, then converted to lower case and lemmatized with NLTK's WordNetLemmatizer. Next, the tweets were vectorized with the GloVe 100-dimensional word vectors from a model pretrained on a corpus of 27 billion tweets with 2 billion tokens (Pennington et al., 2014). GloVe embeddings use the rates of co-occurrence of tokens in the tweets of the dataset that were used to train the GloVe model. Co-occurrence values quantitatively represent semantic meaning. Vectorizing with GloVe embeddings created a two-dimensional float array of size 33 by 100 for each tweet. Preprocessing of the training set's target classes involved creating binary targets (for the replication analysis only), and the primary analyses used one-hot encoding or probabilistic labeling.

In every classification analysis that we report, majority vote was used to determine class labels for validation and testing, but in some instances, there was no majority. No majority means two or three annotators provided a hate speech label, but no two of the labels matched.

[2] It is unclear how cases with one annotation should be treated in a study of annotator disagreements, and how cases with two disagreeing annotations should be treated in the majority voting paradigm. It is possible that all the anomalous cases were the result of transcription errors. Fortunately, the dropped cases only represent 0.05% of the corpus and could not have had any substantive effect on the results.

We refer to these cases as Mixed Hate. In the binary paradigm, the majority vote yields a consensus label of hate for all the Mixed Hate instances even though no two annotators agreed on the type of hate. In the multi-class paradigm, identifying an appropriate approach to producing a consensus label was a dilemma. Alternatives included discarding those cases, retaining the Mixed Hate class as a seventh class, using a new group of annotators to obtain new labels, and including Mixed Hate in the Other Hate class. Each approach would have had its advantages and disadvantages. If the cases had been discarded, the retained cases would be cleaner, but the corpus would sacrifice some fidelity to a real-world classification challenge. Keeping them as a seventh class would involve introducing a class to the analyses with an ambiguous interpretation that would likely be difficult for a classifier to recognize as distinct. Using a new group of annotators would be resource-intensive and might not generate much improvement.

For this study, the cases that had no consensus label were added to the Other Hate class. If the annotators are viewed as the arbiters of ground truth, then the literal interpretation of cases with no consensus is intersectional hate. For instance, if a tweet expressed hate toward women wearing burkas (e.g., Fortuna & Nunes, 2018), given that the annotators had to select one of the predefined labels that did not account for intersectional hate, then conflicting annotations should be expected (e.g., Racist, Sexist, Religion-based, or Other Hate). The Other Hate class as originally provided by Gomez et al. was already a heterogenous class of tweets expressing hate. Adding the cases with no majority vote is unlikely to corrupt the Other Hate class. The class counts are as reported in Table 1. For completeness, Mixed Hate is shown, but we did not run analyses that treated it as a separate class.

Table 1. Annotations by Class and Agreement

				Mixed as Other	
Multi-class Labels	Clean	2 of 3	Majority Vote	Clean	Majority Vote
Non Hate	57,890	54,897	112,787	57,890	112,787
Racist	1,710	8,898	10,608	1,710	10,608
Sexist	407	2,270	2,677	407	2,677
Homophobic	887	2,190	3,077	887	3,077
Religion-based	23	88	111	23	111
Other Hate	1,065	3,797	4,862	5,547	20,489
Mixed Hate[1]	4,482	11,145	15,627		
	66,464	83,285	149,749	66,464	149,749

Binary Labels	Clean	2 of 3	Majority Vote
Non Hate	57,890	54,897	112,787
Hate	8,574	28,388	36,962
	66,464	83,285	149,749

[1] Mixed Hate was added during preprocessing. Instances where all three annotations were a form of hate are shown here as Clean. Instances where one annotation was Not Hate and the other two were different classes of hate are shown as 2 of 3.

3.2 Base Case

Our initial analyses involved running several experimental trials to become familiar with the dataset, to replicate the results reported by Gomez and colleagues, to observe the effects of different model architectures and hyperparameters, and to establish an appropriate base case for comparisons. In an attempt to replicate the findings of Gomez and colleagues, we used the text-only classification model features they described: a single 150-unit long short-term memory (LSTM) hidden state, a Binary Cross-Entropy loss function, sigmoid activation, and an Adam optimizer with a learning rate of 0.0001.

With negative cases outnumbering positives 3-to-1 in the corpus, we used class weighting in training. As previously mentioned, initial trials showed models tended to converge on the larger and easier to classify negative class. We tested whether SMOTE or adaptive synthetic sampling (ADASYN) would be more effective than class weighting, but they were not. We also examined whether a Jensen-Shannon Divergence loss function (also known as Reverse Kullback-Liebler Divergence), which factors the distribution of predicted labels into loss calculations, would reduce the extent to which models would over-predict the large negative class, but it did not perform better than Cross-Entropy (results available upon request).

3.3 Comparative Adjudication of Label Disagreements

After attempting a pure replication, we adapted the model to create a base case that would permit closer comparisons when subsequently conducting trials with clean label filtering and soft labeling. The base case for binary classification has a 128-unit layer, which may make processing and memory use more efficient when compared to the 150 units Gomez et al. used. Additionally, the binary target was one-hot encoded to form two classes. As shown in Table 2, example tweet B received annotations of 3, 0, and 0. The annotations in Tables 2 and 3 are unordered lists as they appear in the JSON file for MMHS150K. The class codes for annotations are 0 for Not Hate, 1 for Racist, 2 for Sexist, 3 for Homophobic, 4 for Religion-based, and 5 for Other Hate.[3] Representing example tweet B's label for binary classification with majority voting involves creating a one-hot encoded vector with 1.0 in index position 0 and 0.0 in index position 1, which is shown in the Majority column under Binary in Table 2. The other example tweets in

[3] By obtaining only one label per annotator per instance, Gomez and colleagues (2020) made each class annotation mutually exclusive.

Table 2. Representation of the Class Labels in Training

		Binary			Multi-Class		
Examples	Annotations	Majority	Clean	Soft	Majority	Clean	Soft
tweet A	[2, 0, 2]	[0.0, 1.0]	out	[0.33, 0.67]	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0]	out	[0.33, 0.0, 0.67, 0.0, 0.0, 0.0]
tweet B	[3, 0, 0]	[1.0, 0.0]	out	[0.67, 0.33]	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]	out	[0.67, 0.0, 0.0, 0.33, 0.0, 0.0]
tweet C	[3, 3, 0]	[0.0, 1.0]	out	[0.33, 0.67]	[0.0, 0.0, 0.0, 1.0, 0.0, 0.0]	out	[0.33, 0.0, 0.0, 0.67, 0.0, 0.0]
tweet D	[3, 3, 3]	[0.0, 1.0]	[0.0, 1.0]	[0.0, 1.0]	[0.0, 0.0, 0.0, 1.0, 0.0, 0.0]	[0.0, 0.0, 0.0, 1.0, 0.0, 0.0]	[0.0, 0.0, 0.0, 1.0, 0.0, 0.0]
tweet E	[5, 5, 5]	[0.0, 1.0]	[0.0, 1.0]	[0.0, 1.0]	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0]	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0]	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0]
tweet F	[0, 1, 4]	[0.0, 1.0]	out	[0.33, 0.67]	[0.0, 0.0, 0.0, 0.0, 0.0, 1.0]	out	[0.33, 0.33, 0.0, 0.0., 0.33, 0.0]

Annotation codes = {0: Not Hate, 1: Racist, 2: Sexist, 3: Homophobic, 4: Religion-based, 5: Other Hate}

Table 3. Representation of the Class Labels in Validation and Testing

Examples	Annotations	**Binary**	**Multi-Class**
tweet X	[1, 1, 0]	[0.0, 1.0]	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0]
tweet Y	[0, 0, 0]	[1.0, 0.0]	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]
tweet Z	[0, 5, 0]	[1.0, 0.0]	[1.0, 0.0, 0.0, 0.0, 0.0, 0.0]

Annotation codes = {0: Not Hate, 1: Racist, 2: Sexist, 3: Homophobic, 4: Religion-based, 5: Other Hate}

Table 2 have at least two annotations of a form of hate speech and are one-hot encoded with 1.0 in index position 1 for binary analyses in order to represent the Hate class. Accordingly, the final layer of the neural network had two units with a Softmax activation rather than sigmoid activation.

The difference between the one-hot encoded binary base case analysis and the clean label filtered analysis was simply filtering of the training dataset. As Table 2 shows, only instances in the training dataset for which all three annotators agreed on a label (e.g., example tweets D and E) were included. Unanimously labeled case were identical in the base case and the clean label training datasets. The model architecture features, the hyperparameters, and the validation and testing datasets (examples in Table 3) were the same for the base and clean labels cases.

The differences between the one-hot encoded binary base case analysis and the probabilistic soft label analysis were features of the training set's target class data. One-hot encoding represents a label as [0.0, 1.0] for a positive hate speech instance whether all three annotations were positive or only two of the three. In contrast, if only two of three annotators provided a hate speech label for an instance, its soft label would be [0.33, 0.67]. This is shown in the Soft column under Binary in Table 2 with examples A, C and F. The model architecture features, and the hyperparameters we used for the soft labels analyses matched those of the base case.

After training models with binary targets (Not Hate and Hate classes) three ways (majority vote, clean filtered, and soft labeled), we trained models three ways with multi-class targets. The six classes were those defined by the annotation scheme Gomez and colleagues provided to annotators. In the examples provided in Table 2, tweet B received one annotation for Homophobic, and tweet C received two annotations for Homophobic. For multi-class majority

vote analyses, tweet B is one-hot encoded as Not Hate (1.0 in index position 0) and tweet C is one-hot encoded as Homophobic (1.0 in index position 3). The label for each class is represented by the index where 1.0 appears in the vector. As previously mentioned, for the majority vote multi-class analyses, we chose to put the hate speech-labeled instances that had no majority class in the Other Hate class. That is illustrated by tweet F in the Majority column under Multi-Class in Table 2.

The representation of labels for multi-class clean filtering and multi-class soft labeling follows the pattern used for the binary representations. As Table 2 shows, since only example tweets D and E received unanimous annotations, they are the only examples that would have been used in multi-class clean filtered training. The label representations in the Soft column under Multi-Class in Table 2 show that labels are represented as the proportions of the three annotations they received for each of the six classes. In the label vector for a given tweet, each proportion represents the probability that the tweet belongs to the class associated with that index position given the annotations for that tweet.

To address the issue of models tending to converge on Not Hate at the cost of poor recognition of hate speech instances, we used F1 for the hate speech classes as one metric for selecting a model. In training with multi-class targets, F1 was computed across all positive classes for this purpose. In Keras model training, callbacks were set to save the model with the best hate speech F1 and the best validation accuracy.

4. Results

4.1 Replication of Gomez and colleagues (2020)

As previously stated, considerable inaccuracy is inevitable given the nature of online hate speech and the MMHS150K corpus. However, superior approaches to classification can improve overall classifier performance and address particular problems in classification. Beginning with our replication of Gomez and colleagues' (2020) analyses, the results show that the weighted average accuracy is low. More importantly, inaccuracy in predicting the hate speech cases is a major problem since that is the main use of the classifier. As Table 4 shows, the base case classifier only recognized hate speech about half of the times when it was present. A recall ratio of 0.514 for the Hate class means that, of the 2,495 test cases that were Hate, the model only recognized 51% of them. With Not Hate being a much bigger and easier to recognize class, the model minimized loss by achieving much higher recall for Not Hate than it did for the target of interest. "Accuracy" in Table 4 is micro accuracy, which is simply the ratio of the number of correct predictions to the number of cases, and it can overlook difficulties a model might have with predicting smaller classes. Weighted average precision, recall and F1-score values can also distract from such problems since they give greater weight to larger classes. Macro averages for each performance metric are an improvement because they give equal weight to each class when averaging metrics across classes. Table 5 presents the confusion matrix.

Table 4. Classification Report for the Replication

	Precision	Recall	F1-score	Support
Not Hate	0.831	0.797	0.814	7,505
Hate	0.457	0.514	0.484	2,495
Accuracy			0.726	10,000
Macro Avg.	0.644	0.655	0.649	10,000
Weighted Avg.	0.738	0.726	0.732	10,000

Table 5. Confusion Matrix for the Replication

		Predicted	
		Not Hate	Hate
Actual	Not Hate	5,982	1,523
	Hate	1,213	1,282

Our results differ from those obtained from the original classification of MMHS150K by Gomez and colleagues who reported accuracy of 0.683, F1 of 0.703, and area under the curve (AUC) of 0.732. Our attempt at replicating their analyses produced micro accuracy of 0.726, weighted average F1 of 0.732, and AUC of 0.656. These differences could be due to random differences in the class distributions of the respective holdout datasets, random differences in the initialization weights, or the choice of a classification performance metric for selecting a model from those created during training. While replicating, we chose the model with the highest validation accuracy across training epochs.

Among the analyses reported in this study, we established a baseline for comparisons and kept the criteria for model selection consistent. In each model fitting, we allowed training to continue until validation loss was clearly trending up. Rather than using the last model produced during training, we used callbacks to save the best model for the specified metric. We kept both the model with the highest accuracy across classes and the model with the highest F1-score for the hate speech classes. We report results for both the maximum accuracy models and the maximum hate speech F1-score models. As we noted, the class imbalance makes the accuracy metric misleading, and precision and recall should be the focus. However, accuracy is the only metric consistently reported in prior MMHS150K studies, so including it here facilitates comparisons to prior work.

4.2 Majority Vote, Binary

Table 6 contains the baseline results for binary classification of MMHS150K. Means and 95% confidence intervals are included to clarify which differences across analyses are unlikely to be due to randomness. The model that achieved the best accuracy across all epochs with the validation data also achieved better accuracy with the test data than the model that maximized F1 across epochs did. However, the model that maximized hate speech F1 achieved higher recall for the hate speech class. Precision for the positive class was lower though with the maximum hate speech F1 model. The AUC values are equivalent. The weighted average F1 value for the model that maximized accuracy is greater than the same metric value for the model that maximized hate speech F1.

Table 6. Performance Metrics for Majority Vote Labels, Binary

	Maximizing Accuracy			Maximizing Hate Speech F1		
	Lower Bound	Mean	Upper Bound	Lower Bound	Mean	Upper Bound
Weighted Avg. F1	0.713	0.722	0.728	0.678	0.688	0.697
Area Under the Curve	0.642	0.651	0.663	0.653	0.663	0.673
Accuracy	0.714	0.715	0.716	0.668	0.669	0.669
Hate Speech Precision	0.438	0.440	0.441	0.398	0.399	0.400
Hate Speech Recall	0.510	0.526	0.542	0.632	0.651	0.671

4.3 Clean Filtered, Binary

Table 7 contains the results for binary classification with the training dataset filtered such that only clean labels—labels on which all three annotators agreed—were retained. As with the maximum validation accuracy model saved from the baseline training trial, the maximum validation accuracy model from the clean filtered training trial achieved better accuracy and hate speech precision with the testing data. The model that maximized hate speech F1 from the clean filtered trial led to higher recall.

Comparing the metrics in Tables 6 and 7, clean label filtering led to equivalent hate speech recall and slightly lower hate speech precision when using the model from each trial that maximized accuracy. The weighted average F1-scores, AUC values and accuracies are nearly identical for the majority vote and clean filtered models. When comparing the models maximizing F1 across the two training data approaches, the clean filtered approach had higher recall, but had slightly lower hate speech precision, weighted average F1 and accuracy. As mentioned in Chapter 2, training the classifier only on the obvious cases might not satisfactorily prepare the classifier to recognize features of the less obvious cases. This might explain the slightly higher hate speech recognition with the majority vote approach. AUC was equal in the two conditions.

Table 7. Performance Metrics for Clean Filtered Training, Binary

	Maximizing Accuracy			Maximizing Hate Speech F1		
	Lower Bound	Mean	Upper Bound	Lower Bound	Mean	Upper Bound
Weighted Avg. F1	0.712	0.720	0.728	0.661	0.667	0.676
Area Under the Curve	0.640	0.650	0.660	0.652	0.662	0.672
Accuracy	0.711	0.712	0.713	0.642	0.643	0.644
Hate Speech Precision	0.435	0.436	0.438	0.380	0.382	0.383
Hate Speech Recall	0.513	0.529	0.546	0.674	0.696	0.714

4.4 Soft Labels, Binary

Table 8 contains the results for binary classification with probabilistic soft labels used during training. Comparing results of the model that maximized accuracy and the model that maximized hate speech F1shows a pattern that was present in the previously discussed models. Maximizing hate speech F1 led to much higher hate speech recall but lower accuracy and lower hate speech precision.

Table 8. Performance Metrics for Soft Labels, Binary

	Maximizing Accuracy			Maximizing Hate Speech F1		
	Lower Bound	Mean	Upper Bound	Lower Bound	Mean	Upper Bound
Weighted Avg. F1	0.742	0.749	0.757	0.674	0.682	0.689
Area Under the Curve	0.636	0.646	0.657	0.653	0.664	0.674
Accuracy	0.750	0.756	0.765	0.652	0.661	0.668
Hate Speech Precision	0.491	0.513	0.533	0.380	0.393	0.407
Hate Speech Recall	0.408	0.427	0.445	0.654	0.672	0.686

The confusion matrices for soft label binary classification trials in Table 9 illustrate the tradeoffs. The model saved from the training trial with soft labeled targets that maximized hate speech F1 in validation made far more positive predictions in the testing set. In doing so, it

produced 59% more true positive predictions and 61% fewer false negative predictions. However, the model that maximized hate speech F1 also made 145% more false positive predictions. Accordingly, Table 6 shows the accuracy for the model that maximized hate speech F1 generated much lower accuracy with the testing data (0.661 versus 0.756).

Table 9. Confusion Matrices for Soft Labels, Binary

Maximizing Accuracy

		Predicted	
		Not Hate	Hate
Actual	Not Hate	6,536	969
	Hate	1,508	987

Maximizing F1

		Predicted	
		Not Hate	Hate
Actual	Not Hate	5,131	2,374
	Hate	926	1,569

When comparing the models trained with majority vote labels represented in Table 6 and the models trained with probabilistic soft labels in Table 8, the models that maximized accuracy in training with soft labels produced lower hate speech recall, but higher hate speech precision and overall model accuracy. On the other hand, the models that maximized hate speech F1 in those trials mainly differed in hate speech recall. The values for hate speech precision and overall model accuracy differed by less than a percent. The accuracy and hate speech precision versus hate speech recall tradeoff previously discussed also occurred when comparing the majority vote

training results to soft label training results. True positives were higher for the soft label trained model that maximized hate speech F1, but accuracy was slightly lower.

The classification metrics for models trained with clean filtered labels reported in Table 7 and for models trained with soft labels reported in Table 8 exhibit the accuracy and hate speech precision versus hate speech recall tradeoff in the accuracy maximizing condition. The model trained with clean labels that maximized accuracy achieved that performance at the cost of hate speech precision. At a precision level of 0.436, that model was usually wrong when predicting that an instance was hate speech. High accuracy was achieved via a higher true positive rate. On the other hand, the model that maximized validation accuracy during training with soft labels had the highest model accuracy across all of our binary model training at 0.756 via lower recall at 0.427, which is undesirable when the goal is hate speech recognition. Comparing the maximizing hate speech F1 conditions between Table 7 and Table 8, we see it is the same pattern as previously discussed when comparing Tables 6 and 8. Hate speech precision can be increased at a cost.

In summary, ranking performance of the various binary classification models, training with probabilistic soft labels and retaining the model that achieved the highest accuracy with the validation dataset led to the highest accuracy. However, the highest accuracy was achieved while failing to recognize most instances of hate speech (recall of 0.427). Across the training runs for majority vote, clean label filtering and soft labeling, the models that optimized hate speech F1 in each trial provided highly similar performance metrics. In each of the three training trials, hate speech recall was higher for the model that maximized hate speech F1, but lower values for precision also resulted. Across majority vote, clean filtering and soft labels trials, clean label filtering achieved the highest recall (0.696) and the lowest precision (0.382).

4.5 Majority Vote, Multi-Class

Switching from binary to multi-class classification, Tables 10 through 13 contain results for multi-class classification of MMHS150K. Table 10 contains the baseline results for multi-class classification in which the annotations were defined by majority vote and one-hot encoded. Even for the model saved during training for having the highest accuracy with the validation data, the accuracy of 0.534 is low. Recall for the Homophobic class is better at 0.776, and for the Other Hate class at 0.885, but precision across all hate speech classes is low. As expected, the model saved from training for having the highest hate speech F1 produced some slightly better hate speech recall values at the cost of lower overall accuracy.

Table 10. Performance Metrics for Majority Vote Labels, Multi-Class

	Maximizing Accuracy			Maximizing Hate Speech F1		
	Lower Bound	Mean	Upper Bound	Lower Bound	Mean	Upper Bound
Weighted Avg. F1	0.602	0.611	0.621	0.557	0.566	0.576
Area Under the Curve	0.812	0.819	0.827	0.815	0.823	0.830
Accuracy	0.525	0.534	0.544	0.470	0.480	0.490
Precision - 'Racist'	0.400	0.410	0.419	0.357	0.367	0.376
Precision - 'Sexist'	0.191	0.199	0.207	0.198	0.206	0.214
Precision - 'Homophobic'	0.337	0.346	0.355	0.337	0.347	0.356
Precision - 'Religion-based'	0.026	0.029	0.032	0.021	0.024	0.027
Precision - 'Other hate'	0.100	0.106	0.112	0.087	0.093	0.099
Recall - 'Racist'	0.173	0.181	0.189	0.187	0.195	0.202
Recall - 'Sexist'	0.233	0.241	0.250	0.248	0.257	0.265
Recall - 'Homophobic'	0.768	0.776	0.784	0.772	0.780	0.788
Recall - 'Religion-based'	0.390	0.400	0.410	0.590	0.600	0.610
Recall - 'Other hate'	0.879	0.885	0.891	0.869	0.876	0.882

4.6 Clean Filtered, Multi-Class

Table 11 contains the results for multi-class classification with the training dataset filtered to use only the unanimously annotated cases. The model from the training trial that maximized accuracy with the clean filtered training dataset achieved higher accuracy than the comparable model in Table 10. The maximum accuracy model in Table 11 also recognized a majority of Sexist, Homophobic, and Other Hate instances, but precision values were low. The model from clean filtered training that maximized hate speech F1 recognized a majority of the Sexist, Homophobic, and Religion-based hate cases. While precision was low across all classes, that model balanced recall and precision for Other Hate.

Table 11. Performance Metrics for Clean Filtered Training, Multi-Class

	Maximizing Accuracy			Maximizing Hate Speech F1		
	Lower Bound	Mean	Upper Bound	Lower Bound	Mean	Upper Bound
Weighted Avg. F1	0.629	0.638	0.648	0.565	0.575	0.585
Area Under the Curve	0.798	0.806	0.814	0.786	0.794	0.802
Accuracy	0.567	0.576	0.586	0.485	0.495	0.505
Precision - 'Racist'	0.175	0.183	0.190	0.144	0.151	0.159
Precision - 'Sexist'	0.113	0.119	0.126	0.108	0.114	0.120
Precision - 'Homophobic'	0.333	0.342	0.352	0.259	0.267	0.277
Precision - 'Religion-based'	0.059	0.063	0.068	0.008	0.010	0.012
Precision - 'Other hate'	0.313	0.322	0.331	0.384	0.394	0.403
Recall - 'Racist'	0.392	0.402	0.412	0.403	0.412	0.422
Recall - 'Sexist'	0.650	0.659	0.668	0.646	0.655	0.664
Recall - 'Homophobic'	0.757	0.765	0.773	0.778	0.787	0.795
Recall - 'Religion-based'	0.390	0.400	0.410	0.590	0.600	0.610
Recall - 'Other hate'	0.598	0.608	0.618	0.421	0.431	0.439

Table 12. Performance Metrics for Soft Labels, Multi-Class

	Maximizing Accuracy			Maximizing Hate Speech F1		
	Lower Bound	Mean	Upper Bound	Lower Bound	Mean	Upper Bound
Weighted Avg. F1	0.742	0.751	0.753	0.736	0.744	0.753
Area Under the Curve	0.723	0.732	0.742	0.827	0.835	0.842
Accuracy	0.818	0.825	0.828	0.714	0.722	0.731
Precision - 'Racist'	0.000	0.000	0.000	0.365	0.375	0.384
Precision - 'Sexist'	0.354	0.364	0.373	0.170	0.177	0.185
Precision - 'Homophobic'	0.000	0.000	0.000	0.320	0.330	0.339
Precision - 'Religion-based'	0.000	0.000	0.000	0.018	0.020	0.023
Precision - 'Other hate'	0.000	0.000	0.000	0.288	0.297	0.306
Recall - 'Racist'	0.000	0.000	0.000	0.190	0.198	0.206
Recall - 'Sexist'	0.015	0.018	0.021	0.318	0.327	0.337
Recall - 'Homophobic'	0.000	0.000	0.000	0.754	0.762	0.770
Recall - 'Religion-based'	0.000	0.000	0.000	0.770	0.778	0.786
Recall - 'Other hate'	0.000	0.000	0.000	0.544	0.554	0.564

4.7 Soft Labels, Multi-Class

Table 12 reports the multi-class classification results for models trained with probabilistic soft labels. The model from the training trial that maximized accuracy did so with a high true negative rate. The accuracy metric looks favorable when compared to other classification accuracies presented here and in other studies, but it is misleading. That model did not, in fact, recognize hate speech. This is illustrated in the confusion matrix presented in Table 13. The model that maximized accuracy during the training trial was produced in the first epoch. It was an outlier compared to the models produced in subsequent epochs of that training trial. The outlier is included here to illustrate that MMHS150K has a large enough class imbalance that maximizing overall accuracy can lead to failure to recognize hate speech. This reinforces the need for MMHS150K studies to report class-wise classification performance metrics.

Table 13. Confusion Matrix Soft Labels, Multi-Class, Maximizing Accuracy

		Predicted					
		Not Hate	Racist	Sexist	Homophobic	Religion-based	Other Hate
Actual	Not Hate	8247	0	7	37	0	1
	Racist	822	0	0	0	0	0
	Sexist	211	0	4	8	0	0
	Homophobic	273	0	0	0	0	0
	Religion-based	9	0	0	0	0	0
	Other Hate	379	0	0	2	0	0

Returning to Table 12, the righthand side shows that the model from the trial that maximized hate speech F1 produced the best overall metrics for multi-class classification. Accuracy at 0.722 was higher than was obtained with majority vote and clean filtered training. More importantly, it achieved a better balance of hate speech recall and precision in many instances than was achieved by the models in Tables 10 and 11. Across all the multi-class hate speech F1 maximizing models (i.e., the righthand sides of the tables), recognition of the Homophobic class was the best. Apparently, models converge more easily on that hate speech class, which was also demonstrated by Shome and Kar (2021). Being a small class, learning to recognize homophobic tweets in the testing dataset does little to boost overall model accuracy.

In summary, ranking the multi-class models, clean label filtered training performed somewhat better than majority vote training at hate speech recognition, and probabilistic soft labeling performed better than both. For the models saved for maximizing hate speech F1 in each of the majority vote and clean filtered training approaches, the most noticeable differences were that clean filtering training cases recognized the Sexist class better and balanced precision and recall for Other hate better. Training with soft labeled training cases, on the other hand, had

clearly different results. First, that trial produced the highest model accuracy at 0.825, but Table 10 shows that it did not actually recognize hate speech. Second, the model from that same training trial that generated the highest hate speech F1 was actually the best multi-class hate speech classifier. It achieved accuracy of 0.722 and it balanced recall and precision for each hate speech class a little better than the majority vote and clean filtered trained models did.

5. Discussion

5.1 Labels and Performance Metrics

The original goal of this study was to compare clean label filtering and probabilistic soft labeling of targets to improve hate speech classification performance. The results show that clean filtering of training datasets to train with only unanimous cases, or representing training labels as probabilities, can improve training accuracy. Surprisingly and somewhat counter-intuitively, models trained with probabilistic soft labels recognized hate speech more effectively than either the clean label filtered training models or the baseline majority vote models. However, the results did not reveal any simple solution to effectively classifying MMHS150K as error rates (as measured appropriately using F1-score, in preference to accuracy, due to imbalanced dataset as explained below) were consistently high.

A problem encountered in both binary and multi-class trials is that models tended to converge on the easier to classify Not Hate class. Also, since 75% of the MMHS150K cases are Not Hate, accuracy around that level can be achieved by not recognizing hate speech (that is labelling everything as Not Hate). To address this pattern, we retained the model from each trial that maximized F1 for hate speech classes during validation and reported testing results for them. As expected, recall for hate speech classes was higher for such models at the expense of lower overall accuracy. Somewhat surprising was that hate speech precision also tended to be lower when testing models that maximized hate speech F1. To identify more hate speech in

MMHS150K, lower precision and lower accuracy would have to be accepted. We revisit tradeoffs such as this in Section 5.2.

Future research with MMHS150K and other hate speech corpora should report class-wise performance metrics. This would give reviewers and readers key insights into hate speech classification effectiveness. For instance, in this study, multi-class classification accuracy in the testing data of 82.5% was generated by a multi-class classification model using soft labeling. That is a significantly higher accuracy than the 68.3% accuracy for binary classification reported by Gomez et al. (2020) in the original MMHS150K study. However, the confusion matrix in our Table 13 demonstrates how misleading that 82.5% accuracy metric is. The model predicted that 99% of the testing cases were Not Hate. This reaffirms that accuracy can be the wrong metric for evaluating classification performance when there are class imbalances (Han et al., 2011; Manning et al., 2008). Confusion matrices for classifying MMHS150K, on the other hand, can be informative, but papers on MMHS150K omit them (e.g., Cheung & Lam, 2022; Prasad et al., 2021; Sai et al., 2022; Shome & Kar, 2021).

5.2 Implications for Practical Applications

Interpreting the significance of the results relative to use cases, the models reported here would be more appropriate for research and analysis applications than for automated content moderation. The probabilistic soft label approach to training led to the best hate speech recognition in the binary paradigm and to potentially useful levels of recognition of sexist and homophobic content in the multi-class paradigm. Such models can be used in applications where inexact classification can be tolerated.

Looking back at the data collection approach used to create MMHS150K, it was 25% accurate. The researchers crawled Twitter collecting tweets that contained terms that are common in hate speech. By keywords alone, the tweets were predicted to be positive for hate speech. The Amazon Turk annotators identified only 25% of the tweets as actually containing hate speech, so the false positive rate for the keyword search was 75%. The models reported here are clearly more efficient at recognizing hate speech than a keyword search for hate speech terminology in social media posts would be. For research and analysis on online hate speech, data could be processed in three stages. As Gomez and colleagues did to create their corpus, an online platform could be crawled to obtain posts with hate speech terminology in the first stage. In the second stage, a classifier could be applied to improve recognition of the target. A third stage of manually evaluating cases that were classified as positive before conducting analyses would be able to further clean the corpus, based on the needs of the application. Although manually processing online posts is resource-intensive, preprocessing with a classifier such as one of models presented here would help lighten the workload. Additionally, the model's threshold for assigning a hate speech label can be adjusted to make the model more or less conservative with respect to false positives. Finally, by applying a Softmax function to the logits of the final layer and outputting them, the values can serve as indicators of the model's confidence in a given prediction, and the user can use that data to choose a cutoff point.

None of the models from this study would be a good choice for an online platform to use as its sole content moderation tool. Twitter, for instance, would be unlikely to find any of the models reported here constructive for algorithmic hate speech blocking. Even the model with the highest accuracy achieved for binary classification has high rates of false negatives and false positives. Many users would be justifiably upset if that model were to be applied. In the case of

false negatives, approximately 18% of the tweets that annotators judged to have contained hate speech would not be recognized as such and thus be posted. Unless there were other mechanisms for controlling those posts, users who viewed them would be offended and potentially harmed. Furthermore, if a content moderation mechanism were known to be in use, and if a hate speech tweet was not blocked by it, the effect would be to signify that the content was socially acceptable. False positives would also create problems for users due to blocking of posts they attempted to make because they were mistaken for hate speech. This would lead to user disenfranchisement. Other efforts to develop algorithmic hate speech detection for content moderation are needed.

A classifier such as those presented here could be used to complement other methods of content moderation. The classifier could also be used for screening posts to identify those that require further review. While manual processes for screening all social media contributions before posting are infeasible for high-volume platforms, posts that are algorithmically classified as containing hate speech can be referred for manual evaluation without waiting for users to report the content based on the application context. In this scenario, the message would not initially be blocked, but human content moderators can be alerted that there is potentially inappropriate content to be evaluated, or flag the account by placing it on a low-severity-watchlist for more careful automatic monitoring for acquiring more reliable signals of hate speech. The more proactive and efficient content moderators are, the shorter the time a hate speech message remains online doing harm.

5.3 Future Research on Hate Speech Classification

Fortunately, in addition to the approaches applied in this study, various other approaches have the potential to improve classification accuracy of MMHS150K and similar corpora. The approach of this study was somewhat narrow with the intent of providing depth. In terms of scope, our intent was to replicate the analyses by Gomez and colleagues, establish binary and multi-class baselines, and experiment with two approaches for improving classification performance by addressing limitations of the annotations. More importantly, our study focuses on the impact of different approaches to resolving annotation discrepancies in crowdsourced annotations and their comparative impact on classifier performance as discussed later. In *post hoc* analyses, we also used a criterion for model selection from training that increased hate speech recognition. Other techniques attempted *post hoc* but not reported here due to their limited value were SMOTE and ADASYN as alternatives to class weights, various model architectures, and numerous hyperparameters (results available upon request). Nevertheless, MMHS150K is a useful corpus, and future research is likely to find additional classification improvements.

Future research with variations of the classification approach presented here might be able to produce superior results. For instance, the Bidirectional Encoder Representations from Transformers pretrained on tweets (BERTweet; Nguyen, Vu, & Nguyen, 2020) could be used in place of GloVe embeddings or in fusion with GloVe embeddings (Eke, Norman, & Shuib, 2021). Additionally, Bidirectional LSTM (BiLSTM) can be used as an alternative to LSTM in order to represent backward dependencies in the input sequences as well as the forward dependencies of standard LSTM (Khan et al., 2022). As a third suggestion, rather than using a single holdout for testing, future research with different testing datasets might lead to different results. For the sake

of consistency, we used the same holdout for testing all six of the class labeling approaches we compared. The holdout was created with the Random method of the Pandas package. With this approach, the expected distribution of classes in the holdout is the distribution of the classes in the corpus. However, the actual distribution can vary. This situation can be addressed in future research either with stratified random sampling to maintain the class proportions of the corpus (Ramezan, Maxwell, & Warner, 2019), or with *k*-fold cross-validation that iterates training and testing on subsets of the corpus until all cases have been used in a testing holdout (Roy, Tripathy, Das, & Gao, 2020). The computational expense of *k*-fold cross-validation on MMHS150K could be very high though. Our focus here was on adjudication of label disagreements, and we followed the approach of Gomez and colleagues (2020)—who used GloVe, LSTM and a single holdout for testing—so that our results could be comparable. Future research may benefit by deviating from that approach while using probabilistic soft labeling, which our study found to perform the best.

Perhaps the greatest opportunity to improve algorithmic hate speech recognition relative to the approaches that this study reports is to improve the crowdsourcing of annotations. Earlier in the paper, we provided a rationale for using the judgments of annotators sourced from the public at large. Motivating annotators such as Mechanical Turk workers to provide high-quality labels will be essential when obtaining labels. To retain the opportunity to participate and contribute annotations, annotators could be asked to demonstrate they are providing high-quality labels. One way to do this is to have them provide annotations for a batch of instances, asking them first to try to provide labels that they think will match the labels "most people" have provided and then to provide labels based on their own interpretation of each message. After completing a batch of about 20, the annotations they provided could be compared to pre-

established classes for at least a portion of the cases, and feedback could be provided on how accurately they predicted those classes. Accuracy could be a criterion for continuing with the task. To retain valid annotator disagreements in the corpus, an annotator would not need to have a high degree of accuracy in predicting the instances with pre-existing labels, but the accuracy rate should not decline. A declining rate of accuracy could indicate fatigue or spamming.

It should be noted, however, that if annotators get too much training, they will then start to be more like experts and less like the public at large, which might not be valid for all use cases. When the concern is to faithfully mirror the reactions of the public at large, retaining and representing all the labels provided (rather than seeking them from the experts) may be the right course of action. Given the diversity of the public at large, hate speech classification should tolerate lower levels of agreement. Classifying for hate speech is different than classifying for other targets such as images of cats. Images of cats have a ground truth that is anchored in the "hard science" of zoology. No such anchor exists for the phenomenon of hate speech. Prior work demonstrates that knowing a tweet's context is essential to accurately classifying it (Wang et al., 2014). It seems unrealistic to expect algorithmic detection of hate speech to achieve high accuracy levels without a deeper modelling of language, context, and current events in practice. When delving into the sociology, political science and history of hate speech, its subjective and evolving nature quickly becomes clear. Readers interested in the social construction of hate speech can see, for instance, Laaksonen et al. (2020). For our purposes, since hate speech is produced and consumed by the public at large, annotations such as those reported by Gomez et al. have relevance, and some disagreements in them are due to disagreements that are actually present in the ground truth.

Annotator-specific data should be collected and reported as it can be used to detect spamming and error rates for use with probabilistic labeling (Hovy et al., 2013). Uma et al. (2021) noted there are many ways to compute class probabilities for soft labeling. However, most of those approaches require data about which annotator provided which label and measuring each annotator's labeling patterns across instances. Without a means of tracing annotations to specific annotators in the MMHS150K corpus, the probabilities we used for the training data input vectors are simply the probability of the instance belonging to the class given the labels that annotators assigned for the instance. Future research that sources new labels for MMHS150K might find that class probability computations that are able to incorporate additional prior information about specific annotators will lead to better classifier performance.

Another potentially useful source of hate speech labels is a large language model (LLM) such as OpenAI's GPT-3.5. In *post hoc* analyses, we used the API for GPT-3.5 (text-davinci-003) to generate labels for 10,000 MMHS150K instances. Since GPT-3.5 is by default oriented to generating varying and less repetitious responses, the model had to be tuned to concisely return one label per instance in a consistent format. After experimenting with several smaller batches to modify the prompt and tune the model, on the first pass through the 10,000 cases, the model returned 9,952 usable labels. We reran the 48 that did not get a class code with 152 others that did get a code to see how consistent GPT-3.5 is with label generation. Once again it produced labels for 99.5% of the input (199 out of 200). Not providing labels for 0.5% of the input was without explanation. There were no errors or warnings reported. Additionally, the 152 labels generated in both runs were not consistent. Cohen's *Kappa* for the first and second runs was 0.73 with the same prompt and tuning. Furthermore, the GPT-3.5 labels did not appear to be better than the crowdsourced labels. Fleiss's *Kappa* for the 9,952 labels originally generated by

GPT-3.5 and the three crowdsourced labels (all in all four labels) was a little lower than *Kappa* for the three crowdsourced labels alone (0.13 and 0.15, respectively). Although the humans originally hired for annotations by Gomez and colleagues did not agree with each other at a high rate, they agreed more with each other than they agreed with GPT-3.5. We also checked to see whether the Keras model trained on the crowdsourced labels could predict the GPT-3.5 labels. The results were better than they would have been for random data as labels, but a little worse than the results for testing with a holdout of crowdsourced labels. These analyses, while not highly rigorous, do not point to GPT-3.5 as an obvious solution to generating training data and classifying MMHS150K. Nevertheless, future research might find value in using GPT-3.5, GPT-4 or another LLM with a novel prompting strategy to improve its success.

Beyond decisions about labels in the context of annotator disagreements, other approaches to classifying MMHS150K's tweet texts could potentially improve classification accuracy. First, there are many configurations of model architectures and hyperparameter settings that might produce better results than those reported here. While we have experimented with several, further experimentation could boost classification performance. We have also not explored knowledge-based approaches or the use of customized vocabularies to skew the results because such approaches have been explored in the past (e.g., Gupta & Joshi, 2017) for other datasets with mixed results. Furthermore, like Gomez and colleagues, we used GloVe embeddings for this study, but other natural language processing approaches such as term frequency – inverse document frequency might be more effective. Additionally, although Gomez and colleagues did not find a multimodal approach using both text and images produced meaningfully higher accuracy than text alone, future research is likely to generate greater classification accuracy using a multimodal approach with the MMHS150K corpus.

6. Conclusion

This study demonstrated two approaches that can be effective at coping with noisy and subjective annotations in deep learning. Filtering a training dataset such that only cases with clean labels are used during training was shown to improve classification performance in the multi-class paradigm using the MMHS150K hate speech dataset when compared to classification with labels based on majority vote. In addition, using soft labels in training that represent annotations as probabilities proved to be the best approach in both binary and multi-class paradigms. It achieved improvements beyond both the baseline approach and the clean filtering approach.

Despite annotation disagreements, classifiers trained on MMHS150K could have practical utility for some use cases. None of the models reported here would be a good choice for standalone algorithmic hate speech blocking by an online platform. An online platform would want to use such a model only as a complement to other means of addressing hate speech. On the other hand, the models were effective enough that they could be useful in research and analysis applications. When compared to using hate speech terminology in a keyword search to collect hate speech instances for analysis, the classifiers reported here are far more efficient.

7. References

Almanea, D., & Poesio, M. (2022, June). ArMIS-The Arabic Misogyny and Sexism corpus with annotator subjective disagreements. *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France,* (pp. 2282-2291).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555-596.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology, 41*, 3-33.

Blair, J.A. (2012). What is bias? In C. Tindale (Ed.), *Groundwork in the theory of argumentation: Argumentation library, 21*. (pp. 23-32), Springer, Dordrecht. https://doi.org/10.1007/978-94-007-2363-4_3

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior, 45*, 163-172.

Boromisza-Habashi, D. (2012). The cultural foundations of denials of hate speech in Hungarian broadcast talk. *Discourse & Communication, 6*(1), 3-20.

Botelho, A., Vidgen, B., & Hale, S. A. (2021). *Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate*. arXiv. https://doi.org/10.48550/arXiv.2106.05903

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior, 58*, 101608.

Cheung, T. H., & Lam, K. M. (2022). Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing, 514*, 1-12.

Eke, C. I., Norman, A. A., & Shuib, L. (2021). Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. *IEEE Access, 9*, 48501-48518.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018, June). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1), 52-61.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys, 51*(4), 1-30.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2), 2053951720943234.

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1470-1478.

Guiora, A., & Park, E. A. (2017). Hate speech on social media. *Philosophia, 45*, 957-971.

Gupta, I., & Joshi, N. (2017, December). Tweet normalization: A knowledge based approach. *Proceedings of the 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)*, 157-162.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Haski-Leventhal, D., Pournader, M., & Leigh, J. S. (2020). Responsible management education as socialization: Business students' values, attitudes and intentions. *Journal of Business Ethics, 176*, 17–35.

Henry, J. S. (2009). Beyond free speech: novel approaches to hate on the Internet in the United States. *Information & Communications Technology Law, 18*(2), 235-251.

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

Janis, I. (1972). *Victims of groupthink: Psychological studies of policy decisions and fiascoes*. Houghton Mifflin Company.

Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *Journal of King Saud University-Computer and Information Sciences, 34*(7), 4335-4344.

Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (2020). *Intersectional bias in hate speech and abusive language datasets*. arXiv. https://doi.org/10.48550/arXiv.2005.05921

Laaksonen, S. M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020). The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data, 3*, https://doi.org /10.3389/fdata.2020.00003

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society, 7*(1), 2053951720920686.

Lommel, A., Popovic, M., & Burchardt, A. (2014, May). Assessing inter-annotator agreement for translation error annotation. *Proceedings of the Language Resources and Evaluation Conference, Reykjavik,* 31-37.

Lupu, Y., Sear, R., Velásquez, N., Leahy, R., Restrepo, N. J., Goldberg, B., & Johnson, N. F. (2023). Offline events and online hate. *PLoS one, 18*(1), e0278511.

Maity, K., Sen, T., Saha, S., & Bhattacharyya, P. (2022). MTBullyGNN: A graph neural network-based multitask framework for cyberbullying detection. *IEEE Transactions on Computational Social Systems*, http://doi.org/10.1109/TCSS.2022.3230974.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366-4383.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). *BERTweet: A pre-trained language model for English tweets*. arXiv. https://doi.org/10.48550/arXiv.2005.10200.

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *Sage Open, 10*(4), 2158244020973022.

Paul, J. (2019, February 10). FBI talk emphasizes hate speech's impact. *UWIRE Text*. https://link.gale.com/apps/doc/A573514230/AONE?u=anon~b954d713&sid=googleScholar&xid=2acd2cf.

Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors for Word Representation* [Data set]. Stanford University. https://nlp.stanford.edu/projects/glove/.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors, 19*(21), 4654.

Prasad, N., Saha, S., & Bhattacharyya, P. (2021, July). A multimodal classification of noisy hate speech using character level embedding and attention. *Proceedings of the 2021 International Joint Conference on Neural Networks*, 1-8.

Ramezan, C. A., Warner, T. A., & Maxwell, A. E. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing, 11*(2), 185.

Ravikiran, M., Muljibhai, A. E., Miyoshi, T., Ozaki, H., Koreeda, Y., & Masayuki, S. (2020). *Hitachi at SemEval-2020 Task 12: Offensive language identification with noisy labels using statistical sampling and post-processing*. arXiv. https://doi.org/10.48550/arXiv.2005.00295

Recasens, M., Hovy, E., & Martí, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua, 121*(6), 1138-1152.

Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access, 8*, 204951-204962.

Schweppe, J., & Perry, B. (2022). A continuum of hate: Delimiting the field of hate studies. *Crime, Law and Social Change, 77*(5), 503-528.

Shewart, W. A., Wilks, S. S., Fleiss, J. L., Levin, B., & Paik, M. C. (2003). The Measurement of Interrater Agreement. *Statistical Methods for Rates & Proportions, Third Edition*, 598–626.

Sai, S., Srivastava, N. D., & Sharma, Y. (2022). Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science, 3*(2), 122.

Shekarpour, S., Alshargi, F., Thirunarayan, K., Shalin, V. L., Sheth, A., & Rezvan, M. (2020). Analyzing and learning the language for different types of harassment. *PLoS ONE*, *15*(3): e0227330.

Shome, D., & Kar, T. (2021, December). ConOffense: Multi-modal multitask Contrastive learning for offensive content identification. *Proceedings of the 2021 IEEE International Conference on Big Data*, 4524-4529.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics, 126*, 157-179.

Tsesis, A. (2017). Social media accountability for terrorist propaganda. *Fordham Law Review, 86*, 605.

Uma, A., Almanea, D., & Poesio, M. (2022). Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence, 5*, 818451.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research, 72*, 1385-1470.

Wang, W., Chen, L., Thirunarayan, K., & Sheth A. P. (2014) Cursing in English on Twitter. *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, 415-425.

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology, 60*(1), 93-117.

Yao, H., Chen, Y., Ye, Q., Jin, X., & Ren, X. (2021). Refining language models with compositional explanations. *Advances in Neural Information Processing Systems, 34*, 8954-8967.

Zhang, X., Wu, X., Chen, F., Zhao, L., & Lu, C. T. (2020, April). Self-paced robust learning for leveraging clean labels in noisy data. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(4), 6853-6860.