



MASTERARBEIT | MASTER'S THESIS

Titel | Title

Multilingual Hate Speech Modeling from a Perspectivist
Viewpoint: Incorporating Inter-Annotator Disagreement into
Model Training and Evaluation

verfasst von | submitted by
Patricia-Carla Grigor BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 066 587

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Joint-Masterstudium Multilingual Technologies

Betreut von | Supervisor:

Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

Acknowledgements

A very special thank you goes to my supervisor and professor, Dagmar Gromann, who supported me throughout the process of writing this thesis, whether it was with ideas, knowledge, or tips on how to reduce my use of italics. Her lectures were one of the highlights of this master's program.

I would also like to thank my second supervisor, Petra Kralj Novak from the Central European University, for giving me the opportunity to contribute to her project and for her tireless support at every stage of the process.

Another person I couldn't have completed this thesis without is Bojan Evkoski from the Central European University, who not only shared his valuable insights and ideas with me, but also supported me with the computational resources and the coding.

My deepest gratitude goes to all of my colleagues who have taken part in this journey with me and who have always offered to help when help was needed. Completing this degree wouldn't have been half as much fun without you, and I'm glad to have made so many new friends along the way. Long live the MLT24 Support Group. Never forget: *you are allowed to think!*

Finally, I couldn't have completed this degree without the support of my family, friends, and pets. Thank you for always being there and keeping me motivated — I couldn't have done it without you.

Abstract

As the number of social media users rapidly increases, so does the amount of toxic content published on social networking sites. As a result, the Machine Learning (ML) community has begun to take an interest in the automatic detection of hate speech, which is typically approached as a classification task. While modern ML algorithms are known to provide nearly human-like results for a variety of downstream Natural Language Processing (NLP) tasks, the classification of hate speech is still an open challenge due to its problematic annotation process, which often leads to disagreement between annotators. To address this issue, perspectivist approaches depart from the gold standard by introducing methods for learning from disagreement as opposed to disregarding it. Such approaches recognize the inherent subjectivity and ambiguity of hate speech and believe that including multiple perspectives in the training pipeline of ML models helps them learn from conflicting annotations, ultimately leading to improved performance in real-world scenarios. This thesis follows such a perspectivist approach by fine-tuning multilingual language models on the hate speech detection task using diamond standard data, as well as using appropriate evaluation metrics which take disagreement into account to measure their performance. The main contributions of this research include the demonstration that these multilingual models outperform their monolingual counterparts and perform competitively with models trained on gold standard data, underscoring the effectiveness of multilingual and perspectivist approaches in the nuanced task of hate speech detection.

Zusammenfassung

Mit der raschen Zunahme der Social-Media-Nutzenden steigt auch die Menge an Hassrede in sozialen Netzwerken. Daher hat die Machine-Learning-Community begonnen, sich intensiv mit der automatischen Erkennung von Hassrede zu befassen, die meist als Klassifizierungsaufgabe behandelt wird. Moderne Machine-Learning-Algorithmen sind dafür bekannt, nahezu menschenähnliche Ergebnisse bei verschiedenen Natural-Language-Processing-Tasks zu erzielen. Die Klassifizierung von Hassrede bleibt jedoch eine Herausforderung, da der problematische Annotationsprozess oft zu Unstimmigkeiten zwischen den Annotierenden führt. Um dieses Problem zu lösen, setzen perspektivische Ansätze nicht auf den traditionellen Goldstandard, sondern führen alternative Methoden ein. Diese Methoden lernen aus den Unstimmigkeiten, anstatt sie zu ignorieren. Solche Ansätze erkennen die inhärente Subjektivität und Mehrdeutigkeit von Hassrede an und gehen davon aus, dass das Einbeziehen verschiedener Perspektiven in die Trainingspipeline von Machine-Learning-Modellen dazu führt, dass aus widersprüchlichen Annotationen gelernt werden kann, was letztlich zu besseren Ergebnissen in realen Anwendungen führt. Diese Arbeit verfolgt einen perspektivischen Ansatz, um mehrsprachige Sprachmodelle zur Erkennung von Hassrede mithilfe von Diamond Standard Data zu optimieren. Dabei werden auch geeignete Evaluationsmetriken eingesetzt, die Unstimmigkeiten berücksichtigen. Eine der wichtigsten Erkenntnisse dieser Forschung ist der Nachweis, dass die untersuchten mehrsprachigen Modelle ihre einsprachigen Gegenstücke übertreffen und mit Modellen konkurrieren, die auf Goldstandard-daten trainiert wurden. Dies unterstreicht die Effektivität mehrsprachiger und perspektivischer Ansätze bei der Erkennung von Hassrede.

Contents

List of Tables	9
List of Figures	11
1. Introduction	1
2. Background	5
2.1. Text Classification	5
2.1.1. Traditional Machine Learning Approaches	5
2.1.2. Deep Learning-Based Approaches	8
2.1.3. Evaluation Metrics	11
2.2. Transfer Learning	13
2.3. Hate Speech	14
2.3.1. Definitions of Hate Speech	14
2.3.2. Approaches to Automatic Hate Speech Detection	22
2.3.3. Challenges of Automatic Hate Speech Detection	27
2.4. Handling Annotator Disagreement	32
2.4.1. Measures of (Dis)agreement	32
2.4.2. Common Areas of Disagreement	34
2.4.3. Sources of Disagreement	35
2.4.4. Incorporating Disagreement in the Training and Evaluation Pipeline	38
3. Related Work	41
4. Method	43
4.1. Datasets	43
4.2. Model Selection	45
4.3. Fine-Tuning	46
4.4. Evaluation	48
5. Results and Analysis	53
5.1. Inter-Annotator and Model-Annotator Agreement	53
5.2. Model Comparison	54
5.3. Statistical Significance	57

Contents

5.4. Qualitative Analysis	58
6. Discussion	63
7. Conclusion	67
Bibliography	69
A. Appendix	81
A.1. Automatic Hate Speech Detection Resources and Approaches	81
A.2. Label Distribution of Monolingual Hate Speech Datasets	89

List of Tables

1.	Confusion matrix	12
2.	Definitions of hate speech as defined on the social media platforms Facebook, Twitter and YouTube	20
3.	Common areas of disagreement	35
4.	Distribution of labels across the multilingual training sets	45
5.	Distribution of labels across the monolingual evaluation sets	46
6.	Inter-Annotator Agreement (IAA) as opposed to model-annotator agreement in terms of Krippendorff’s ordinal alpha (α) and Accuracy (Acc.) for the models Multilingual Duplicate All (MDA), Multi- lingual Duplicate Disagreement (MDD), Multilingual Remove Dis- agreement (MRD) and Multilingual Duplicate All No Re-Weighting (MNR) models based on the language-specific evaluation sets	54
7.	Model evaluation and inter-annotator agreement results in terms of class-specific F1 scores on the English dataset. The Total score was calculated using the weighted F1 score. The first three models rep- resent the English (EN), Italian (IT), and Slovenian (SL) baselines. The subsequent models represent the multilingual models Multi- lingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All (No-Reweightings) (MNR).	56
8.	Model evaluation and inter-annotator agreement results in terms of class-specific F1 scores on the Italian dataset. The Total score was calculated using the weighted F1 score. The first three models rep- resent the English (EN), Italian (IT), and Slovenian (SL) baselines. The subsequent models represent the multilingual models Multi- lingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All (No-Reweightings) (MNR).	56

List of Tables

9.	Model evaluation and inter-annotator agreement results in terms of class-specific F1 scores on the Slovenian dataset. The Total score was calculated using the weighted F1 score. The first three models represent the English (EN), Italian (IT), and Slovenian (SL) baselines. The subsequent models represent the multilingual models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All (No-Reweightings) (MNR).	57
10.	Average scores calculated by averaging the weighted F1 scores of each model.	57
11.	Examples of implicit and explicit hate speech (HS) and model predictions used for the qualitative analysis	61
12.	Hate speech resources	82
13.	Approaches to automatic hate speech detection	86

List of Figures

1.	Transfer learning framework (Ruder, 2019, 43)	14
2.	The OpSaLa approach (Pelicon et al., 2023, 356)	29
3.	Images from the LabelMe dataset alongside their gold labels (Uma et al., 2021, 1388)	37
4.	Critical difference diagram according to the Nemenyi test	59
5.	Label distribution for the English dataset in Kralj Novak et al. (2022)	89
6.	Label distribution for the Italian dataset in Kralj Novak et al. (2022)	89
7.	Label distribution for the Slovenian dataset in Kralj Novak et al. (2022)	90

1. Introduction

The phenomenon of hate speech, which is typically defined as offensive or derogatory language targeting individuals or groups based on characteristics such as race, religion, ethnic origin, sexual orientation, disability, or gender, has become a significant problem on social networks in recent years, leading to an increase in the level of toxicity within online communities. Social media platforms, such as Twitter (now X), Facebook, or YouTube, are known to facilitate a rapid dissemination of this form of speech across diverse language communities, underscoring its multilingual nature and transforming it into a worldwide issue. To mitigate the spread of hate speech, the Machine Learning (ML) and computational linguistics communities have taken to developing content moderation strategies that effectively use the power of advanced algorithms and Natural Language Processing (NLP) techniques to detect this type of content (e.g. Pérez et al. 2023; Kralj Novak et al. 2022; Poletto et al. 2021). These strategies are designed not only to accurately identify hate speech, but also to adapt to the nuanced variations and contexts that differ from one language to another, ensuring broader and more effective moderation across diverse user bases. However, this has led to the identification of a second challenge, namely the subjectivity of hate speech. It is known that training reliable ML models requires extremely large amounts of data. In the case of hate speech classifiers, this is mostly annotated data, as the majority of approaches are rooted in the supervised ML paradigm. In turn, annotated data implies the inclusion of humans into the annotation process.

That being said, a number of research results have led to the conclusion that, in the case of subjective tasks, such as the annotation of hate speech, annotators often disagree when it comes to deciding on the gold label. This phenomenon can be attributed to the fact that annotators often come from different backgrounds and have different conceptions of what should be considered hateful and what should not. Consequently, in more recent publications, researchers have proposed alternative methodologies for ground truthing, including the incorporation of diverse perspectives into the training and evaluation pipelines of ML models. One such approach is introduced by Kralj Novak et al. (2022), who train monolingual hate speech classifiers in several languages directly on datasets which include disagreement. Such data can be referred to as diamond standard data, an alternative to the gold standard based on the assumption that more than one single truth exists. In terms of evaluation, the researchers focus on the evaluation of models from the

1. Introduction

perspective of disagreement, with the ultimate goal of estimating the agreement between the annotators themselves, as well as between models and annotators by using the appropriate metrics. The researchers' main findings indicate that disagreement between human annotators represents an intrinsic limitation to the performance that can be achieved through automated systems. This thesis introduces a similar approach that attempts to utilize the capabilities of pre-trained multilingual language models for the development of multilingual hate speech classifiers fine-tuned on diamond standard datasets. Therefore, at the basis of this thesis lies the following research question:

How does the performance of multilingual hate speech classifiers trained on diamond standard data compare to that of monolingual models trained under the same conditions, and how can inter-annotator disagreement effectively be incorporated into the fine-tuning process?

In light of the this research question, one of the primary motivations for this thesis is the subjective and culturally nuanced nature of hate speech, which, in turn, necessitates a rethinking of the gold standard for ML applications. Furthermore, the existing research on the topic of hate speech has demonstrated that the available datasets are severely imbalanced, which represents a significant challenge for effective classifiers. A potential solution is therefore the use of multilingual pre-trained models, which have the ability to leverage their pre-existing knowledge of languages in the resolution of classification tasks. Finally, hate speech has an extremely negative impact on online communities by disrupting healthy user interactions and creating divisive environments. In consequence, the implementation of measures to prevent and remove such content is an essential and immediate measure that must be taken in order to maintain the integrity of online community interactions.

Having outlined the motivation and objectives that have led to the writing of this thesis, the expected results are twofold. First, it is anticipated that multilingual hate speech classifiers trained on diamond standard data will outperform their monolingual counterparts in terms of accuracy and robustness, due to their ability to leverage cross-linguistic knowledge. Second, the incorporation of inter-annotator disagreement into the fine-tuning process is anticipated to enhance the models' sensitivity to nuanced hate speech, thereby facilitating the development of more reliable and robust classifiers. These findings have the potential to be beneficial to a number of communities and stakeholders. From the perspective of computational linguistics, the insights into multilingual model performance and the alternative approach to handling inter-annotator disagreement could prove beneficial in future research, not only in the context of hate speech, but more broadly in tackling the issue of annotator disagreement. Moreover, the findings presented in this thesis

could prove beneficial for social media providers, as they offer novel ideas for the development of more robust content moderation algorithms, which could ultimately contribute to maintaining the integrity of online communities.

To conclude, this thesis makes a contribution to the existing body of research on the subject of automatic hate speech detection by (1) developing multilingual solutions to detect hate speech¹ in several languages, i.e., English, Italian, and Slovenian, (2) developing solutions based on the diamond standard, which represents a relatively under-researched area, and (3) providing an extensive theoretical framework on the topics of hate speech and learning from disagreement. The code for the presented experiments is available at: <https://tinyurl.com/multilingual-hs>.

¹This thesis contains examples of hate speech and abusive language. These are used solely for the purpose of illustrating the severity of such expressions and do not reflect the opinions or beliefs of the author.

2. Background

This chapter provides the foundational basis necessary for answering the previously formulated research question. It begins with a brief introduction to the underlying task of text classification and the different algorithms and approaches used to tackle it in Section 2.1. Section 2.2 briefly discusses transfer learning, a technique often used to improve the performance of language models on specific tasks, such as hate speech detection. Section 2.3 offers an introduction to the topic of hate speech and its automatic detection using various approaches. Finally, Section 2.4 introduces the topic of annotator disagreement and discusses methods for incorporating it into the development of automated systems.

2.1. Text Classification

Text classification is a fundamental task in NLP that involves categorizing text into predefined classes and serves as the basis for a number of tasks, including automatic hate speech detection. The process underlying text classification algorithms employs various machine learning techniques to analyze and categorize text data, thereby enabling systems to automatically identify and categorize content based on its linguistic characteristics. This section provides a concise overview of models and techniques used for the task of text classification, starting with traditional ML approaches in Subsection 2.1.1. Subsection 2.1.2 discusses more recent methods that have advanced text classification algorithms by leveraging deep neural networks, with a particular focus on the Transformer architecture (Vaswani et al., 2017) and its application in state-of-the-art language models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and its multilingual variant, Cross-lingual Language Model - RoBERTa (XLM-R) (Conneau et al., 2019).

2.1.1. Traditional Machine Learning Approaches

ML is a subfield of Artificial Intelligence (AI) that focuses on the development and study of systems capable of learning from data and generalizing to unseen data, thereby performing tasks without explicit instructions. ML systems may be grouped into three primary paradigms, based on the type of learning scenario they address (Goodfellow et al., 2016). The first paradigm is supervised learning,

2. Background

where models trained on labeled datasets learn relationships between inputs and outputs, which enables them to predict outcomes and recognize patterns in new, unseen data. The second paradigm is unsupervised learning, where models are trained to identify patterns and intrinsic structures within unlabeled data. Finally, the third paradigm is self-supervised learning, which is a recent and increasingly influential ML paradigm based on learning the supervision signal dynamically from unlabeled data, allowing models to learn robust representations from large amounts of unlabeled text.

The task that underlies automatic hate speech detection is text classification. Text classification is a task that is predominantly categorized under the supervised learning paradigm of ML and involves the categorization of text into one or more predefined categories. In the context of automatic hate speech detection, the most commonly employed traditional text classification models include Naïve Bayes (NB), logistic regression, Support Vector Machines (SVM) (Cortes and Vapnik, 1995), and decision trees, as will be discussed in Subsection 2.3.2.

Naïve Bayes. NB classifiers are probabilistic models operating under Bayes’ Theorem with a strong (naïve) assumption of independence among features, assuming that each feature, e.g. a word, is independent of the others, given the class label (Jurafsky and Martin, 2023). In this context, the equation for Bayes’ Theorem is articulated as depicted in Equation 1.

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c)P(c)}{P(\mathbf{x})} \quad (1)$$

In this equation, $P(c \mid \mathbf{x})$ is the posterior probability of class c given predictor vector $\mathbf{x}$, $P(c)$ is the prior probability of class c , $P(\mathbf{x} \mid c)$ is the likelihood which is the probability of the predictor given the class, and $P(\mathbf{x})$ is the prior probability of the predictor.

Logistic Regression. Logistic regression is a statistical method which extends traditional linear regression by modeling the probability that a given input belongs to a particular category (Jurafsky and Martin, 2023). The core of logistic regression is the logistic function, which is used to model the probability p that an instance $\mathbf{x}$ belongs to the default class. This probability can be expressed as in Equation 2.

$$p(\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}} \quad (2)$$

In this equation, $\beta_0, \beta_1, \dots, \beta_n$ are the coefficients of the model, and e is the base of the natural logarithm. This function ensures that the output of the logistic

regression model lies between 0 and 1, making it suitable for estimating probabilities.

Support Vector Machines. SVMs (Cortes and Vapnik, 1995) operate on the principle of finding a hyperplane that optimally separates different classes in a high-dimensional space. Their objective is to maximize the margin between the nearest data points of any class and the decision boundary, which is the hyperplane. Mathematically, this is achieved by solving the optimization problem depicted in Equation 3.

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \quad (3)$$

In this equation, $\mathbf{w}$ is the weight vector, b is the bias, ξ_i are the slack variables allowing for misclassification of difficult or noisy examples, C is the regularization parameter, $\mathbf{x}_i$ are the training examples, and y_i are the labels.

To handle non-linearly separable data, SVMs use kernel methods. Kernel methods are algorithms for pattern analysis, which involves identifying relationships between data points in a given dataset. By employing a kernel function, SVMs can operate in a high-dimensional, implicit feature space without the need of computing the coordinates of the data in that space. This allows SVMs to perform non-linear classification with high efficiency. There are several types of kernels that can be used, including linear, polynomial, and radial basis functions. The choice of the kernel function can have a significant impact on the performance of the SVM.

Decision Trees. Decision trees are hierarchical models that build on tree-like structures for decision support, where each node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. The decision to split at each node is made based on the criterion that best separates the classes, typically measured by impurity metrics such as Gini impurity or entropy. The general form of these metrics for a classification tree can be expressed as in Equations 4 and 5.

$$\text{Entropy}(S) = - \sum_{i=1}^c p_i \log_2 p_i \quad (4)$$

$$\text{Gini impurity}(S) = 1 - \sum_{i=1}^c p_i^2 \quad (5)$$

2. Background

In these equations, p_i is the proportion of the samples that belong to class i in a given subset S . It is important to note that decision trees also serve as the basis for more advanced ensemble methods such as random forests and gradient boosting machines. These methods build on the concept of decision trees to create more robust and complex predictive models.

2.1.2. Deep Learning-Based Approaches

Deep Learning (DL) is a subset of ML that utilizes artificial neural networks with multiple layers and the backpropagation of gradients to model complex patterns in data, incorporating non-linear transformations at each layer. These networks are capable of learning hierarchical representations of data, which is particularly advantageous for tasks that involve large and complex datasets, such as those used for the training of hate speech classifiers. Consequently, DL distinguishes itself from traditional ML approaches primarily through its ability to automatically extract and learn representations from data. In contrast to traditional ML techniques, which often rely on handcrafted features that require domain expertise and extensive pre-processing, DL-based models can learn to identify useful features directly from the data itself. This capability stems from the use of multiple layers in neural networks, each layer building on its predecessor to progressively extract higher-level features from raw input, as well as the use of non-linear activation functions for the individual layers. This learning process enables DL models to handle complex and high-dimensional datasets more effectively, often outperforming traditional ML approaches when applied on the same tasks. This subsection provides a brief introduction to the most common DL-based architectures used for text classification, with a focus on state-of-the-art language models.

Convolutional Neural Networks. Convolutional Neural Networks (CNNs) are a class of deep neural networks originally developed by LeCun et al. (1998) for computer vision tasks and have since been adapted to a wide range of domains, including NLP. In NLP, CNNs are utilized to perform text classification by leveraging convolutional layers to capture local features within text, such as n-grams, without the need for manual feature engineering. Instead, local feature detection is achieved through the application of multiple filters that perform convolutions over the input data, effectively extracting significant features, such as word or phrase patterns essential for understanding textual context. That said, the typical CNN architecture for NLP includes several convolutional layers followed by pooling layers, which serve to reduce dimensionality and highlight dominant features, thus forming robust feature maps. These feature maps are subsequently interpreted by dense layers that perform the final classification in relation to specific tasks.

Recurrent Neural Networks. Introduced in their modern form by Rumelhart et al. (1986), Recurrent Neural Networks (RNNs) are a specialized class of neural networks particularly well-suited for sequential data processing, such as time series or natural language. In NLP, RNNs are primarily implemented for text classification tasks due to their architecture, which allows them to maintain a memory of previous inputs by looping output back into inputs. This enables RNNs to process and make predictions based on the entire sequence of data, capturing temporal dependencies and contextual relationships within text. Their architecture includes a loop that reuses the same weights with each time step, allowing for information persistence, a critical feature for tasks like sentiment analysis or topic detection, where understanding context and sequence is vital. However, traditional RNNs often struggle with long-range dependencies due to the vanishing gradient problem, where gradients used in the network’s training process can disappear over time, making learning ineffective for distant data points. This issue was notably addressed with the introduction of Long Short-Term Memory (LSTM) networks by Hochreiter and Schmidhuber (1997), which significantly improve the network’s ability to remember information over longer periods. Further advancements in RNN architecture include the development of Gated Recurrent Units (GRUs) by Cho et al. (2014). GRUs simplify the LSTM design by combining the forget and input gates into a single update gate and merging the cell state and hidden state. This modification not only reduces model complexity, but also helps in mitigating the vanishing gradient problem, making GRUs efficient and effective for many sequence modeling tasks.

The Attention Mechanism. The attention mechanism, initially proposed by Bahdanau et al. (2014) in the context of neural machine translation, is an advanced computational technique employed in neural networks to dynamically focus on specific elements within the input data. This mechanism represents a significant advancement over preceding methodologies, such as CNNs and RNNs, by enhancing the model’s ability to process and interpret complex and lengthy data sequences. This is accomplished by selectively prioritizing which parts of the input sequence are most relevant at different stages of processing through the computation of attention weights. Subsequently, these weights are used to compute a weighted sum of features at different positions. Consequently, the attention mechanism can be defined as in Equation 6.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (6)$$

In this equation, Q , K , and V represent the query, key, and value matrices which are derived from the input data. Specifically, a query represents a specific word in

2. Background

the input sequence, a key is the set of all words in the input sequence, and a value is also a representation of all words in the input sequence. The attention scores are computed by taking the dot product of the query with all keys, dividing each by the square root of the dimensionality of the keys ($\sqrt{d_k}$) to stabilize the training process, and applying a softmax function to ensure the scores sum to one. The output is then a weighted sum of the values, where the weights are the attention scores. This process enables the model to allocate attention differentially across the sequence and to focus more on relevant parts for predicting each element in the output sequence.

Transformer-Based Models. Transformer-based models make use of attention mechanisms, in particular self-attention, to enhance the capabilities of language modeling. Self-attention is a key innovation within the Transformer architecture that allows each token in the input sequence to be evaluated in the context of all other tokens, rather than solely in relation to its adjacent ones. This way, the mechanism computes attention weights that reflect the relevance of every other token to a given token, facilitating a comprehensive contextual understanding across the entire sequence. Furthermore, the Transformer architecture supports parallel processing, which accelerates training and inference compared to traditional recurrent neural network approaches. The original Transformer model, as introduced in Vaswani et al. (2017), is an encoder-decoder architecture comprising multiple layers of multi-head attention and feedforward networks. Multi-head attention allows the model to focus on different parts of the input simultaneously, enhancing its ability to capture various aspects of the data. Each layer within the encoder and decoder has several heads, each learning different representations of the input data, thus improving the model’s overall understanding. Consequently, the capacity to process all tokens simultaneously allows Transformer-based models to efficiently manage dependencies and nuances in language data, which is crucial for complex NLP tasks. In contrast, subsequent models often specialize by adopting either the encoder-only structure, as seen in BERT (Devlin et al., 2018), or the decoder-only structure, as in GPT models (Radford et al., 2019).

BERT A state-of-the-art language model based on the Transformer architecture is BERT, developed by Devlin et al. (2018). Unlike previous approaches that process text in a single direction, BERT considers both left and right contexts simultaneously in all layers, which is referred to as bidirectional training. This context awareness significantly improves the model’s performance on downstream tasks, such as named entity recognition, sentiment analysis, and question answering, by providing a deeper, more nuanced semantic representation of the input text. Furthermore, BERT is pre-trained using a combination of masked language modeling

and next sentence prediction tasks. In masked language modeling, random words in a sentence are replaced by a [MASK] token, and the model is trained to predict the original word based only on its context. The next sentence prediction task trains the model to predict whether one sentence logically follows another, ultimately helping it understand how sentences relate to each other. In consequence, this type of pre-training helps the model acquire a rich representation of language before being fine-tuned for specific downstream tasks with smaller datasets, such as hate speech datasets, which is often sufficient to achieve state-of-the-art results due to the model’s transfer learning capabilities.

XLM-R Building on the foundational methodologies established by BERT, the XLM-R model introduced by Conneau et al. (2019) extends the bidirectional training framework to a multilingual context, thus effectively enhancing the model’s capabilities across different languages. Based on the RoBERTa architecture (Liu et al., 2019), which itself refines BERT’s pre-training methods by removing the next-sentence prediction task and intensifying the masking strategy, the XLM-R model is pre-trained on a corpus consisting of texts in more than 100 different languages. This extensive pre-training is designed to give the model a deep, nuanced understanding of syntactic, semantic, and contextual differences across languages, enabling it to tackle complex cross-lingual classification tasks.

2.1.3. Evaluation Metrics

After defining text classifiers and the various approaches used for their implementation, this section presents a brief overview of the most common evaluation metrics used to assess their performance. A starting point from which subsequent evaluation metrics can be derived is the confusion matrix. The confusion matrix is a tabular representation of the actual versus predicted classifications made by a model (Jurafsky and Martin, 2023). It provides a comprehensive view of the performance of a classification model by displaying the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), as shown in Table 1. More specifically, TP represent the instances correctly classified as positive; FP represent the number of instances incorrectly classified as positive; TN represent the number of instances correctly classified as negative; FN represent the number of instances incorrectly classified as negative.

An initial metric which can be derived from the confusion matrix is precision (Jurafsky and Martin, 2023). Precision measures the accuracy of the positive predictions made by a model. It is defined as the ratio of TP to the total number of positive predictions and can be calculated using Equation 7.

2. Background

Table 1.: Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

A second metric is recall, also known as sensitivity. Recall measures a model’s ability to correctly identify all actual positive instances. It is defined as the ratio of TP to the total number of actual positives, as shown in Equation 8.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

The third metric is accuracy (Jurafsky and Martin, 2023). Accuracy measures the overall correctness of the model by calculating the ratio of correctly classified instances to the total number of instances, as shown in Equation 9. However, accuracy may not always be the most reliable measure of model performance, particularly in cases of class imbalance, where high accuracy may be misleading if the model predicts predominantly the majority class (Kralj Novak et al., 2022).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

Finally, an alternative to accuracy is the F1 score (Jurafsky and Martin, 2023). The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it particularly useful in cases of imbalanced class distributions. (Kralj Novak et al., 2022). The general form of the F1 score is shown in Equation 10.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

Moreover, it is important to note that there are several variants of the F1 score, each addressing different aspects of model performance. The first one of these is the micro-averaged F1 score, which calculates a global F1 score by counting the sums of the TP, FN, and FP. This is achieved by first summing the respective TP, FN, and FP values across all classes and then applying the F1 score formula. The second variant of the F1 score is the macro-averaged F1 score. This variant calculates the F1 score for each class independently and then takes the average, thereby treating all classes equally regardless of their support. In this context,

support refers to the number of actual instances of each class in the dataset. This variant provides a measure of performance across classes by considering the balance among them. The final variant is the weighted F1 score. The weighted F1 score addresses imbalances in class distribution by incorporating the number of instances of each class into the calculation. This is achieved by first multiplying the F1 score of each class by its support. Subsequently, the final weighted F1 score is calculated by averaging the resulting adjusted F1 scores. This approach ensures that classes with a greater number of instances have a proportionately greater influence on the overall metric, thus allowing for a more balanced evaluation in the context of unbalanced datasets.

2.2. Transfer Learning

Having established that pre-trained language models can effectively be utilized for the development of hate speech classifiers through transfer learning, it is important to discuss what this method entails. Transfer learning is an ML method which involves adapting a model developed for a general task to a specific, more narrowly defined task (Pan and Yang, 2009). Alternatively, it can also be defined as the transfer of knowledge acquired by a model trained on a source task, domain, or language(s) to a new, target task, domain, or language(s) (see Figure 1). In the context of language models, transfer learning typically consists of two phases: the pre-training and the fine-tuning phase. In the pre-training phase, a language model is trained on a large, diverse corpus of text, thus enabling the model to gain a comprehensive understanding of general linguistic patterns. This results in pre-trained language models, such as BERT (Devlin et al., 2018) or XLM-R (Conneau et al., 2019). In the subsequent fine-tuning phase, such models are fine-tuned to a specific task, for example hate speech detection, by utilizing task-specific datasets. In most cases, such datasets are significantly smaller and are annotated with task-specific labels. During the fine-tuning process, the weights of the pre-trained model are adjusted in order to better align with the patterns of the target task, with the aim of enhancing the model’s predictive accuracy on this task. It must be noted, however, that fine-tuning is by no means the only transfer learning approach, but rather one of many that are typically used in tasks such as hate speech classification, where models need to adapt to specific nuances and context of the language used. Other transfer learning methods include, for example, feature extraction, where a pre-trained model is used to extract general features that are then fed into another model, or domain adaptation, where a model is adjusted to perform well across different but related domains. In conclusion, transfer learning is an effective ML method used for improving the performance of models on specialized tasks without the prohibitive costs and time associated with

2. Background

training complex models from scratch.

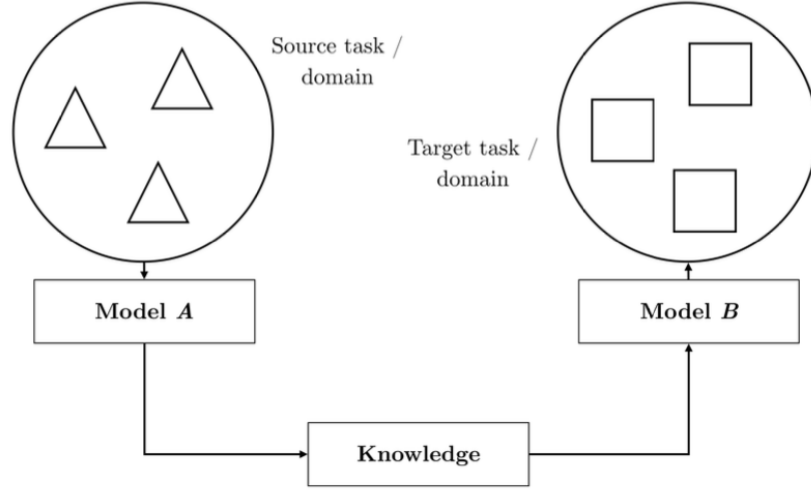


Figure 1.: Transfer learning framework (Ruder, 2019, 43)

2.3. Hate Speech

Performing the task of automatic hate speech detection requires a fundamental understanding of the concept of hate speech itself, as well as a thorough exploration of the multifaceted challenges it poses to automated systems. This chapter provides an overview of the task at hand and its underlying theoretical concepts, starting with an examination of the concept of hate speech based on linguistic and philosophical conceptual analysis in Subsection 2.3.1. This is followed by a discussion of different approaches to automated hate speech detection and existing resources in Subsection 2.3.2. Finally, Subsection 2.3.3 concludes by reviewing a set of challenges encountered in the implementation of hate speech detection systems, as identified in contemporary research.

2.3.1. Definitions of Hate Speech

To better grasp the concept of hate speech, it is important to analyze it from a linguistic perspective. Language is known to be the primary tool for creating and altering interpersonal relationships, and words can be utilized as a means of triggering negative reactions when formulated as offensive statements (O’Driscoll, 2020). Depending on their severity, reactions to such offences can vary in type and intensity, ranging from feelings of humiliation to threats to a recipient’s sense

of security. Consequently, language that has the potential to cause offence or has caused offence in a particular situation falls under the category of offensive language. O’Driscoll (2020, 16) describes it as:

any word or string of words which has or can have a negative impact on the sense of self and/or wellbeing of those who encounter it – that is, it makes or can make them feel, mildly or extremely, discomfited and/or insulted and/or hurt and/or frightened

Additionally, offensive language can be divided into two subcategories: taboo language, which is language that may cause offence, and offending utterances, which consist of language that has caused offence in a particular scenario (O’Driscoll, 2020).

Taboo Language. According to O’Driscoll (2020), taboo language refers to specific linguistic sequences that are abstracted from the context of time, place, and participants. The author argues that the offensiveness of such sequences, e.g. swear words, is dependent on the value attached to them within a particular linguistic community, rather than any inherent offensiveness in the words themselves. Therefore, defining this type of language requires consideration of the prevailing values of the society in which it is produced. The proposed definition of taboo language is:

any (string of) words whose production is transgressive of polite social norms (O’Driscoll, 2020, 5)

Offending Utterances. In contrast, O’Driscoll (2020) uses the term offending utterances when referring to actual offences. He defines these as statements that are experienced by their recipient(s) as being:

deleterious to (one or more aspects of) their sense of self and/or wellbeing – that is, it makes them feel, mildly or extremely, discomfited and/or insulted and/or hurt and/or frightened (O’Driscoll, 2020, 5)

On the basis of this classification, it can be argued that hate speech falls under the category of offending utterances, due to its inherent nature of being an attack targeting individuals or collectives with the sole purpose of inciting negative reactions. However, Brown (2017) suggests that rather than having a single, precise definition, hate speech is a phenomenon that is subject to multiple interpretations. In his analysis of the term, the author distinguishes between the legal and the broader conceptualizations of hate speech. He notes that the legal definition is

2. Background

narrow and primarily addresses „incitement to hatred“ (2017: 595). Conversely, the more common understanding of the term, which is often used by the general public, is found to be ambiguous and should be regarded as a family resemblance, as coined by Wittgenstein (1953) (Brown, 2017). In an attempt to grasp the latter and explore its multitude of emerging definitions, Brown (2017) proposes four methods of philosophical conceptual analysis: purpose-oriented analysis, folk platitudes analysis, analysis of cases based on intuition, and ordinary language analysis.

As regards the purpose-oriented analysis, its aim is to examine how the term is being used by the general public and the purposes it serves. Five of its primary purposes have been identified (Brown, 2017, 569):

- Hate speech highlights forms of harmful speech that disproportionately harm already disadvantaged or victimized members of society.
- Hate speech points out socially disruptive forms of speech.
- Hate speech identifies forms of speech that can challenge people’s sense of equality.
- Hate speech articulates norms of civility.
- Hate speech labels forms of speech that undermine democracy.

In addition, the term may be used more broadly to express disapproval of certain forms of speech. It can therefore be understood as a pejorative term that aims to condemn and identify speech that deserves censure. However, such an understanding of hate speech is problematic because, as Brown (2017) explains, it can be misused by undemocratic or authoritarian governments, which could harm society by restricting legitimate forms of public discourse and self-expression. On the contrary, it is well known that any attempt to legally censor or restrict hate speech, even in democratic systems or online platforms, is always in opposition to free speech rights.

The analysis of folk platitudes concentrates on common generalizations of the term. Brown (2017, 574) identifies the following four:

- The term hate speech carries a negative connotation as it implies a normative judgement.
- Hate speech is speech that targets individuals based on certain characteristics, such as race, ethnicity, religion, sexual orientation, gender identity, and disability.

2.3. Hate Speech

- Hate speech is an emotionally charged term and can incite hatred.
- Hate speech is a form of speech that can be subject to regulation and does not have the protection of an individual's right to freedom of expression.

It is important to recognize that the boundaries of what constitutes hate speech are not limited to traditional categories, such as those previously outlined. Any speech that deliberately targets aspects of an individual's identity, including their name, nickname, or visual representation in online formats, as well as direct or indirect familial descent, may also be considered hate speech. This broader interpretation acknowledges the diverse and evolving ways in which identity can be constructed and targeted in various contexts.

The analysis of cases based on intuition examines whether given examples are primarily instances of hate speech based on linguistic cues. Following this approach, Brown (2017, 581) identifies five broad types of hate speech:

- Any form of insult or words of disparaging abuse or vilification targeting members of groups or classes of people identified by protected characteristics.
- Any expressions suggesting moral inferiority, non-humanity, or alienness of historically oppressed or victimized groups or groups identified by protected characteristics.
- Any form of group defamation, such as damaging statements that aim to diminish the reputation of individuals who belong to groups or classes of people identified by protected characteristics.
- Negative stereotypes or generalizations about any group or class of people identified by protected characteristics.
- Language that incites, stirs up, or advocates hatred, discrimination, or violence against individuals or groups based on protected characteristics, or that threatens violence against such groups, or that justifies or glorifies discrimination or violence against them.

It should be noted that while general characteristics may refer to any notable attributes, protected characteristics are legally recognized categories that safeguard individuals from discrimination. These typically include age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation¹ (The National Archives, 2010).

¹The characteristics listed are derived from the Equality Act 2010 (The National Archives, 2010), which applies specifically to the UK; these protections may not be directly applicable in other international contexts.

2. Background

To perform ordinary language analysis, Brown (2017) draws on Austin’s speech act theory (Austin, 1975), which categorizes speech into three acts: (1) locutionary, the actual utterance and its literal meaning; (2) illocutionary, the intended significance of the utterance; and (3) perlocutionary, the effect the utterance has on the listener. He emphasizes that hate speech can be described as a multitude of illocutionary acts, referring to the communicative force that an utterance has in a given context. Examples for this are speeches that perpetuate or facilitate acts of subordination or oppression, speeches that actually constitute acts of subordination or oppression, as well as illocutions that persecute and degrade. In a similar fashion, O’Driscoll (2020) argues that, while potentially insulting language is a matter of locution, actual insulting language can be understood as a matter of perlocution, influenced by the insulted person’s perception of the illocutionary force of the insult.

Family Resemblance. Although analytical approaches may be useful in outlining the general concept of hate speech, the term remains ambiguous due to the multiplicity of definitions available. To address this issue, Brown (2017) suggests that the concept of hate speech can be understood as a family resemblance, drawing on an idea introduced by Wittgenstein in 1953. Wittgenstein (1953) proposed the theory of family resemblances, using games as an example to demonstrate that similar entities in a category might not share a single common feature but are connected through a series of overlapping similarities. Following his line of argumentation, Brown (2017) applies this concept to hate speech, illustrating that although individuals may not agree on a precise definition, they can competently use the term because they recognize these overlapping characteristics. This perspective suggests that definitions of hate speech lack a central common feature and instead rely on family resemblances, where everything that falls under a given concept shares at least one similar characteristic with at least one other instance under the same concept. Additionally, he emphasizes that, unlike the more narrowly defined legal term, the general term hate speech has multiple definitions, resulting in a complex array of overlapping similarities and internal diversity. Moreover, the concept of hate speech can take various forms, such as written or spoken words, symbols, gestures, or pictures, thereby encompassing a diverse range of phenomena that share family resemblances.

Apart from being a concept of linguistic and legal interest, hate speech can be more concretely described as a real-life phenomenon, which nowadays strongly unfolds on online social networks (Kralj Novak et al., 2022; Arango et al., 2019). That being said, the plurality of meanings behind the term remains unchanged even in the online medium, where it appears under forms such as toxic comments, abusive language, discrimination, cyberbullying, or flaming (Chhabra and Vishwakarma, 2023). Other publications broaden such definitions by stating that online

hate speech is directed towards individuals or groups protected by national or international law, and can even lead to the glorification of war crimes and crimes against humanity (Kralj Novak et al., 2022; Pérez et al., 2023). Thus, hate speech negatively affects the dynamics of online social communities (Arango et al., 2019). To counteract and limit this phenomenon, several internet companies employ community guidelines which are to be followed by the platform’s users. Such guidelines provide concrete definitions for hate speech with the ultimate goal of categorizing specific comments as such and removing them (Brown, 2017). Table 2 shows definitions of hate speech collected from three major social media platforms: Facebook², Twitter³, and YouTube⁴. An analysis of these definitions reveals that, although similar in some respects, each platform employs a nuanced approach to defining hate speech, reflecting its distinct user base and regulatory history. Facebook’s definition is notably detailed, encompassing a wide range of protected characteristics, including caste and serious disease, and explicitly categorizing forms of hate speech to include violent or dehumanizing speech and harmful stereotypes. Such a comprehensive understanding of the term could be indicative of an attempt to provide clear, actionable guidelines within its diverse global community. While the definitions provided by Twitter⁵ shares many protected characteristics with Facebook, it is less precise, omitting specific examples like dehumanizing comparisons. This lack of detail could compromise the effectiveness of content moderation algorithms and influence the clarity with which users understand and adhere to the platform’s policies. In contrast, the definition provided by YouTube distinguishes itself by including unique categories such as immigration and veteran status, thereby addressing broader societal issues that may be overlooked by other platforms. Consequently, there is considerable variation in the way hate speech is defined on such platforms, once again highlighting the heterogeneous nature of the term. This not only illustrates the diversity of policies shaped by specific incidents and user interactions, but also demonstrates the difficulty of effectively moderating online environments.

Finally, this chapter would be incomplete without a definition of hate speech from a computational linguistics perspective. In the field of computational linguistics, the detection of hate speech is most commonly treated as a classification task, as will be discussed in Subsection 2.3.2. Therefore, establishing what hateful text constitutes is extremely important for enabling the classification of text instances

²<https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>

³<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

⁴https://support.google.com/youtube/answer/2801939?ref_topic=9282436

⁵Twitter has been officially renamed to X. This thesis continues to use the name Twitter and to refer to postings as tweets throughout, as this was the official name at the time the data for the experimental part was collected, making it the most appropriate and clear reference for the analysis conducted.

2. Background

Table 2.: Definitions of hate speech as defined on the social media platforms Facebook, Twitter and YouTube

Platform	Hate Speech Definition
Facebook	We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanising comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence. [...] (Facebook, 2024)
Twitter	You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. [...] (Twitter, 2024)
YouTube	Hate speech is not allowed on YouTube. We don't allow content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube's policy: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status (YouTube, 2024)

as such. A review of relevant literature on this topic reveals that the majority of definitions align with the ones published on social media platforms, emphasizing that hate speech is an insult directed towards individuals or groups of people based on certain characteristics. Depending on the country, such characteristics may be referred to as protected characteristics. A definition of hate speech that is frequently encountered in approaches to automatic hate speech detection is:

any communication that disparages a person or a group on the basis of some characteristic, such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic (Nockleby (2000), as cited in Warner and Hirschberg (2012, 1))

Alternatively, Davidson et al. (2017, 1) differentiate between offensive language and hate speech, defining the former as „language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group“. Another common topic in computational linguistics publications on hate speech is the distinction between explicit and implicit forms of hate speech (Yin and Zubiaga, 2021). Explicit hate speech consists of expressions that are typically unambiguous and require minimal interpretation (Waseem and Hovy, 2016) since explicit linguistic markers, i.e., words that can be interpreted as hate speech are used. Forms of speech that fall under this category include, but are not limited to, the following: (1) racial slurs, (2) sexual orientation and gender slurs, (3) profanity, (4) religious bigotry, and (5) ethnic slurs.

- (1) „@eBeZa: Stupid f*cking n*gger LeBron. You flipping jungle bunny monkey f*ggot.“ (Davidson et al., 2017, 514)
- (2) „My advice of the day: If your a tranny...go f*ck your self!“ (Davidson et al., 2017)
- (3) „Somebody told me that Little Debbie likes to take it up the A.\$\$.“ (Nobata et al., 2016, 149)
- (4) „Don’t try to explain—Irish Catholics are just idiots“ (Fortuna and Nunes, 2018, 6)
- (5) „The JEWS are Market Manipulators Hate them freaking Jews !! To bad Hitler did not finish them off !!!“ (Nobata et al., 2016, 149)

In contrast, implicit hate speech is more nuanced and frequently manifested through indirect expressions of hatred. Such instances may necessitate an understanding of contextual and cultural nuances to discern them, as they frequently rely on insinuations, stereotypes, or coded language (Ocampo et al., 2023; Wiegand et al., 2021). Examples of implicit hate speech include, among others: (1) sarcasm, (2) stereotypes, (3) dehumanization, (4) comparisons, and (5) rhetorical questions.

- (1) „It’s always fun watching sports with a woman in the room.“ (Wiegand et al., 2021, 580)
- (2) „Women should all stay at home in an apron, chained to the oven!“ (Wiegand et al., 2021, 578)
- (3) „Headscarfed muslims waddle around our streets all over.“ (Wiegand et al., 2021, 579)
- (4) „You sing like a dying bird.“ (Wiegand et al., 2021, 579)

2. Background

(5) „Did Stevie Wonder choose these “models”?“ (Wiegand et al., 2021, 580)

To conclude, the term hate speech has been investigated across several academic disciplines, resulting in a diverse range of possible definitions rather than a single, unified one. This underscores not only the inherent complexity of the term, but also the fundamentally subjective nature of classifying any form of linguistic expression. The following chapters further illustrate this subjectivity. It is important to note that this thesis primarily examines the phenomenon of hate speech as it manifests on online social networks, i.e., hate speech appearing in the form of comments or tweets designed to attack individuals or groups based on protected characteristics.

2.3.2. Approaches to Automatic Hate Speech Detection

Having defined hate speech, the following subsections explore its importance in the NLP community by discussing existing types of resources necessary for the training of hate speech classifiers, as well as presenting current methods for its automatic detection.

As previously discussed, hate speech is a phenomenon that rapidly spreads on social media platforms, particularly with the number of users constantly increasing (Poletto et al., 2021). Consequently, the topic has received considerable attention in the field of AI, as AI-driven solutions, especially NLP-based approaches, are considered to be a viable solution to restrict and regulate the spread of hate speech within online platforms (Arango et al., 2019; Poletto et al., 2021). The idea behind such solutions is to leverage the capabilities of ML and NLP for creating effective large-scale applications for content moderation, which automatically identify and remove content deemed to fall under the category of hateful speech. Due to their high potential, there has been a notable surge in advancements in automatic hate speech detection approaches in recent years. However, despite significant improvements, the implementation of hate speech classifiers remains a challenge for several reasons, as Poletto et al. (2021) explain. Firstly, identifying hateful speech is a complex and context-dependent task due to the evolving and adaptive nature of hate speech. Secondly, detecting such content relies not only on linguistic elements but also on cultural contexts, demanding a grasp of linguistic variations and cultural nuances. Thirdly, the subjective nature of hate speech makes annotation a particularly challenging task in itself. Finally, training hate speech classifiers is highly susceptible to biases, often leading to overfitting and poor generalization of models to new data. However, these challenges underscore the importance of research in this field. The ultimate goal is to develop applications that not only generalize better but also support multiple languages, acknowledging hate speech as a global issue.

Existing Resources. Like any task involving training ML-based or neural models, reliable hate speech classifiers require large amounts of data. Therefore, it is important to first discuss the available types of resources for this task, as found in existing literature. Poletto et al. (2021) introduce five different dimensions for categorising resources for automatic hate speech detection.

The first dimension pertains to the resource type. In their literature review on existing resources for automatic hate speech detection, the researchers draw a key conclusion that the most common resources are annotated hate speech datasets (e.g. Albadi et al. 2018), suggesting that most approaches in this area are based on supervised learning. Moreover, some of the discovered datasets are part of publicly available benchmarks, allowing researchers to test their systems and compare results in open scientific competitions, such as SemEval Task 5 on multilingual detection of hate speech against immigrants and women in tweets (Basile et al., 2019). In addition, the analysis also uncovered several lexica (e.g. Davidson et al. 2017; Wiegand et al. 2018), which are resources consisting of words and phrases related to a specific semantic field, in this case, hate speech.

A second approach to categorization focuses on classifying hate speech resources by their topical foci, all of which are considered sub-categories of hate speech. Given that the perception of the term is highly subjective and its meanings are heterogeneous, the choice of thematic foci may be influenced by the targets of hate speech, the reach of the concept, the cultural aspects involved, the language, individual perception, and so on. Consequently, the imprecise boundaries and ambiguous areas between hate speech definitions have resulted in many approaches moving beyond the binary classification of hate and non-hate (e.g. Vidgen and Yasseri 2020; Kralj Novak et al. 2022). Instead, several existing approaches investigate more detailed classifications, taking into account factors such as the severity, targets, cultural nuances, the time period the speech appeared in, or specific events that the speech relates to. That said, examples of possible methods for identifying topical foci in contemporary works include analysing rhetoric devices, such as aggressive language (e.g. Sanguinetti et al. 2018), threats (e.g. Ljubešić et al. 2022), slurs (e.g. Nobata et al. 2016), and sarcasm (e.g. Sanguinetti et al. 2018). These approaches can be used to identify explicit forms of hate speech, such as profanity, offensive or violent language, as well as implicit forms, such as irony, sarcasm, or stereotypes. Another method is to focus on the groups of people being targeted, which can lead to different subcategories of hateful speech, such as racism, sexism, homophobia, or anti-semitism.

Another distinction can be made between the sources of hateful speech. Most approaches to automatic hate speech detection focus on hate speech that appears on online social media platforms (Poletto et al., 2021). Therefore, the data is typically sourced from platforms, such as Twitter (e.g. Pérez et al. 2023), Reddit

2. Background

(e.g. Olteanu et al. 2018), Facebook (e.g. Mandl et al. 2019), Instagram (e.g. Corazza et al. 2019) or YouTube (e.g. Kralj Novak et al. 2022), and consists of user-generated content that has been made publicly available on one of these platforms. Retrieval methods for such data often consist of keyword-based search algorithms. Some approaches use more complex strategies, such as retrieving entire timelines from users, threads, or videos that are considered hateful (e.g. Pérez et al. 2023; Ljubešić et al. 2022). Such approaches allow for more contextual information to be retrieved, which becomes beneficial in the annotation process.

In terms of the annotation scheme employed, Poletto et al. (2021) identify three principal strategies: the binary scheme, the non-binary scheme and multi-level annotation. Approaches that use the binary scheme focus on determining whether sentences in the dataset are hateful or not. This scheme consists of two mutually exclusive values, for example [homophobic] and [not homophobic] (Akhtar et al., 2019). The second type of scheme allows for a more fine-grained classification based on several aspects, as discussed before. For example, some approaches focus on the level of hate present in an instance, e.g. [strong hate], [weak hate], [no hate] (Vigna et al., 2017); [acceptable], [inappropriate], [offensive], [violent] (Kralj Novak et al., 2022). Alternatively, other approaches focus on the different phenomena that can fall under the concept of hate speech, e.g. [hate speech], [aggressiveness], [offensiveness], [irony], [stereotype] (Sanguinetti et al., 2018); [racism], [sexism], [both], [neither] (Waseem and Hovy, 2016). Multi-level annotation schemes are the most complex. They consist of fine-grained schemes that focus on several phenomena by using multiple layers of annotation. For instance, Nobata et al. (2016) categorize instances first into [clean] and [abusive] language. Then, the latter category can be further divided into [hate speech], [derogatory], and [profane] speech.

Another important aspect to consider in this dimension of categorization is the type of annotators involved in the annotation process. Poletto et al. (2021) identify three main types of annotators: experts, non-experts, or automatic classifiers. In the first case, the experts are either the developers themselves or people with knowledge in the field. In the second case, non-professional annotators, often crowdsourced workers or students, are either employed or recruited as volunteers for the annotation process. In the third case, no human annotators are involved. As Poletto et al. (2021) report, several publications have adopted combinations of the three annotation strategies discussed.

Finally, resources for automatic hate speech detection can be categorized according to the language of the data. English is the primary language for scholars, particularly in technical fields, hence it is not surprising that most of the existing resources for automatic hate speech detection are in English (Poletto et al., 2021). Another trend observed by Poletto et al. (2021) is the emergence of multilingual

hate speech datasets (e.g. Chung et al. 2019; Steinberger et al. 2017), which include not only high-resource languages, such as English (e.g. Chung et al. 2019), but also lower resourced languages, such as Slovenian (e.g. Kralj Novak et al. 2022), Bengali (e.g. Ishmam and Sharmin 2019), or Swedish (e.g. Fernquist et al. 2019). It is worth noting that the continuous development of multilingual datasets for hate speech detection is extremely important, as combating hate speech is of global concern. In this regard, one publicly available catalog⁶ of hate speech datasets in different languages published by Vidgen and Derczynski (2020) is worth mentioning.

For better visualization, all publications given as examples in this subsection are listed in tabular form along with their main characteristics, such as resource type, source of the data, annotation scheme employed, labels, types of annotators, and language(s) in Table 12 in Appendix A.1.

Existing Approaches. After discussing various resources used for the development of automatic approaches to hate speech detection, this subsection summarizes the prevailing approaches and models in this field, as identified in existing literature. It starts with traditional, feature-based ML techniques employed in the early 2010s, which mark the initial formulation of the task (Yin and Zubiaga, 2021, 2), and advances to the most recent state-of-the-art approaches based on deep neural networks and LLMs.

Most existing approaches to automatic hate speech detection are supervised classification tasks, i.e., a model that is trained in a supervised scenario is expected to learn connections between the features and the actual predictions present in the training data, with the ultimate goal of generalising well to new, unseen datasets (Yin and Zubiaga, 2021). That said, the first hate speech classifiers date back to the early 2010s, one of the very first implementations being the SVM developed by Warner and Hirschberg (2012). Since back then, state-of-the-art models were based on traditional, statistical ML approaches, several classifiers based on algorithms such as SVM (e.g. Burnap and Williams 2014; Davidson et al. 2017), Naïve Bayes (e.g. Kwok and Wang 2013; Liu and Forss 2014), logistic regression (e.g. Djuric et al. 2015; Davidson et al. 2017), or decision trees (e.g. Burnap and Williams 2014, 2015; Agarwal and Sureka 2017) followed. While the model architectures differ, one thing all these models have in common is the fact that they are based on hand-crafted features. In their review, Schmidt and Wiegand (2017) list types of features typically used for traditional ML-based hate speech classifiers: character-level features and n-grams, often used in combination with other features (Warner and Hirschberg, 2012; Burnap and Williams, 2015; Nobata et al., 2016); word clusters induced using, for example, Brown clustering (Warner and Hirschberg, 2012) or Latent Dirichlet Allocation (Xiang et al., 2012); incorporating information stemming from

⁶<https://hatespeechdata.com>

2. Background

sentiment analysis algorithms (Dinakar et al., 2012); lexicons consisting of negative words, such as slurs or insults (Nobata et al., 2016; Burnap and Williams, 2015); linguistic features such as dependency relationships (Chen et al., 2012); information stemming from knowledge bases (Dinakar et al., 2012); meta-information, such as background information about users (Waseem and Hovy, 2016); multimodal information, such as images, video, or audio content which accompanies hateful social media posts (Hosseinmardi et al., 2015).

Another methodology worth mentioning is the implementation of ensemble methods, which are based on the idea of applying the wisdom of the crowd (Mullah and Zainon, 2021, 5). Ensemble methods consist of aggregated classifiers that often perform better than individual machine learning algorithms since they combine the advantages of several models to improve performance on a specific task. In the case of hate speech classifiers, the most common ensemble method approaches are random forests (Agarwal and Sureka, 2017; Burnap and Williams, 2015; Xiang et al., 2012), as well as bagging or boosting methods (Mullah and Zainon, 2021). That being said, ensembles are not limited to traditional ML approaches and can be used in combination with DL-based models. For example, Paschalides et al. (2020) implement a hate speech detection model using an ensemble method of combining CNNs and RNNs.

More recently, as deep neural networks have revolutionized the field of ML, a number of researchers have started developing DL-based hate speech classifiers (Mullah and Zainon, 2021; Alkomah and Ma, 2022). For example, several publications report the use of CNN architectures in combination with features such as word embeddings, TF-IDF vectors, or n-grams (Paschalides et al., 2020; Zimmerman et al., 2018; Zhang et al., 2018; Gambäck and Sikdar, 2017), all of which show an improvement in terms of quality over traditional ML methods, such as the ones previously listed. Similarly, RNN-based architectures, for example RNNs (Sharma et al., 2017), LSTMs (Ni and Cao, 2020), Bi-LSTMs (Badjatiya et al., 2017; Vigna et al., 2017) and GRUs (Mossie and Wang, 2020) have reportedly led to significant advances for hate speech classification without any explicit linguistic or other features. Moreover, a number of publications present combinations of such architectures. For example, Zhang et al. (2018) implement a CNN-GRU model, while Badjatiya et al. (2017) train a gradient-boosted decision tree using word embeddings produced by an LSTM model.

The currently most dominant architecture in DL is the Transformer. Transformer-based models, such as BERT (Devlin et al., 2018), have set a new standard for many ML tasks. This is also true for hate speech classification, as shown by leaderboards of shared tasks, such as OffensEval 2020 (Zampieri et al., 2020) or EVALITA2020 (Fersini et al., 2020). As a result, most current approaches to hate speech classification are focused on the fine-tuning of LLMs, such as variants of

the BERT model, using hate speech datasets. For example, Koufakou et al. (2020) fine-tune the baseline BERT model using lexicon-based lexical features, Fortuna et al. (2021) test the generalization capabilities of BERT and ALBERT (Lan et al., 2019), Glavaš et al. (2020) compare the performance of BERT, ROBERTA (Liu et al., 2019), mBERT (Devlin et al., 2018), and XLM-R (Conneau et al., 2019) models across multiple domains and languages.

In summary, most existing approaches to hate speech classification treat it as a supervised classification task. The focus on solving this task has shifted over the years from using traditional ML classifiers, such as SVMs or random forests, to more complex approaches that utilize deep neural networks to learn patterns more effectively from the training data. While CNN and RNN architectures were considered the preferred approaches in previous years, current research shows that the Transformer architecture is regarded as the current state-of-the-art approach to hate speech classification. For a more detailed view of all the publications listed as examples in this subsection, see Table 13 in Appendix A.1.

2.3.3. Challenges of Automatic Hate Speech Detection

Being a relatively new task within the NLP research community, and given the heterogeneous nature of hate speech, the task of automatic hate speech classification faces several challenges. This subsection outlines some of the most common ones, from data collection to model validation, as well as possible mitigating strategies.

Dataset Imbalance. Despite the continuous spread of hate speech in online communities, several existing datasets used for training hate speech classifiers, particularly those involving pre-trained models, are severely imbalanced, with hateful instances representing the minority class(es) (Pelicon et al., 2023; Kralj Novak et al., 2022; Arango et al., 2019). This is particularly evident in datasets with finer-grained, non-binary classifications, such as the [Violent] class in Kralj Novak et al. (2022), which are often the most underrepresented. It is important to note that a critical aspect of this imbalance issue is not solely attributable to the limitations of the datasets themselves, but also to the patterns inherent in the majority of available data. These patterns, which LLMs typically uncover and replicate, serve to exacerbate the issue. Consequently, classifiers, especially those based on pre-trained models, may be biased towards predicting the majority class(es), potentially overlooking more severe instances of hate speech. This imbalance often leads to suboptimal performance in real-world scenarios, where models encounter novel hate speech instances that differ from the constrained examples they were trained on. Therefore, current research suggests that the hypothesis of training models on datasets that reflect the natural distribution of the data needs to be

2. Background

re-evaluated in cases of extreme class imbalance (Pelicon et al., 2023). Following this line of argumentation, Pelicon et al. (2023) describe various approaches for handling dataset imbalance.

Firstly, cost-sensitive re-weighting is a method which manipulates the loss function during the training of neural ML models by increasing the misclassification cost of instances belonging to the minority class(es). This causes the model in question to attend more to instances belonging to the underrepresented class(es), potentially improving its performance, without compromising the overall accuracy. A common re-weighting technique is allocating weights that are inversely proportional to the frequency of each class (Pelicon et al., 2023; Huang et al., 2016).

Another common mitigating strategy is data resampling. Two main methods are oversampling the minority class(es) and undersampling the majority class(es). Oversampling is a technique used to balance out class distribution by augmenting the minority class(es) through replicating existing instances. This allows classifiers to improve learning from classes that were previously underrepresented by increasing their instance count. However, there is a risk of overfitting as models may memorize the replicated examples instead of learning their underlying distribution. Conversely, undersampling involves reducing the size of the majority class(es) by eliminating instances until the class distribution becomes more balanced. According to Pelicon et al. (2023), the latter method has proven to be most efficient. Nevertheless, it is not applicable to datasets that are extremely skewed, as it would result in the loss of significant amounts of data. This, in turn, could impair the model’s ability to capture the full complexity of the majority class(es).

Thirdly, several publications report using data augmentation techniques to balance out skewed datasets. This approach utilizes methods of artificially expanding the dataset by creating modified versions of existing instances belonging to the minority class(es). Pelicon et al. (2023) provide a list of data augmentation techniques, including back-translation, substituting specific words with synonyms from WordNet (Fellbaum, 1998) or their nearest neighbours, and using language models, such as BERT to generate artificial examples.

Finally, Pelicon et al. (2023) introduce Optimized data Selection before Labeling (OpSaLa), their novel three-step approach for optimizing the data sampling process prior to labeling (see Figure 2), proposed as an alternative to the aforementioned mitigation strategies. The first step involves creating a large unlabeled corpus by collecting more data than the annotation budget allows, i.e., only a proportion can be manually annotated. In the second step, external sources are used to identify and collect instances that are more likely to belong to the underrepresented class(es). Sources can range from simple keyword lists or dictionaries to more complex methods, such as probabilistic binary classifiers. The third step involves randomly sampling the test set to maintain the natural class distribution, allowing

for an objective evaluation of models. The advantage of this approach is that the external sources used in the second step allow for the creation of a minority class-rich training dataset using real examples instead of artificial ones. However, it is important to note that such an approach is only applicable in cases where large amounts of unlabeled data are available. Having conducted comparative studies, Pelicon et al. (2023) demonstrate that the use of the OpSaLa approach leads to enhanced model performance for transformer-based hate speech classifiers when dealing with dataset imbalance. More precisely, it enhances the performance of minority classes without compromising the overall performance of the model. Nevertheless, it must be emphasized that this is not the only viable solution and that it may not be applicable to all cases of hate speech.

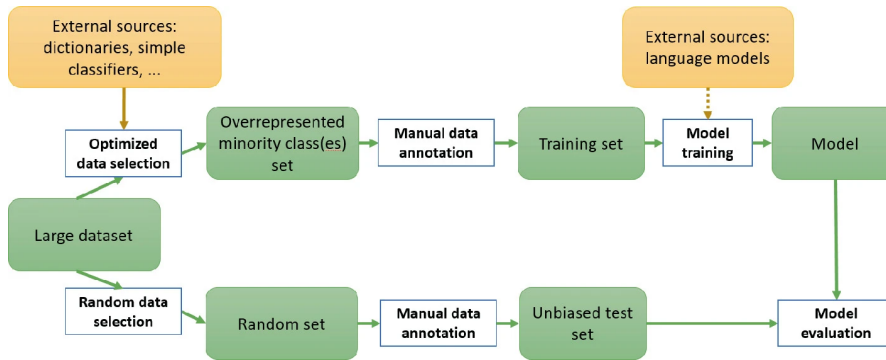


Figure 2.: The OpSaLa approach (Pelicon et al., 2023, 356)

To sum up, dataset imbalance is an issue which might affect the generalization ability of models to real-world data, hence it should be tackled through techniques, such as re-weighting, resampling, data augmentation, or optimization of the data selection process.

Hate Speech Annotation. Another major challenge for the automatic hate speech detection task is the annotation of hate speech datasets, which is most commonly done at sentence level (Kralj Novak et al., 2022; Pérez et al., 2023; Ljubešić et al., 2022). The reason for this is the fact that annotators often disagree, especially in the case of highly subjective tasks, such as the annotation of hate speech. This happens either due to differences in thinking patterns, misunderstandings or badly specified annotation schemes (Kralj Novak et al., 2022; Uma et al., 2021). Either way, disagreement between annotators can lead to ambiguous results, which in turn affect the performance of automatic classifiers. This has led to the emergence of different approaches for tackling disagreement.

On the one hand, several publications attempt to increase inter-annotator agreement and thus improve the quality of annotations by providing annotators with

2. Background

additional context for the instances to be annotated (Pérez et al., 2023; Ljubešić et al., 2022). For example, Pérez et al. (2023) provide their annotators, who follow a multi-level annotation scheme, with Twitter replies as well as the articles posted by the Twitter accounts of news outlets to which they refer. This way, annotators can make use of the full context when deciding whether a reply is hateful or not, and when classifying it into one of the following categories: [RACISM], [APPEARANCE], [CRIME], [POLITICS], [WOMEN], [CLASS], [LGBTI], [DISABLED]. In a similar manner, Ljubešić et al. (2022) explore the impact of providing their annotators with context for the tweets and YouTube comments they are annotating, concluding that in-context annotation not only results in higher inter-annotator agreement, but also leads to higher levels of toxicity. Thus, their research also demonstrates that annotators are more likely to misinterpret hateful comments as appropriate ones when lacking context. This suggests that providing more context may be beneficial for the annotation of implicit forms of hate speech.

On the other hand, some publications consider disagreement between annotators to be a valuable source of information for training ML models (Kralj Novak et al., 2022; Uma et al., 2021). These approaches move away from the gold standard and argue that it idealizes how humans perceive the world, while in reality, disagreement is more natural (Kralj Novak et al., 2022). In such cases, disagreeing labels are kept throughout the entire ML pipeline, leveraging different techniques for training and evaluating automatic classifiers. Section 2.4 discusses how disagreement can be leveraged for the training of models in more detail.

To summarize, annotating hate speech can be highly subjective, often leading to disagreements between annotators. This can be addressed by providing additional information to annotators or by leveraging the disagreement to train hate speech classifiers.

Generalization to New Data. Although many hate speech classifiers are claimed to perform almost perfectly, cross-dataset testing techniques demonstrate otherwise. In reality, high performance scores on a specific dataset do not necessarily indicate real predictive accuracy, but rather an over-optimization for the characteristics of that particular dataset. Thus, when applied to different datasets, several models that have been considered state-of-the-art in recent years tend to generalize poorly (Yin and Zubiaga, 2021; Arango et al., 2019). This is considered a problem in current research, as it often leads to the overestimation of hate speech classifiers, whereas the ultimate goal of mitigating the harm of online hate speech remains unresolved. In their research, Yin and Zubiaga (2021) list a number of reasons that lead to over-optimization and propose solutions for the implementation of more robust solutions.

First of all, one main obstacle to generalization is the presence of non-standard

grammar and vocabulary in the language used on social networks, and hence present in hate speech datasets. Not only do social media users often use alternative spelling and abbreviations to save space or symbols, e.g. emojis, to express their feelings, but also to evade content moderation. The latter has led to the use of obfuscated spelling and code words, such as *ni9 9er*, *hoopiuglyniggerratgolberg*, *JOOZ* (Nobata et al., 2016, 5). Such language poses significant challenges for traditional ML-based classifiers and early word embeddings approaches, which often consider obfuscated words to be out-of-vocabulary or fail to grasp the context-dependent hateful meanings of code words. These methods typically rely on exact word matches or limited contextual understanding, making them ill-equipped to interpret the nuanced and dynamically changing use of language on social media. To address this issue, Yin and Zubiaga (2021) identify several solutions in existing research. Some publications report attempts to identify and create lists of commonly used code words to enrich hate speech lexicons used for model training (e.g. Taylor et al. 2017; Magu and Luo 2018). Other research directions focus on the use of character-level features to deal with obfuscated language (e.g. Van Aken et al. 2018). Sentence embeddings can solve the out-of-vocabulary words problem (e.g. Indurthi et al. 2019). In contrast, more recent approaches based on LLMs, which integrate broader context and deeper linguistic patterns, are better suited to understanding linguistic complexities, such as those previously mentioned. However, they still face challenges with highly inventive language use. In such cases, the most viable solution to improving model performance is fine-tuning pre-trained models on domain-specific data, for example hate speech, social media, or sentiment analysis datasets (e.g. Vidgen and Derczynski 2020).

Another factor that leads to overfitting of hate speech models is the size of the training datasets. Yin and Zubiaga (2021) argue that existing hate speech resources are still limited in size due to the novelty of the field. Seeing as most state-of-the-art hate speech classifiers are based on LLMs, mitigation strategies include fine-tuning pre-trained models on multiple hate speech datasets (e.g. Caselli et al. 2020a), or performing transfer learning to transfer knowledge from closely related fields, such as sentiment analysis (e.g. Cao et al. 2020).

Implicit expressions of hate pose another significant challenge to automatic hate speech classifiers. As discussed at the beginning of this chapter, in addition to explicit forms, hate speech can take on implicit forms, such as stereotypes, sarcasm, or irony (Yin and Zubiaga, 2021). As many existing models struggle to identify such implicit patterns, proposed solutions range from incorporating context into the training pipeline (e.g. Pérez et al. 2023) to developing annotation schemes that differentiate between implicit and explicit hate to create clearer distinctions (e.g. Caselli et al. 2020b). A different approach is presented by Sap et al. (2019): in a sequence-to-sequence task, their model is trained not only to classify hateful content,

2. Background

but also to generate paraphrased statements of offensive instances. Furthermore, Ocampo et al. (2023) enhance the performance of pre-trained hate speech models, such as HateBERT (Caselli et al., 2020a), on implicit hate speech instances by retraining them using adversarial examples. Adversarial examples are strategically modified inputs designed to expose and exploit vulnerabilities in ML models, thereby inducing misclassification. This type of learning is beneficial as it improves the robustness and generalization capabilities of the models, making them more effective in real-world scenarios.

In addition, Arango et al. (2019) discuss a further phenomenon observed in their research: when analysing the Waseem and Hovy (2016) dataset, a commonly used dataset for automatic hate speech detection, they find that most hateful tweets are produced by the same few users. This suggests that models trained on this dataset may be learning to identify users as a secondary task. To address this issue, the researchers suggest enriching the dataset with external data and restricting the number of tweets per user.

In summary, hate speech classifiers can become overly specialized to their training data, leading to an overstatement of their performance in real-world scenarios. To address this issue, adopting more realistic evaluation methods, such as cross-dataset testing, is crucial for accurately assessing the generalization ability of models. As for improving it, strategies to handle non-standard language, dataset size, implicit forms of hate speech or user bias, such as the ones discussed in this subsection, can be taken into consideration to develop more robust solutions.

2.4. Handling Annotator Disagreement

While the majority of supervised learning tasks are based on the assumption that a single, ground truth exists, some recent research directions depart from this premise to investigate scenarios in which human annotators disagree. This chapter addresses the topic of disagreement by presenting measures used to quantify it in Subsection 2.4.1, identifying tasks where disagreement can typically be observed in Subsection 2.4.2, listing potential sources of disagreement in Subsection 2.4.3, and discussing methods of leveraging disagreement in the training and evaluation of automated systems in Subsection 2.4.4.

2.4.1. Measures of (Dis)agreement

Inter-annotator agreement, often referred to as inter-rater reliability in related literature, is a metric used to measure the extent to which different annotators give consistent ratings to the same items. This metric not only reflects the complexity of a task for humans but also sets an upper limit for model performance on the

2.4. Handling Annotator Disagreement

same task (Basile et al., 2020). To measure it, various statistical measures can be employed. This subsection discusses three of the most commonly used ones: *Cohen’s Kappa* (Cohen, 1960), *Fleiss’ kappa* (Fleiss, 1971), and *Krippendorff’s alpha* (Krippendorff, 2018).

Cohen’s Kappa. Cohen’s kappa is a statistical measure that quantifies the degree of agreement between two raters who each classify a set of items into mutually exclusive categories, adjusting for agreement that could occur by chance. This measure is particularly valuable when assessing the reliability of categorical data in scenarios where subjective judgments might lead to variability (Cohen, 1960). The equation for Cohen’s kappa (κ) is presented in Equation 11.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (11)$$

In this equation, κ represents the Cohen’s kappa coefficient, P_o is the observed agreement between the raters, and P_e is the expected agreement by chance. The formula calculates the agreement adjusted for the agreement that might occur just by chance, providing a more accurate measure of the actual agreement between the raters on a scale from -1 (perfect disagreement) to 1 (perfect agreement), where 0 indicates the level of agreement that could be expected from random chance.

Fleiss’ Kappa. Fleiss’ kappa is an extension of Cohen’s kappa for assessing the reliability of agreement between multiple raters, rather than just two. It is particularly useful when several raters evaluate the same set of items into mutually exclusive categories (Fleiss, 1971). The equation for Fleiss’ kappa (κ) is similar in structure to Cohen’s kappa, but adapted to multiple raters, as depicted in Equation 12.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (12)$$

In this equation, $\bar{P}$ is the average proportion of raters who agree on an item, and $\bar{P}_e$ is the hypothetical probability of chance agreement using a multinomial distribution.

Krippendorff’s Alpha. Krippendorff’s alpha is a flexible statistical measure designed to assess the reliability of data annotated by multiple raters that supports various scales of measurement, including nominal, ordinal, interval, and ratio levels. Additionally, it is particularly effective for handling missing values, as it can accurately compute reliability estimates without requiring complete data entries from all raters (Krippendorff, 2018). The general equation for Krippendorff’s alpha (α) is depicted in Equation 13.

2. Background

$$\alpha = 1 - \frac{O}{E} \quad \text{where} \quad O = \sum_{d=1}^D o_d \quad \text{and} \quad E = \sum_{d=1}^D e_d \quad (13)$$

In this equation, O represents the total observed disagreement among the raters, and E indicates the total disagreement expected by chance. The observed disagreement o_d quantifies how much raters disagree on a particular unit of analysis, while the expected disagreement e_d reflects how much disagreement would be anticipated if the ratings were purely random. The index d encompasses all possible pairs of units being evaluated, providing a comprehensive assessment of agreement across all units and categories of analysis. When annotators are in complete agreement, the α value equals one. If the level of agreement is no better than what would be expected by chance, the α value would be close to zero. When there is systematic disagreement among the raters, the α value would be below zero (Kralj Novak et al., 2022).

This formulation highlights the significance of Krippendorff’s alpha in evaluating the extent to which raters agree beyond what might be expected by chance. It thus serves as a significant measure of reliability in the annotation process within research methodologies involving multiple annotators.

2.4.2. Common Areas of Disagreement

Several publications in the fields of NLP and Computer Vision (CV) argue that the existence of a singular label representing the ground truth in classification tasks, the gold standard, is an idealization (Basile et al., 2020; Kralj Novak et al., 2022; Uma et al., 2021). In reality, it is often the case that people tasked with annotating datasets, regardless of their level of experience, disagree for a variety of reasons. That said, previous research was focused on addressing such disagreements, which were regarded as noise, by employing aggregation strategies, such as the majority vote, to reach a consensus on the gold label. However, more recent studies have begun to recognise the value of incorporating certain types of disagreement into the training of automatic classifiers, particularly in the case of inherently subjective tasks, where the gold label represents merely a subjective interpretation over all the available alternatives (Basile et al., 2020; Uma et al., 2021).

This has resulted in the creation of datasets and methodologies for learning from disagreement in numerous sub-fields of NLP and CV. Table 3 summarizes examples based on the literature review conducted by Uma et al. (2021). As can be observed, disagreement between annotators is a common phenomenon in NLP tasks, such as part-of-speech tagging, information status classification, relation extraction, and natural language inference. According to the publications that have addressed these tasks, disagreement primarily occurs due to the intrinsic ambiguity of certain

2.4. Handling Annotator Disagreement

linguistic expressions. Nevertheless, these types of disagreements are still based on objective judgements made by the annotators. This is also the case in image classification, where disagreements are said to stem either from vaguely defined annotation schemes or difficulty in recognizing the subjects to be annotated due to low resolution. In contrast, disagreements stemming from subjective tasks, such as sentiment analysis or the classification of offensive language, are typically attributed to differences in the annotators’ personal perceptions. Focusing on the annotation of hate speech, several publications (e.g. Kralj Novak et al. 2022; Basile et al. 2020; Akhtar et al. 2019) demonstrate that people often have different perceptions of what is to be considered offensive and what is not. An example for this is the following tweet taken from the dataset introduced by Waseem and Hovy (2016). While some annotators perceived this tweet as sexist, others did not.

@ XXX uh... did you watch the video? one of the women talked about how it's assumed she's angry because she's latina. (Akhtar et al., 2019, 599)

In consequence, due to their inherently subjective nature, tasks such as these are said to pose the biggest challenge to the concept of the single, ground truth, since admitting to one would be a purely arbitrary decision (Uma et al., 2021).

Table 3.: Common areas of disagreement

Field	Task	Publication(s)
NLP	Part-of-speech tagging	Plank et al. (2014)
NLP	Information status classification	Poesio et al. (2019)
NLP	Relation extraction	Dumitrache et al. (2018)
NLP	Natural language inference	Zaenen et al. (2005)
NLP	Sentiment analysis	Kenyon-Dean et al. (2018)
NLP	Hate speech annotation	Basile et al. (2020); Kralj Novak et al. (2022)
CV	Image classification	Russell et al. (2008); Peterson et al. (2019)

2.4.3. Sources of Disagreement

After a discussion of the areas where disagreement among annotators is commonly encountered, it is essential to identify the reasons behind it in order to determine its potential benefits for the training of ML models. Although the preceding subsection briefly addressed some sources of disagreement, it is crucial to acknowledge that this phenomenon can be influenced by a multitude of factors, ranging from annotator

2. Background

proficiency and task complexity to subjective perceptions and cultural contexts. This complexity underscores that disagreement is not merely a challenge to be overcome, but rather a dynamic aspect of human annotation that can offer valuable insights into the underlying data and the interpretative flexibility of language itself. The following paragraphs discuss commonly encountered reasons for disagreement in NLP and CV tasks, based on the survey conducted by Uma et al. (2021).

An initial source of disagreement are errors or slips made by annotators during the annotation process. These errors often result from lapses in attention, misinterpretation of the annotation guidelines, or simple human error. They do not provide insights into the task itself and, as a consequence, are not useful for the training of models. Nevertheless, frequent annotator errors can serve as an indicator of the reliability of the annotators. If certain annotators consistently make more errors than others, this could indicate issues with their training, understanding, or suitability for the specific tasks they are assigned. Additionally, a high rate of errors across multiple annotators might highlight problems with the annotation interface itself. Issues such as poor interface design, confusing user interactions, or inadequate feedback mechanisms can significantly contribute to annotator errors by making the task more difficult than necessary. For instance, in Poesio et al. (2019), the interface that falsely marked portions of the texts to be annotated led to disagreements between expert annotators. Consequently, while annotator errors are not informative about the task itself, analysing them can provide valuable information about the effectiveness of the annotation process and the necessity for improvements in training, guidelines, or interface design.

Similarly, poorly designed annotation schemes or guidelines may result in disagreements between annotators (Uma et al., 2021). This phenomenon is particularly prevalent in CV tasks, such as image classification, where the classes in question are often poorly defined, overlap, or lack mutual exclusivity. Consequently, the selection of each annotator is likely to be arbitrary. An illustrative example for this can be seen in Figure 3, which is taken from the LabelMe dataset (Russell et al., 2008). As can be observed, the gold standard labels [inside city], [street] and [tall building] are not mutually exclusive, which in turn leads to the annotators having to make an arbitrary decision.

That being said, numerous studies have demonstrated that the inherent ambiguity and complexity of items to be annotated represent a significant source of disagreement in NLP and CV tasks (Uma et al., 2021). The former is often encountered in NLP tasks such as the ones outlined in Table 3. For instance, in Plank et al. (2014), expert annotators re-annotated the POS dataset published by Gimpel et al. (2011) and found multiple instances to be linguistically debatable, referring to them as hard cases. Particularly, most disagreements stemmed from the annotation of sequences as either [NOUN] or [PRON], which has proven to be highly linguistically



Figure 3.: Images from the LabelMe dataset alongside their gold labels (Uma et al., 2021, 1388)

ambiguous. In a similar fashion, linguistic ambiguity has been identified as one of the sources of disagreement in anaphoric annotation (Poesio et al., 2019) and relation extraction (Dumitrache et al., 2018) tasks. Moving on to item complexity, Uma et al. (2021) argue that it differs from ambiguity in that, while ambiguous items lack a clear interpretation due to the context they appeared in, difficult items either lack an interpretable meaning altogether or their interpretation may be extremely challenging to identify. This is often the case in CV image classification tasks, particularly in cases when the images are distorted or their resolution is so low that the subjects to be annotated are extremely difficult to identify, ultimately leading to disagreements between annotators (Peterson et al., 2019). Furthermore, item difficulty also led to disagreements in the study conducted by Zaenen et al. (2005) on natural language inference. In this study, difficult items required extra-contextual information, such as world knowledge, in order to be correctly classified. Nevertheless, the entailment relations in question could only be either true or false, thus not ambiguous.

Finally, a major source of disagreement in tasks like sentiment analysis or hate speech detection is subjectivity. In such tasks, human judgment is said to be inherently influenced by the individual perceptions of the judges themselves, rather than by the linguistic phenomena being evaluated (Basile et al., 2020, 34). For example, the annotation of hate speech is highly dependent on the individual perceptions of offensiveness, as demonstrated in the example in Subsection 2.4.3. Annotators themselves often come from a variety of backgrounds, and their cultural, social, and personal experiences can significantly influence their interpretations of what constitutes hate speech. Consequently, the perception of offensive language may vary considerably, with one annotator perceiving a statement as culturally typical speech, while another may perceive it as hate speech (Basile et al., 2020). Furthermore, differing levels of familiarity with specific linguistic or social conventions can influence a person’s perception of offensiveness, resulting in discrepancies in the identification and categorization of hate speech (Kralj Novak et al., 2022).

2. Background

Consequently, disagreements arising from subjective tasks present a significant challenge in the development of effective automated solutions, as discussed in the following sections.

2.4.4. Incorporating Disagreement in the Training and Evaluation Pipeline

Given that disagreement can serve as an indicator for the ambiguity or subjectivity inherent in certain tasks, as well as the reliability of annotators, a number of research directions (e.g. Kralj Novak et al. 2022; Fleisig et al. 2023; Fornaciari et al. 2021) integrate it within the training and evaluation pipelines of automated classifiers. Such approaches can broadly be classified into two categories: approaches based on the gold standard and approaches which deviate from the concept of a singular truth. Some approaches belonging to the latter category introduce alternative standards in lieu of the gold standard, one notable example being the diamond standard. This subsection provides a concise review of these categories, discussing strategies for incorporating disagreement into the training and evaluation of ML models.

Gold Standard Approaches. One common method for addressing annotator disagreement, particularly in cases where datasets have been annotated by crowdworkers, is the aggregation of multiple annotations into a single, ground truth label (Uma et al., 2021). This is typically achieved through the use of strategies such as majority voting, thereby eliminating the necessity for manual adjudication or expert judgments in determining the gold label. While such aggregation methods are effective in generating what is sometimes referred to as silver labels according to Uma et al. (2021), it is important to note that they do not account for the varying difficulty of the instances to be annotated, nor differences in the reliability of annotators. Consequently, these factors could potentially lead to inaccurate labels. More advanced alternatives to the majority vote include probabilistic (Dawid and Skene, 1979), heuristic, and metrics-based aggregation methods (Sheshadri and Lease, 2013; Aroyo and Welty, 2015), which offer a more nuanced understanding of label accuracy by taking into account different parameters, such as item difficulty or annotator reliability.

An alternative category of approaches based on the gold standard involves leveraging the level of annotator disagreement to filter or weight data used for the training of ML models (Uma et al., 2021). The idea behind such methods is to identify items with substantial disagreement, suggesting that they are either too difficult or ambiguous. Consequently, these approaches propose that such items should be excluded or down-weighted during training (Reidsma and op den Akker,

2008). Alternatively, Reidsma and op den Akker (2008) suggest a second approach which involves the training of a classifier for each of the annotators, followed by a voting classifier that makes a prediction only when there is unanimous agreement among all classifiers on a certain class label. Both these methods have shown high precision but low recall during evaluation. In a similar vein, Klebanov and Beigman (2014) develop classifiers to distinguish between easy and hard items, considering the latter as potential noise that could hinder effective model training, and therefore excluding these instances from the training set.

Alternative Approaches. While the preceding category of methods does make use of disagreement, they do not actively include it in the training and evaluation pipeline of ML models. In contrast, several research directions that are focused precisely on training models that learn directly from the disagreement between annotators exist. An initial approach proposed by Kralj Novak et al. (2022) utilizes so-called diamond standard data for the training and evaluation of hate speech classifiers. The diamond standard contrasts with the traditional gold standard by recognizing that more than one truth can exist. This makes it particularly useful in the context of subjective tasks, such as the annotation of hate speech, where consensus among annotators on the optimal gold label is often lacking. Considered a perspectivist approach to ground truthing, the diamond standard allows multiple labels to be maintained in the training pipeline (Kralj Novak et al., 2022). That said, the objective of the methodology presented in Kralj Novak et al. (2022) is to enhance the robustness of models by using data that replicates the complexity and variability inherent in human judgment. Additionally, in the final model evaluation step, measures commonly used to assess agreement among human annotators, such as Krippendorff’s alpha, the F1 score, and accuracy, are also applied to evaluate the alignment between model classifications and the manually annotated data, thereby simulating model evaluation through the lens of disagreement.

An alternative approach involves aggregating annotations into a probabilistic distribution, also referred to as soft labels. This technique involves learning directly from these distributions via loss functions such as cross-entropy or mean squared error. Peterson et al. (2019) have demonstrated that in image classification tasks, utilizing cross-entropy to compare soft labels with model-predicted probabilistic distributions, achieves superior generalization to new data compared to traditional gold label-based methods. Similarly, Uma et al. (2020) have applied the same approach to POS tagging, achieving comparable results. In this approach, the researchers assess model performance using both traditional hard evaluation metrics, i.e., accuracy, and soft evaluation metrics, i.e., cross-entropy. They conclude that although the soft label approach does not outperform gold standard methods when evaluated using hard metrics, it more accurately reflects the label distribution

2. Background

produced by annotators.

Furthermore, approaches that integrate both hard and soft labels into the training of models exist. One such approach is proposed by Lalor et al. (2017), who use both hard and soft labels at different times when training models on the natural language inference task. For instance, they propose a model that was trained using hard labels in one epoch and soft labels in the next, as well as a model that was trained only using hard labels and then fine-tuned using soft labels. During evaluation, both approaches demonstrated enhanced generalisation performance when compared to baselines trained exclusively on gold labels on the same task.

In a more recent approach, Fornaciari et al. (2021) employ multi-task learning to incorporate annotator disagreement in the training of neural models. Using multi-task frameworks, separate models are trained on the POS tagging and morphological stemming tasks. Each model is trained to predict gold labels as its main task, while concurrently addressing an auxiliary task of predicting soft label distributions. These distributions represent the probability distributions over the labels provided by annotators. As a result, the incorporation of soft label distributions enables each model to capture the disagreement among annotators on certain labels, thereby allowing them to learn a more nuanced representation of the data. Additionally, the researchers tested various loss functions for soft labels, including cross-entropy, the standard Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), and the inverse KL-divergence. Their findings demonstrate that, in all cases, the multi-task models outperformed their single-task counterparts trained on the same tasks. Moreover, the researchers report that the auxiliary task served to mitigate overfitting, functioning as a regularizer, particularly in the case of ambiguous cases.

In conclusion, there is a considerable range of methodologies employed for integrating disagreement into the development of ML models. In traditional approaches based on the gold standard, disagreement is aggregated into single labels or used to filter out difficult items. In contrast, more recent approaches recognise and make use of the diversity of annotator perspectives by employing alternative standards and training techniques with the objective of developing models that are more robust in real-world scenarios.

3. Related Work

While there are several approaches to incorporating disagreement into the training pipeline of ML models, only a few specifically target the task of hate speech detection. One such approach is introduced by Kralj Novak et al. (2022), who train monolingual hate speech classifiers on English, Italian, and Slovenian diamond standard datasets. For the English classifier, they fine-tuned the BERT base model (Devlin et al., 2018) on a dataset consisting of YouTube comments. Each comment has been annotated by two annotators, which ultimately led to cases of disagreement. In the case of Italian, the ALBERTo model (Polignano et al., 2019) was fine-tuned on a dataset consisting of Italian YouTube comments, annotated following the same procedure. For Slovenian, the same procedure was repeated using the CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020) and a dataset of Slovenian tweets. For evaluation, the researchers used metrics such as Krippendorff’s ordinal alpha, the F1 score, and accuracy to assess both inter-annotator agreement and model performance. Thus, an evaluation from the perspective of disagreement was simulated. Their main findings indicate that, according to the accuracy scores, the annotators demonstrated a high degree of agreement in approximately 80% of the cases across all three datasets. Nevertheless, in terms of the Krippendorff’s ordinal alpha score, which considers both class imbalance and the ordering of classes (from least to most severe), the agreement score was reduced to approximately 60% in all three languages. Furthermore, the evaluation results indicated that the performance of each model aligned with the inter-annotator agreement, both in terms of accuracy and the alpha score. This implies that the performance of models is inherently constrained by the level of agreement among annotators. Consequently, when trained on diamond standard data, it is unlikely that the performance of these models can significantly surpass human performance. Finally, the researchers indicate that class imbalance represents another limitation that requires attention in future research.

Another novel approach to hate speech modeling which makes use of annotator disagreement is presented in Fleisig et al. (2023). This approach involves the development of two interconnected modules designed to effectively address hate speech annotated by individuals from different demographic groups. The initial module employs a RoBERTa-based model (Liu et al., 2019) that has been fine-tuned on two hate speech datasets. This module makes use of demographic information

3. *Related Work*

and survey responses from annotators in order to predict the individual ratings that they would assign to instances which could potentially be considered hateful. The second module employs a GPT-2-based model (Radford et al., 2019) trained to identify the demographic groups targeted by the potentially hateful instances. In combination, these modules are designed to first predict the target group of the input text and then to estimate the ratings that members of this predicted target group would assign to the text. This approach enables the identification of instances that might not be considered hateful by the majority but are perceived as hateful by annotators belonging to the targeted groups, thereby enhancing the model’s classification accuracy. Their main findings highlight the benefits of incorporating demographic information, survey responses, and target group predictions to improve model performance, mitigate annotator disagreement, and address ethical considerations in subjective tasks that involve diverse annotator perspectives. However, one main limitation of this approach is that it is currently limited to English data.

4. Method

This chapter outlines the method employed for the development and evaluation of the three multilingual hate speech classifiers which were developed according to the theoretical foundations presented in the earlier chapters of this thesis. Section 4.1 provides a comprehensive examination of the diamond standard datasets utilized for both training and evaluation purposes. Section 4.2 outlines the selection of the pre-trained language model upon which the hate speech classifiers were built. Section 4.3 describes the fine-tuning process. Section 4.4 discusses the evaluation strategies employed.

4.1. Datasets

The three monolingual datasets introduced by Kralj Novak et al. (2022) served as the starting point for the datasets developed in this thesis. The first dataset contains 124,812¹ English comments posted on YouTube between February 2020 and May 2020, focusing on content related to the COVID-19 pandemic. Similarly, the second dataset comprises 140,812² Italian YouTube comments on the same topic, posted between January 2020 and May 2020. The third dataset consists of 120,000³ Slovenian tweets on general topics, posted between December 2017 and October 2020. The annotators were recruited from Slovenia and Italy. The requirements were an excellent knowledge of the target language and an interest in hate speech and social media topics. They were provided with annotation guidelines in their first languages, which included information on the available labels, as well as access to a support group to which they could direct questions related the annotation task. The distribution of data among the annotators was constrained by the following conditions (Kralj Novak et al., 2022, 685):

- Each social media item should be annotated by two annotators independently.
- Each annotator should receive approximately the same number of items to annotate.

¹62,414 unique comments, the majority annotated by two annotators independently

²70,406 unique comments, the majority annotated by two annotators independently

³60,000 unique comments, annotated by two annotators independently

4. Method

- All pairs of annotators should have roughly the same overlap in the number of items assigned to them.
- For YouTube comments, threads should be kept intact; each annotator should be assigned both long and short threads.

Each social media item could be assigned one of four available classes: [Appropriate], [Inappropriate], [Offensive], and [Violent]⁴. The definition of each class is as follows (Kralj Novak et al., 2022, 683):

- **Acceptable:** does not present inappropriate, offensive or violent elements.
- **Inappropriate:** contains terms that are obscene or vulgar; but the text is not directed at any specific target.
- **Offensive:** includes offensive generalizations, contempt, dehumanization, or indirect offensive remarks.
- **Violent:** threatens, indulges, desires or calls for physical violence against a target; it also includes calling for, denying or glorifying war crimes and crimes against humanity.

Additionally, the original annotation scheme for the [Offensive] and [Violent] classes included a target for each instance of hate speech. In a subsequent annotation step, one of ten pre-specified targets was to be assigned to each instance belonging to these two classes. The available targets were as follows: [Racism], [Migrants], [Islamophobia], [Antisemitism], [Religion (other)], [Homophobia], [Sexism], [Ideology], [Media], [Politics], [Individual], and [Other]. In the case of Italian, an additional [North vs. South] target was included. These targets were not within the scope of this thesis and therefore not considered.

Another important aspect of these datasets is that, while the evaluation sets were randomly sampled, the data selection process was optimized for all the training sets using the OpSaLa approach introduced in Pelicon et al. (2023) and outlined in Subsection 2.3.3. This was done in order to address dataset imbalance and resulted in the creation of training sets with approximately two times more hate speech instances belonging to the [Violent] class than could have been collected from a random sample. Nevertheless, the proportion of [Violent] hate speech instances remained underrepresented compared to the rest, as can be seen in Figures 5 - 7 in Appendix A.2. This once again confirms the observation made in Subsection 2.3.3 that more severe forms of hate speech are usually the most underrepresented in hate speech datasets. Furthermore, it can be observed that the [Inappropriate] class is

⁴[Appropriato], [Nespodobni govor], [Offensivo], [Violento] for Italian; [Ni sporni govor], [Inappropriato], [Žalitev], [Nasilje] for Slovenian

also underrepresented in all of the three languages, suggesting that the majority of hateful content found on social media platforms is directed at specific individuals or groups. This is a defining characteristic of hate speech, as discussed in Section 2.3.1.

The combination of the three monolingual datasets described above led to the creation of three multilingual datasets which were used to train the classifiers presented in this thesis. The first dataset (*Duplicate All*) contains all instances from the three monolingual datasets, merged and shuffled. In the second dataset (*Duplicate Disagreement*), instances on which annotators disagreed appear twice with their respective labels, while instances that they agreed upon appear only once. The idea behind this dataset was to create a more balanced training set that reflects both agreement and disagreement equally, potentially preventing the models from being biased towards instances where annotators agree. The third dataset (*Remove Disagreement*) consists only of instances which were agreed upon, each appearing only once. Thus, the first two datasets contain diamond standard data, while the third dataset can be considered a gold standard dataset. Additionally, during the training phase of each model, a validation set comprising 10% of the training dataset was automatically generated using the `train_test_split` function of the `scikit-learn` library. It is important to note that all instances in these datasets have undergone several preprocessing steps, such as lower-casing and replacing links and usernames with placeholders. The latter step was undertaken to mitigate any potential biases associated with certain names, as discussed in Garg et al. (2018). Table 4 presents an overview of the label distribution across the three multilingual training sets. The datasets used for evaluation are visualized in Table 5. These reflect the unmodified monolingual evaluation sets presented in Kralj Novak et al. (2022).

Table 4.: Distribution of labels across the multilingual training sets

Dataset	Acceptable	Inappropriate	Offensive	Violent
Duplicate All	191,677	11,005	112,833	7,145
Duplicate Disagreement	111,324	8,346	72,706	4,992
Remove Disagreement	80,573	2,661	40,255	2,161

4.2. Model Selection

For the purpose of this thesis, four hate speech classifiers have been developed by fine-tuning the pre-trained XLM-R model (Conneau et al., 2019) using each of the

4. Method

Table 5.: Distribution of labels across the monolingual evaluation sets

Dataset	Acceptable	Inappropriate	Offensive	Violent
English	15,934	183	5,310	76
Italian	15,956	770	4,082	264
Slovenian	13,273	285	6,373	69

datasets presented in the previous section and illustrated in Table 4. The XLM-R model was selected as the baseline for the development of the multilingual hate speech classifiers due to its demonstrated effectiveness in cross-lingual understanding and its capability to process a wide range of languages, providing a solid foundation for subsequent fine-tuning and optimization. More specifically, English, Italian and Slovenian, the languages used for fine-tuning, were also among the languages used in the pre-training of the XLM-R model.

To answer the research question posed at the beginning of this thesis and compare the performance of the multilingual models against that of monolingual ones, the three monolingual models, i.e., the English, Italian, and Slovenian models developed by Kralj Novak et al. (2022) were selected as the monolingual baselines.

4.3. Fine-Tuning

As a preliminary remark, it should be noted that, although each classifier was fine-tuned on a distinct dataset, the fine-tuning process remained consistent across all models. In an initial step of the fine-tuning process, the datasets were converted into a format compatible with the Hugging Face `datasets` library (Lhoest et al., 2021) using the `Dataset.from_pandas` function. This allowed for efficient handling and manipulation of the data by leveraging the library’s built-in functions for tokenization, batching, and parallel processing. Each dataset was then partitioned into training and evaluation subsets, with 90% allocated for training and 10% for evaluation. The text data was then tokenized using the XLM-R tokenizer, with sequences padded and truncated to a fixed length of 512 tokens. The pre-trained XLM-R model, configured for sequence classification, was then loaded for fine-tuning, repeating the process for each dataset. The fine-tuning process was implemented using PyTorch (Paszke et al., 2017). To address class imbalance and enhance model performance on minority classes, a custom training loop with a weighted cross-entropy loss function was implemented, as proposed in (Pelicon et al., 2023). The class weights were calculated to be inversely proportional to the frequency of each hate speech class within the training data. Fine-tuning was

conducted using the `Trainer` class from the Hugging Face `transformers` library (Wolf et al., 2020), modified to incorporate the specified class weights (see Algorithm 1). To assess the effectiveness of the re-weighting method, one additional model was fine-tuned on the *Duplicate All* dataset using the regular configuration of the `Trainer` class. This model served as a benchmark for comparison with the other models. The hyperparameters for the fine-tuning process included a learning rate of 6×10^{-6} , a batch size of 8, and 3 training epochs. It is important to note that, due to computational constraints, each epoch was trained independently over a two-day period for each model. During the training phase, the AdamW optimizer was employed to optimize the model parameters.

Algorithm 1 Custom Trainer class for weighted loss

```

1: Initialize:
2:   class_weights  $\leftarrow$  None
3:   loss_fct  $\leftarrow$  CrossEntropyLoss
4:
5: function init_weights(class_weights):
6:   if class_weights is not None:
7:     class_weights  $\leftarrow$  tensor(class_weights, dtype=float).to(device)
8:
9: function compute_loss (model, inputs, class_weights=None, re-
   return_outputs=False):
10:  labels  $\leftarrow$  inputs.pop("labels")
11:  outputs  $\leftarrow$  model(forward_pass(inputs))
12:  logits  $\leftarrow$  extract_logits(outputs)
13:
14:  if class_weights is not None:
15:    loss_fct  $\leftarrow$  CrossEntropyLoss(weight=class_weights)
16:  else:
17:    loss_fct  $\leftarrow$  CrossEntropyLoss()
18:
19:  loss  $\leftarrow$  loss_fct(logits.view(-1, num_labels), labels.view(-1))
20:
21:  if return_outputs is True:
22:    return loss, outputs
23:  else:
24:    return loss

```

4. Method

4.4. Evaluation

In terms of evaluation, the approach introduced in Kralj Novak et al. (2022) was replicated in order to compare the performance of the multilingual classifiers to human judgement from the perspective of disagreement. This was achieved by employing identical measures to estimate the agreement between human annotators, as well as the agreement between annotators and models. That said, the measures employed in Kralj Novak et al. (2022) and subsequently in this thesis are Krippendorff’s ordinal alpha (Krippendorff, 2018), the F1 score and accuracy. In this particular case, Krippendorff’s ordinal alpha becomes extremely relevant since it takes into consideration the fact that the hate speech classes are ordered, i.e., have different magnitudes, respectively $[\text{Appropriate}] < [\text{Inappropriate}] < [\text{Offensive}] < [\text{Violent}]$. To initiate the evaluation process, each monolingual diamond standard dataset is represented using a $\mathbf{n} \times \mathbf{m}$ reliability matrix, $\mathbf{n}$ being the number of items labeled, and $\mathbf{m}$ the number of annotators. In this matrix, an entry $c_{iu} \in \mathcal{C}$ represents a label assigned by annotator $i \in \{1, \dots, \mathbf{m}\}$ to item $u \in \{1, \dots, \mathbf{n}\}$. The reliability matrix may contain missing values, i.e., items that have not been labeled. Subsequently, a $\mathbf{k} \times \mathbf{k}$ coincidence matrix, $\mathbf{k} = |\mathcal{C}|$ being the number of possible values of $\mathcal{C}$, is built. This matrix tabulates all the combined values of $\mathbf{c}$ assigned by two different annotators. A possible representation is as depicted in Equation 14.

$$\begin{array}{c|cc|c}
 & c' & & \Sigma \\
 \hline
 & \cdot & \cdot & \cdot \\
 c & \cdot & N(c, c') & \cdot \\
 & \cdot & \cdot & \cdot \\
 \hline
 \Sigma & \cdot & N(c') & \cdot \\
 & & & N
 \end{array} \tag{14}$$

In the coincidence matrix, an entry $N(\mathbf{c}, \mathbf{c}')$ represents all occurrences where one annotator has assigned the label $\mathbf{c}$ and the other $\mathbf{c}'$ across all items. $N(\mathbf{c})$ and $N(\mathbf{c}')$ represent the totals for each label, while N represents the grand total. The coincidences $N(\mathbf{c}, \mathbf{c}')$ are computed using Equation 15.

$$N(c, c') = \sum_u \frac{N_u(c, c')}{m_u - 1} \quad c, c' \in \mathcal{C} \tag{15}$$

In this equation, $N_u(\mathbf{c}, \mathbf{c}')$ is the number of $(\mathbf{c}, \mathbf{c}')$ pairs for the item u . m_u is the number of labels assigned to the item u . When calculating $N_u(\mathbf{c}, \mathbf{c}')$, each pair of annotations is considered twice, once as a $(\mathbf{c}, \mathbf{c}')$ pair, and once as a $(\mathbf{c}', \mathbf{c})$ pair. As a result, the coincidence matrix is symmetric around the diagonal, which contains all the matching labels. Utilizing these values as a starting point, each of the aforementioned evaluation metrics can be calculated to quantify both the agreement among annotators and the agreement between the models and the

annotators.

Firstly, in the case of Krippendorff's ordinal alpha (previously defined in Equation 13), the total observed disagreement between annotators can be defined as depicted Equation 16.

$$O = \frac{1}{N} \sum_{c, c'} N(c, c') \cdot \delta^2(c, c'), \quad (16)$$

The total disagreement expected by chance is defined as in Equation 17.

$$E = \frac{1}{N(N-1)} \sum_{c, c'} N(c) \cdot N(c') \cdot \delta^2(c, c') \quad (17)$$

In Equations 16 and 17, $\delta(c, c')$ represents the difference function between the values of c and c' , and is dependent on the type of the preferred decision variable c (nominal, ordinal, interval, etc.). Seeing as in this case, c is an ordinal variable, δ is defined as in Equation 18 (Kralj Novak et al., 2022).

$$\delta(c, c') = \sum_{i=c}^{i=c'} N(i) - \frac{N(c) + N(c')}{2} \quad \text{e.g. } c, c', i \in \{1^{st}, 2^{nd}, 3^{rd}, 4^{th}\}. \quad (18)$$

In this thesis, Krippendorff's ordinal alpha is employed to measure not only the disagreement among annotators, but also the disagreement between the models' predictions and the annotators' labels. In case of the latter, a distinct coincidence matrix is constructed. This matrix tabulates all possible pairs of model-annotator labels, which differs from the initial matrix that tabulates all possible pairs of labels assigned by different annotators.

Secondly, the F1 score is employed to calculate class-specific scores. It can be calculated for a given class c using Equation 19.

$$F_1(c) = \frac{2 \times \text{Pre}(c) \times \text{Rec}(c)}{\text{Pre}(c) + \text{Rec}(c)} \quad (19)$$

However, in the scenario where the F1 score is applied to the symmetric coincidence matrix defined in Equation 14, precision and recall become equal, since false positives and false negatives both point to disagreement. In this case, the F1 score formula can be simplified to Equation 20.

$$F_1(c) = \frac{N(c, c)}{N(c)} \quad (20)$$

4. Method

In assessing annotator agreement, $F_1(c)$ represents the proportion of items equally labeled by annotators out of all items assigned label c .

Finally, accuracy is also employed to quantify annotator disagreement, while acknowledging its limitations in not accounting for disagreement due to chance or the order of classes. To define observed disagreement using accuracy, Equation 21 was used.

$$\text{Acc} = 1 - D_o = \frac{1}{N} \sum_c N(c, c) \quad (21)$$

Having defined proper evaluation measures, the evaluation process began by running each of the fine-tuned classifiers on the three monolingual evaluation sets visualized in Table 5. The resulting predictions were used to compute both accuracy and class-specific F1 scores. Additionally, the weighted F1 score for each model was calculated to determine their relative performance. Regarding inter-annotator agreement, the previously described evaluation metrics were applied to each of the evaluation sets. To assess model-annotator agreement, coincidence matrices were constructed using the predictions from each model alongside the manual annotations. Subsequently, to determine whether the differences in performance among the models were statistically significant, two statistical tests, i.e. the Friedman (Friedman, 1937) and the Nemenyi (Nemenyi, 1963) tests, were employed. The Friedman test is a non-parametric statistical test used to detect differences in methods across multiple test attempts. It is particularly useful when the similar models are exposed to different methods, and the goal is to determine if there are statistically significant differences between these methods. Unlike parametric tests, the Friedman test does not assume normality of the data, making it suitable for a wide range of applications. The Nemenyi test is a post-hoc analysis test used following the Friedman test when significant differences are detected. It is designed to perform pairwise comparisons between methods to identify which specific groups differ from each other. The Nemenyi test accounts for multiple comparisons, making it a robust method for identifying statistically significant differences between classifiers.

The final evaluation step involved a qualitative analysis, through which the model that had been deemed to perform best on average according to the previous evaluation metrics was subjected to manual inspection. The objective of this analysis was to assess the model’s actual performance on various types of hate speech that were not included in its training or evaluation data. The examples were drawn from previously discussed publications and comprised both explicit and implicit instances of hate speech. This analysis was designed to demonstrate the model’s capacity to handle a range of hate speech instances, thereby providing

a more comprehensive assessment of its robustness when confronted with scenarios that deviate from its training distribution, as would be expected in a real-world context.

5. Results and Analysis

This chapter presents the results of the models developed with the purpose of answering the research question posed at the beginning of this thesis. It starts with an evaluation from the perspective of inter-annotator and model-annotator agreement in Section 5.1. Section 5.2 presents class specific evaluation results, as well as a model comparison based on the models’ average scores. Based on these scores, Section 5.3 discusses whether the differences in model performance are statistically significant by employing two statistical significance tests. Finally Section 5.4 presents and discusses the results of a qualitative analysis applied on the best performing model.

5.1. Inter-Annotator and Model-Annotator Agreement

An initial evaluation step focuses on comparing the agreement between the annotators themselves and the agreement between models and annotators. Inter-annotator agreement was computed for the evaluation sets in each language using Krippendorff’s alpha and accuracy. The same measures were used to compute the agreement between the annotators and the models, following the approach described in Section 4.4. The results are visualized in Table 6.

First, it can be observed that, in the case of inter-annotator agreement, annotators agree around 80% of the times in terms of accuracy, with an accuracy score between 79% and 82% across all three datasets. However, accuracy does not take into account the fact that the datasets are highly imbalanced, nor the ordering of classes. A more accurate estimate of the agreement between annotators is offered by Krippendorff’s ordinal alpha. In terms of the alpha score, the annotators agree nearly 60% of the times, with an alpha score between 56% and 58% across the three languages.

Second, in regards to the agreement between annotators and models, the same metrics were applied following the approach delineated in Section 4.4. The results show a consistent agreement between models and annotators in all of the cases. According to their accuracy scores, all models agree with at least one of the annotators around 80% of the times. Once again, the alpha score is lower (around 60%). Moreover, it can be observed that in some cases, for example in the case of the

5. Results and Analysis

Multilingual Duplicate All (No-Reweighting) model, the accuracy and alpha scores representing the model-annotator agreement are higher than the inter-annotator agreement scores. This phenomenon may indicate that the model is effectively learning consistent patterns or biases that align well with one or more of the annotators. This is to be expected in cases where annotator disagreement is primarily influenced by subjective interpretation and should not be interpreted as the model being inherently better, but rather efficient in modeling predominant patterns found in the training data.

Table 6.: Inter-Annotator Agreement (IAA) as opposed to model-annotator agreement in terms of Krippendorff’s ordinal alpha (α) and Accuracy (Acc.) for the models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All No Re-Weighting (MNR) models based on the language-specific evaluation sets

Dataset	IAA		MDA		MDD		MRD		MNR	
	α	Acc.	α	Acc.	α	Acc.	α	Acc.	α	Acc.
English	58.19	82.91	55.89	79.97	50.18	76.47	57.90	81.41	60.58	83.13
Italian	57.00	81.79	58.29	82.00	56.15	80.43	57.84	82.69	59.08	83.61
Slovenian	56.62	79.43	55.74	78.60	52.95	76.52	55.15	78.84	54.97	79.84

5.2. Model Comparison

To evaluate the performance of the models across the four hate speech classes, the F1 score was used. Additionally, the weighted F1 score was computed for each model to assess their overall performance. To determine the best-performing model, the weighted F1 scores were averaged across all three datasets. Table 7 shows the results achieved by each of the models on the English evaluation set. As mentioned before, the first three models represent the monolingual models developed by Kralj Novak et al. (2022). In the case of the English dataset, the results show that several multilingual models outperform the baseline monolingual English model across all classes except the [Appropriate] class, a case in which they still perform competitively. Another phenomenon to be observed is that all multilingual models outperformed the monolingual baselines in case of the minority classes [Inappropriate] and [Violent]. The model which achieved the highest score on these classes is the *Multilingual Duplicate All model*, with an F1 score of 39.16% for the [Inappropriate] class and an F1 score of 27.82% for the [Violent] class. This could be considered as proof that the re-weighting method was effective in

improving performance on underrepresented classes. Nevertheless, the best overall score according to the weighted F1 measure was achieved by the *Multilingual Duplicate All* model with an F1 score of 83.61%. As this model was fine-tuned without class re-weighting, its performance indicates that while the re-weighting method might be useful in terms of minority classes, it might lead to a performance decrease in the case of the remaining classes.

Moving on to the Italian dataset, similar patterns can be identified when comparing model performance, as shown in Table 8. While coming close, none of the multilingual models outperformed the Italian baseline on the [Appropriate] class. That being said, the highest scores on all the remaining classes was achieved by the *Multilingual Duplicate All* model, once again showing that re-weighting might potentially be an effective solution for the handling of minority classes. It is to be noted that all other multilingual models surpass the performance of the monolingual baseline on these classes as well, the *Multilingual Duplicate All (No-Reweighting)* model achieving the highest weighted F1 score of 83.14%. Nevertheless, the monolingual Italian model has the highest weighted F1 score with an overall result of 83.22%.

In the case of the Slovenian dataset, the observed phenomena slightly differ from the previous ones. The evaluation results are presented in Table 9. At least one multilingual model outperforms the monolingual Slovenian model on each class, but the [Violent] one, where the model trained without re-weighting achieved the best F1 score of 27.18% after the monolingual one (F1 score of 29.12%). In terms of the second minority class, the [Inappropriate] one, the *Multilingual Duplicate All* model scores best with an F1 score of 45.22%. The best overall performance in terms of the weighted F1 score is achieved by the *Duplicate All (No-Reweighting)* model with an F1 score of 79.29%.

Finally, Table 10 shows the average scores of all models, achieved by averaging their weighted F1 scores across all three datasets. It is the *Multilingual Duplicate All (No-Re-Weighting)* model that achieves the best overall performance with an F1 score of 82.01%. On the one hand, this shows that the re-weighting technique employed might hinder the overall performance score, as previously discussed. On the other hand, it shows that a model trained using diamond standard data outperformed not only all of the monolingual baselines, but also the model trained on gold standard data. Nevertheless, it is important to be cautious when making a comparison such as this and acknowledge the fact that the *Multilingual Remove Disagreement* model was not only fine-tuned using re-weighting, but also on a smaller dataset. For a future fairer comparison, the fine-tuning process on gold standard data should be adjusted accordingly.

5. Results and Analysis

Table 7.: Model evaluation and inter-annotator agreement results in terms of class-specific F1 scores on the English dataset. The Total score was calculated using the weighted F1 score. The first three models represent the English (EN), Italian (IT), and Slovenian (SL) baselines. The subsequent models represent the multilingual models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All (No-Reweightings) (MNR).

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN Model	89.38	28.95	68.36	24.17	83.44
IT Model	85.25	13.81	0.41	0.00	63.39
SL Model	88.01	25.17	49.69	2.88	77.71
MDA Model	86.10	39.16	68.24	27.82	81.09
MDD Model	83.33	34.16	65.07	24.52	78.20
MRD Model	87.43	29.90	69.02	27.27	82.18
MNR Model	88.81	36.76	70.43	26.66	83.61
Inter-Annotator	89.14	24.04	67.19	18.42	–

Table 8.: Model evaluation and inter-annotator agreement results in terms of class-specific F1 scores on the Italian dataset. The Total score was calculated using the weighted F1 score. The first three models represent the English (EN), Italian (IT), and Slovenian (SL) baselines. The subsequent models represent the multilingual models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All (No-Reweightings) (MNR).

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN Model	86.27	1.28	1.05	0.00	67.42
IT Model	91.32	58.46	59.02	40.34	83.22
SL Model	86.23	0.76	3.25	0.00	65.95
MDA Model	89.77	58.45	60.42	44.97	82.38
MDD Model	88.95	56.04	58.31	39.85	81.19
MRD Model	90.41	55.46	59.49	38.78	82.50
MNR Model	90.97	56.70	60.03	44.16	83.14
Inter-Annotator	89.84	52.99	54.47	54.75	–

5.3. Statistical Significance

Table 9.: Model evaluation and inter-annotator agreement results in terms of class-specific F1 scores on the Slovenian dataset. The Total score was calculated using the weighted F1 score. The first three models represent the English (EN), Italian (IT), and Slovenian (SL) baselines. The subsequent models represent the multilingual models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), Multilingual Remove Disagreement (MRD) and Multilingual Duplicate All (No-Reweightings) (MNR).

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN Model	79.93	3.98	2.34	0.00	53.84
IT Model	79.84	3.80	1.24	0.00	53.43
SL Model	85.70	43.69	65.26	29.12	78.39
MDA Model	84.30	45.22	69.69	24.79	78.88
MDD Model	82.33	43.39	68.59	23.84	77.19
MRD Model	84.98	38.47	68.40	15.50	78.80
MNR Model	85.98	44.00	67.50	27.18	79.29
Inter-Annotator	85.45	40.70	69.10	34.78	—

Table 10.: Average scores calculated by averaging the weighted F1 scores of each model.

Model	Average Weighted F1 Score
EN Model	68.23
IT Model	66.68
SL Model	74.02
MDA	80.78
MDD	78.86
MRD	81.16
MNR	82.01

5.3. Statistical Significance

Given the fact that the differences in performance discussed in the previous section are minimal across most of the multilingual models, it is generally considered beneficial to employ statistical significance tests when deciding on the best performing model. In this thesis, two such tests, the Friedman (Friedman, 1937) and the Nemenyi (Nemenyi, 1963) statistical tests, were employed. First, the Friedman test was applied for evaluating the performance of multiple hate speech classifiers

5. Results and Analysis

across the four hate speech classes. The initial implementation step consisted in creating tables where each row corresponds to a combination of hate speech classes and languages, and each column represents a classifier. This resulted in three different matrices, one for each language. The class-specific F1 scores presented in Tables 7 - 9 represented the elements of these matrices. The Friedman test was then applied on these matrices using the `scipy.stats` library to statistically evaluate the ranks of the classifiers across all combinations and determine if there are significant differences in their performances. The results yielded a statistic of 15.10 and a p-value of 0.0017, indicating that the differences in classifier performance are statistically significant, as the p-value was below the common significance threshold of 0.05. Subsequently, to assess whether the observed differences in classifier ranks are significant beyond chance, the Nemenyi test available in the `scikit_posthocs` package was performed. The results are visualized in the critical difference diagram in Figure 4. As can be observed, the performances of the *Multilingual Duplicate All* and the *Multilingual Duplicate Disagreement* models (p-value of 0.0085), as well as the ones between the *Multilingual Duplicate Disagreement* and *Multilingual Duplicate All (No Re-weighting)* models (p-value of 0.0142), differ significantly from each other. Taking into consideration the evaluation results discussed previously, it can be observed that, across all multilingual models, it is the *Multilingual Duplicate Disagreement* model that performs worst on average. Therefore, it can be concluded that (1) the method of duplicating only instances where annotators disagree in the dataset has proven to be least effective, and (2) there are no statistically significant differences between the rest of the models. That said, it is the *Multilingual Duplicate All (No-Reweighting)* model that achieved the highest average score (0.820 F1-points) and will therefore be considered the best performing model in this thesis. The *Multilingual Duplicate All (No-Re-Weighting)* is available at: <https://tinyurl.com/HF-multilingual-hs-model>.

5.4. Qualitative Analysis

Having assessed the performance of the multilingual hate speech classifiers based on scores achieved on the evaluation set, manual inspection has been employed as a final qualitative analysis step. This step was taken in order to assess how the model considered best on average would perform in detecting different types of hate speech drawn from different sources. To achieve this, the *Multilingual Duplicate All (No-Reweighting)* model was tested on the examples used to illustrate different types of implicit and explicit hate speech in Subsection 2.3.1¹. The results are presented in Table 11.

¹Due to limited knowledge of Italian and Slovenian, only English examples were used for this analysis.

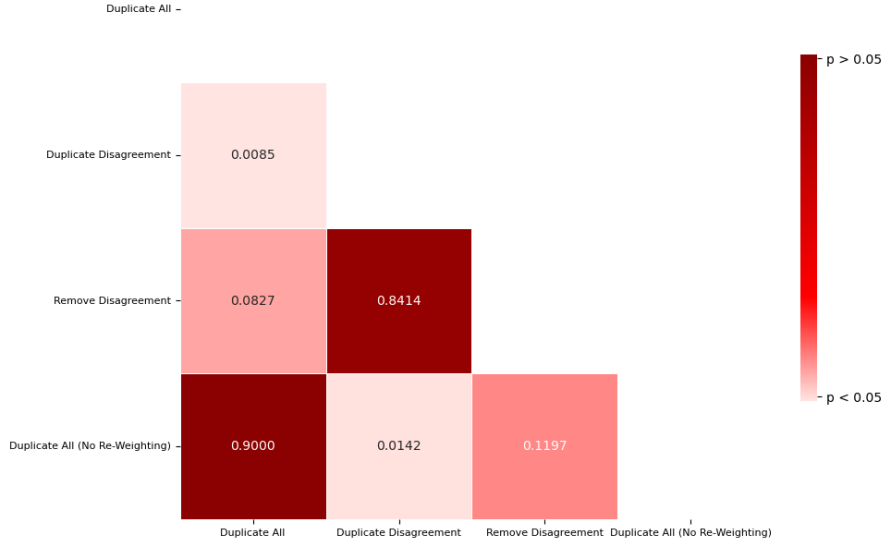


Figure 4.: Critical difference diagram according to the Nemenyi test

On the one hand, it can be observed that the model did not recognize any of the implicit forms of hate speech, but one. Given the fact that all hate speech classes used for fine-tuning the model referred to explicit forms of hate speech, these results were expected. Moreover, it is well known that recognizing specific instances of implicit hate speech as such requires a considerable amount of world knowledge, such as the fifth example of implicit hate speech, which refers to the fact that Stevie Wonder was blind. Nevertheless, the model classified the third sentence as offensive, which could be considered a correct categorization given the description of the [Offensive] class that includes forms of dehumanization (see Section 4.1). In an attempt of estimating which words triggered the model’s prediction, several relevant n-grams of the sentence were input into the model individually. While the noun *muslims* itself was classified as [Appropriate], both the combinations *headscarfed muslims* and *muslims waddle* were categorized as [Offensive]. A possible explanation for this is that the term *headscarf* is typically associated with Muslim women and carries significant cultural and religious connotations. The use of the adjective *headscarfed* in the sentence implies a direct reference to a specific demographic group, which could be recognized by the model as a potential indicator of hate speech, especially in datasets where religious attire is contextually related to negative sentiments. Another possible explanation lies in the contextual learning abilities of the model’s architecture. The contextual combination with *waddle*, a verb that might imply derogatory action or attribute negative characteristics, could

5. Results and Analysis

be recognized by the model as matching patterns similar to those found in training examples of explicit hate speech. This sensitivity to context suggests that while the model struggles with subtle implicit cues, it can effectively detect offensive content when specific phrases align closely with derogatory contexts seen during training.

On the other hand, the model correctly predicted the majority of explicit hate speech instances as belonging to the [Offensive] class. In case of the third explicit hate speech sentence, which the model classified as [Appropriate], the reason for the missclassification might be attributed to the coded language that was used. Since the model was not explicitly trained to recognize such coded terms or neologisms, the sentence was rewritten using the actual word implied. This time, the model classified the sentence as [Inappropriate]. Nevertheless, the more accurate label would have been [Offensive], since the insult targets a specific person. Finally, the last sentence should have been categorized as [Violent], since it implies calls for physical violence against a target. The model might have missed this cue due to two main reasons: the negation *did not*, and the more euphemistic, yet violent expression *finsh them off*. To test this hypothesis, the second sentence of the comment has been rewritten to *Hitler should have killed them all !!!*. This time, the model correctly classified the comment as [Violent].

In conclusion, while the best performing multilingual hate speech classifier developed in this thesis did not recognize implicit forms of hate speech, it was able to identify different forms of explicit hate speech effectively. That being said, it seems that its performance is limited by its training data, which may not include sufficient examples of nuanced language, coded expressions, or neologisms in offensive language. Consequently, future research could focus on enhancing the model's ability to interpret such subtleties by integrating a broader collection of linguistic contexts and training examples to improve its overall sensitivity to different forms of hate speech.

Table 11.: Examples of implicit and explicit hate speech (HS) and model predictions used for the qualitative analysis

Text	HS Type	Model Prediction
„It’s always fun watching sports with a woman in the room.“ (Wiegand et al., 2021, 580)	Implicit	Appropriate
„Women should all stay at home in an apron, chained to the oven!“ (Wiegand et al., 2021, 578)	Implicit	Appropriate
„Headscarfed muslims waddle around our streets all over.“ (Wiegand et al., 2021, 579)	Implicit	Offensive
„You sing like a dying bird.“ (Wiegand et al., 2021, 579)	Implicit	Appropriate
„Did Stevie Wonder choose these “models”?“ (Wiegand et al., 2021, 580)	Implicit	Appropriate
„@eBeZa: Stupid f*cking n*gger LeBron. You flipping jungle bunny monkey f*ggot.“ (Davidson et al., 2017, 514)	Explicit	Offensive
„My advice of the day: If your a tranny...go f*ck your self!“ (Davidson et al., 2017)	Explicit	Offensive
„Somebody told me that Little Debbie likes to take it up the A.\$\$.“ (Nobata et al., 2016, 149)	Explicit	Appropriate
„Don’t try to explain—Irish Catholics are just idiots“ (Fortuna and Nunes, 2018, 6)	Explicit	Offensive
„The JEWZ are Market Manipulators Hate them freaking Jews !! To bad Hitler did not finish them off !!!“ (Nobata et al., 2016, 149)	Explicit	Offensive

6. Discussion

In recent years, the role of automated hate speech detection systems in moderating online content has gained significant attention due to its potential to mitigate the spread of hate speech, a phenomenon that negatively impacts the social dynamics of online communities. To address this challenge and answer the research question posed at the beginning of this thesis, four multilingual hate speech classifiers have been developed. Based on the results presented in Chapter 5, the main findings of this research are manifold.

First, the consistent agreement between the annotators themselves and the annotators and models leads to one main conclusion: inter-annotator agreement serves as an intrinsic limit for the performance achievable by the models. This is due to the fact that the models are constrained by the quality of the data. Consequently, limitations to the accuracy with which a model can predict previously unseen data can be observed in the case of instances where the data exhibits significant variability or inconsistency in annotations. Nevertheless, analyzing areas of disagreement and including them into the development of automated models may lead to more robust and nuanced understandings of complex phenomena such as hate speech, resulting in models with an improved ability of handling ambiguous cases commonly encountered in real-world scenarios.

A second main conclusion that can be drawn from the model comparison results is that each of the multilingual models demonstrated superior performance to the monolingual baselines. This indicates that the models are capable of leveraging patterns learned from the multilingual datasets to achieve enhanced results in each of the three languages, i.e., English, Italian, and Slovenian. This finding can be used to answer the first part of the research question: on average, multilingual hate speech classifiers perform better than their monolingual counterparts when trained on diamond standard data. With regard to the second part of the question, the evaluation results demonstrate that inter-annotator agreement can effectively be incorporated into the training of ML models through the use of diamond standard datasets. Furthermore, based on these results, it becomes evident that multilingual pre-trained models, such as the XLM-R model (Conneau et al., 2019), are highly advantageous in the development of multilingual solutions, particularly when working with a combination of high and low-resource languages, as has been done in this thesis. This superior performance can be led back to a number of reasons. Firstly, multilingual models have the capacity to transfer knowledge across

6. Discussion

languages, enabling patterns and features acquired in one language to inform and enhance learning in another. This phenomenon has also been observed in the evaluation of the multilingual classifiers developed, which demonstrated enhanced performance across all languages, in contrast to the monolingual models, which only performed well on the language they have explicitly been trained on. This cross-lingual transfer becomes particularly beneficial in the case of low-resource languages, where data scarcity usually leads to lower model performance. By sharing information across languages, multilingual models effectively compensate for the lack of extensive datasets in lower-resourced languages. Again, this was the case for the developed classifiers, given that English is regarded as a language with considerable resources, whereas Italian and Slovenian are considered to be lower-resourced. That said, the ability to generalize across languages leads to more robust models that can handle a wider variety of linguistic nuances and contexts, making them more efficient in real-life scenarios, such as the detection of hate speech on social media platforms with an international reach.

Third, the qualitative analysis performed as part of the evaluation process has shown that, while the best performing model is effective in the detection of explicit hate speech, the detection of implicit forms of hate requires more sophisticated training and nuanced approaches. Implicit hate speech often relies on context, cultural references, and subtleties in language that are not as overt as explicit hate speech. This makes the task more challenging because it necessitates not only an understanding of linguistic patterns but also extensive world knowledge and contextual awareness. Implicit expressions of hate speech can be masked in humor, sarcasm, or coded language, making them harder to identify without a deep comprehension of the underlying implications. Consequently, improving the detection of implicit hate speech may involve integrating broader datasets, employing advanced natural language processing techniques, and potentially leveraging external knowledge bases to provide the necessary contextual information.

Having discussed the major findings of this thesis, it becomes evident that it makes a significant contribution to the existing body of work presented in Chapter 3, specifically in regard to hate speech classification in a multilingual context. In particular, this research makes a significant contribution by introducing a multilingual variant of the approach presented by Kralj Novak et al. (2022), which served as the baseline. In doing so, it has been demonstrated that the XLM-R model (Conneau et al., 2019) not only outperforms the BERT-based (Devlin et al., 2018) monolingual baselines, but also effectively leverages diamond standard data during the fine-tuning process. This finding underscores the second major contribution of this thesis: the incorporation of inter-annotator disagreement into the model development process. While most existing approaches to hate speech detection rely on gold standard data, the validity of a gold standard is questionable

in subjective tasks such as hate speech detection, as discussed in the previous chapters. Consequently, exploring alternative methods like the diamond standard becomes crucial, as they may offer significant benefits for the development of robust, real-world solutions.

That being said, it is extremely important to also critically evaluate and discuss the methods used for the development of the models presented in this thesis, examining both their ethical and technical implications and limitations. On the one hand, the development of hate speech classifiers has serious ethical implications, as they directly affect the inclusivity and safety of online communities, while potentially reinforcing societal biases. The reinforcement of existing societal biases and stereotypes is known to be an inadvertent effect of ML models, due to their tendency to learn and reinforce the biases present in their training data. It is therefore essential to approach the design and implementation of such models with full awareness of their potential impact. That said, implementing rigorous bias mitigation strategies and continuously evaluating the fairness of the models are important steps to ensure that the existing biases and stereotypes are not being perpetuated or exacerbated.

On the other hand, there are several technical limitations inherent in the datasets, computational resources, and evaluation methods used for the training and evaluation of the models presented in this thesis. First, the structure of the comments used for fine-tuning the models varies significantly across the monolingual datasets, seeing as they have been collected from different social media platforms. Each of these platforms has its own linguistic conventions and user demographics, which can affect the applicability and effectiveness of the models. However, this diversity in the training data can also be advantageous, since it may potentially enhance the models' ability to generalize to broader contexts and handle a wider range of linguistic patterns.

Second, computational resource constraints may have also potentially affected the robustness of the training process, limiting the complexity of the models developed and the extent of hyperparameter tuning that could be performed. This factor also limited the exploration of further methods for handling minority classes, which remains an open research gap.

A third limitation concerns the selection and comparison of models. While all employed baseline models utilize the BERT architecture (Devlin et al., 2018), the developed multilingual classifiers leverage the more advanced XLM-R architecture (Conneau et al., 2019). This inherent difference in architectures and pre-training methodology introduces a potential bias in the comparison of models. Consequently, the observed performance improvements of the developed models may, in part, be attributable to the underlying architecture rather than solely reflecting the effectiveness of the proposed multilingual classifiers. To address this limitation, future

6. Discussion

research could explore comparisons between models with equivalent architectural complexity, allowing for a more isolated evaluation of the proposed multilingual adaptation techniques.

Finally, regarding the evaluation measures employed, the weighted F1 score, commonly used due to its balance between precision and recall, is not always ideal. In scenarios where class imbalance is pronounced, the weighted F1 score might still favor the majority class, leading to overestimated performance metrics. Additionally, the intrinsic limit of performance set by inter-annotator agreement challenges the feasibility of achieving perfect accuracy, as the models can only learn to predict as reliably as the data allows.

Nevertheless, training models with diamond standard datasets introduces a novel paradigm for model evaluation and training. Although models trained on gold standard data often achieve superior performance according to traditional metrics, such performance may not generalize well to the complexities of real-world data, particularly in the case of highly subjective tasks, such as hate speech detection. Training with disagreement, therefore, potentially offers a more realistic scenario for model training. That being said, it also presents challenges in evaluation, where traditional metrics may fail to capture a model’s effectiveness in handling ambiguous inputs. This highlights the need to explore alternative evaluation strategies and metrics, such as those proposed in this thesis.

7. Conclusion

This thesis has established a cohesive theoretical foundation on the topics of hate speech and learning from disagreement. It has also contributed to the development of automatic hate speech classifiers by proposing four novel multilingual models. For the development of these models, diamond standard data was used to incorporate inter-annotator disagreement into the training pipeline, providing an alternative approach to models based on the gold standard. In addition, the evaluation strategy employed in this work extended beyond traditional metrics by incorporating Krippendorff’s alpha (Krippendorff, 2018) to assess both inter-annotator and model-annotator agreement, thus simulating an evaluation from a disagreement perspective. The resulting multilingual models outperformed several monolingual baselines, supporting the claim that multilingual model architectures, such as XLM-R (Conneau et al., 2019), lead to superior model performance in diverse linguistic contexts. Consequently, this thesis highlights the efficiency of using diamond standard data for the development of hate speech classifiers. However, disagreement between annotators has also proven to be an inherent limitation to model performance.

Looking ahead, several directions exist for extending and improving this research in future work. First, evaluating the developed multilingual classifiers on languages other than the ones used for fine-tuning would assess their generalizability across diverse linguistic contexts. Second, exploring alternative baseline models and conducting extensive hyperparameter tuning could further optimize performance. Third, exploring alternative ways of incorporating annotator disagreement into the training process, potentially through methods such as logistic regression, could provide further valuable insights into handling subjective tasks, such as hate speech. Lastly, future approaches could benefit from more complex techniques for handling minority classes, which remains a common challenge in hate speech datasets.

In conclusion, as the number of individuals engaging with social media platforms on a global scale continues to grow, so too does the amount of hate speech posted online. Consequently, there is a pressing need for robust and effective content moderation systems capable of handling diverse linguistic contexts, as hate speech has a negative impact on the dynamics of online communities and can lead to social fragmentation and psychological harm. While this thesis has presented a valuable starting point for automatic multilingual hate speech detection, further research is required to address several identified research gaps. That said, potential

7. Conclusion

areas for future research include extending the training and evaluation pipelines to encompass a greater number of languages, exploring alternative baseline models, and investigating additional methods for addressing annotator disagreement, minority classes, and model evaluation. Addressing these areas in future work will not only enhance the technical capabilities of hate speech detection models, but will also facilitate the development of more robust and complex approaches to online content moderation. Ultimately, this will ensure a secure and respectful online environment, while also promoting the well-being of users in digital spaces.

Bibliography

- Swati Agarwal and Ashish Sureka (2017). Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr micro-blogging website. *CoRR*, *abs/1701.04931*.
- Sohail Akhtar, Valerio Basile, and Viviana Patti (2019). A new measure of polarization in the annotation of hate speech. In *AI* IA 2019—Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, pages 588–603. Springer.
- Nuha Albadi, Maram Kurdi, and Shivakant Mishra (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 69–76. ACM.
- Fatimah Alkomah and Xiaogang Ma (2022). A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.
- Aymé Arango, Jorge Pérez, and Barbara Poblete (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.
- Lora Aroyo and Chris Welty (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- John Langshaw Austin (1975). *How to do things with words*. Harvard university press.
- Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, *abs/1409.0473*.

Bibliography

- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Valerio Basile et al. (2020). It’s the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR workshop proceedings*, volume 2776, pages 31–40. CEUR-WS.
- Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava (2018). A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media*, pages 36–41.
- Alexander Brown (2017). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, 36:561–613.
- Pete Burnap and Matthew L Williams (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.
- Peter Burnap and Matthew Leighton Williams (2014). Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making.
- Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang (2020). Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, pages 11–20.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer (2020a). Hatebert: Retraining BERT for abusive language detection in English. *CoRR*, abs/2010.12472.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer (2020b). I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6193–6202.
- Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 international*

- conference on privacy, security, risk and trust and 2012 international conference on social computing*, pages 71–80. IEEE.
- Anusha Chhabra and Dinesh Kumar Vishwakarma (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, *abs/1406.1078*.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini (2019). Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *CoRR*, *abs/1910.03270*.
- Jacob Cohen (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, *abs/1911.02116*.
- Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata (2019). Cross-platform evaluation for Italian hate speech detection. In *CLiC-it 2019-6th Annual Conference of the Italian Association for Computational Linguistics*.
- Corinna Cortes and Vladimir Vapnik (1995). Support-vector networks. *Machine learning*, 20:273–297.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Alexander Philip Dawid and Allan M Skene (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*.

Bibliography

- Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.
- Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.
- Anca Dumitrache, Lora Aroyo, and Chris Welty (2018). Crowdsourcing semantic label propagation in relation classification. *CoRR*, abs/1809.00537.
- Facebook (2024). Facebook community standards. <https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/>. Accessed: 2024-04-10.
- Christiane Fellbaum (1998). *WordNet: An electronic lexical database*. MIT press.
- Johan Fernquist, Oskar Lindholm, Lisa Kaati, and Nazar Akrami (2019). A study on the feasibility to detect hate speech in Swedish. In *2019 IEEE international conference on big data (Big Data)*, pages 4724–4729. IEEE.
- Elisabetta Fersini, Debora Nozza, and Paolo Rosso (2020). AMI @ EVALITA2020: Automatic misogyny identification. *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*.
- Eve Fleisig, Rediet Abebe, and Dan Klein (2023). When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.
- Joseph L Fleiss (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Paula Fortuna and Sérgio Nunes (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

- Paula Fortuna, Juan Soler-Company, and Leo Wanner (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.
- Milton Friedman (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.
- Björn Gambäck and Utpal Kumar Sikdar (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- Kevin Gimpel, Nathan Schneider, Brendan O’connor, Dipanjan Das, Daniel P Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47.
- Goran Glavaš, Mladen Karan, and Ivan Vulić (2020). Xhate-999: Analyzing and detecting abusive language across domains and languages. Association for Computational Linguistics.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Sepp Hochreiter and Jürgen Schmidhuber (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra (2015). Detection of cyberbullying incidents on the Instagram social network. *CoRR*, abs/1503.03909.
- Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384.
- Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma (2019). FERMI at SemEval-2019 Task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 70–74.

Bibliography

- Alvi Md Ishmam and Sadia Sharmin (2019). Hateful speech detection in public Facebook pages for the Bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.
- Daniel Jurafsky and James H Martin (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.
- Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, et al. (2018). Sentiment analysis: It’s complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895.
- Beata Beigman Klebanov and Eyal Beigman (2014). Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396.
- Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti (2020). HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.
- Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo (2022). Handling disagreement in hate speech modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695. Springer.
- Klaus Krippendorff (2018). *Content analysis: An introduction to its methodology*. Sage publications.
- Solomon Kullback and Richard A Leibler (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Irene Kwok and Yuzhou Wang (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1621–1622.
- John P Lalor, Hao Wu, and Hong Yu (2017). Soft label memorization-generalization for natural language inference. *CoRR*, abs/1702.08563.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2019). Albert: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. (2021). Datasets: A community library for natural language processing. *CoRR*, abs/2109.02846.
- Shuhua Liu and Thomas Forss (2014). Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *Special Session on Text Mining*, volume 2, pages 530–537. SCITEPRESS.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Nikola Ljubešić, Igor Mozetič, and Petra Kralj Novak (2022). Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, pages 1–14.
- Rijul Magu and Jiebo Luo (2018). Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.
- Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel (2019). Overview of the HASOC track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.
- Zewdie Mossie and Jenq-Haur Wang (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.
- Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon (2021). Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 9:88364–88376.
- Peter Bjorn Nemenyi (1963). *Distribution-free multiple comparisons*. Princeton University.
- Ru Ni and Huan Cao (2020). Sentiment analysis based on GloVe and LSTM-GRU. In *2020 39th Chinese control conference (CCC)*, pages 7492–7497. IEEE.

Bibliography

- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.
- John T Nockleby (2000). Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.
- Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata (2023). Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772.
- Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney (2018). The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media*, volume 12.
- Jim O’Driscoll (2020). *Offensive language: Taboo, offence and social control*. Bloomsbury Publishing.
- Sinno Jialin Pan and Qiang Yang (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D Dikaiakos, and Evangelos Markatos (2020). Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–21.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer (2017). Automatic differentiation in Pytorch.
- Andraž Pelicon, Syrielle Montariol, and Petra Kralj Novak (2023). Don’t start your data labeling from scratch: Opsala-optimized data sampling before labeling. In *International Symposium on Intelligent Data Analysis*, pages 353–365. Springer.
- Juan Manuel Pérez, Franco M Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, et al. (2023). Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.
- Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626.

- Barbara Plank, Dirk Hovy, and Anders Søgaard (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.
- Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz (2019). A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.
- Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, et al. (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Dennis Reidsma and Hendrikus JA op den Akker (2008). Exploiting ”subjective” annotations. In *Workshop on Human Judgements in Computational Linguistics, Coling 2008*, pages 8–16. Coling 2008 Organizing Committee.
- Sebastian Ruder (2019). *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman (2008). Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77:157–173.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci (2018). An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bibliography

- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi (2019). Social bias frames: Reasoning about social and power implications of language. *CoRR*, *abs/1911.03891*.
- Anna Schmidt and Michael Wiegand (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Yash Sharma, Gaurav Agrawal, Pooja Jain, and Tapan Kumar (2017). Vector representation of words for sentiment analysis using GloVe. In *2017 international conference on intelligent communication and computational techniques (icct)*, pages 279–284. IEEE.
- Aashish Sheshadri and Matthew Lease (2013). Square: A benchmark for research on computing crowd consensus. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, pages 156–164.
- Josef Steinberger, Tomás Brychcín, Tomás Hercig, and Peter Krejzl (2017). Cross-lingual flames detection in news discussions. In *RANLP*, pages 694–700.
- Jherez Taylor, Melvyn Peignon, and Yi-Shin Chen (2017). Surfacing contextual hate speech words within social media. *CoRR*, *abs/1711.10093*.
- The National Archives (2010). Equality act 2010. <http://www.legislation.gov.uk/ukpga/2010/15/contents>. Accessed: 2024-04-24.
- Twitter (2024). Hateful conduct. <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>. Accessed: 2024-04-11.
- Matej Ulčar and Marko Robnik-Šikonja (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 104–111. Springer.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio (2020). A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

- Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser (2018). Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, *abs/1809.07572*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Bertie Vidgen and Leon Derczynski (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- Bertie Vidgen and Taha Yasseri (2020). Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.
- Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Italian Conference on Cybersecurity*.
- William Warner and Julia Hirschberg (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.
- Zeerak Waseem and Dirk Hovy (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder (2021). Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg (2018). Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.
- Ludwig Wittgenstein (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al.

Bibliography

- (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose (2012). Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.
- Wenjie Yin and Arkaitz Zubiaga (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.
- YouTube (2024). Hate speech policy. <https://support.google.com/youtube/answer/2801939?reftopic%20=%209282436>. Accessed: 2024-04-11.
- Annie Zaenen, Lauri Karttunen, and Richard Crouch (2005). Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 31–36.
- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin (2020). Semeval-2020 Task 12: Multilingual offensive language identification in social media (OffensEval 2020). *CoRR*, abs/2006.07235.
- Ziqi Zhang, David Robinson, and Jonathan Tepper (2018). Detecting hate speech on Twitter using a Convolution-GRU based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.
- Steven Zimmerman, Udo Kruschwitz, and Chris Fox (2018). Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A. Appendix

The following appendix provides supplementary material that support and extend the findings presented in the main body of this thesis.

A.1. Automatic Hate Speech Detection Resources and Approaches

The following tables hold detailed information about the hate speech resources and approaches presented in Chapter 2.3. Specifically, Table 12 provides a detailed overview of the hate speech resources discussed. It presents each resource in a systematic manner, including the type of resource, the data source, the annotation scheme, the labels used, the type of annotators involved, and the languages employed. Table 13 provides a comprehensive overview of the approaches to automatic hate speech detection that have been discussed, including the types of classifiers and the features employed.

Table 12.: Hate speech resources

Publication	Resource type	Data source	Annotation scheme	Labels	Annotator type	Language
(Albadi et al., 2018)	Dataset+lexica	Twitter	Multi-level	hateful, not hateful, unclear or unrelated; Muslims, Jews, Christians, Atheists, Sunnis, Shia, Other	Crowdsourcing	Arabic
(Basile et al., 2019)	Benchmark	Twitter	Multi-level	hateful, non-hateful; individual, generic; aggressive, non-aggressive	Crowdsourcing	Italian
(Davidson et al., 2017)	Dataset+lexica	Twitter	Non-binary	hate, offensive, neither	Crowdsourcing	English
(Wiegand et al., 2018)	Lexica	Not specified	Binary	abusive, not abusive	Crowdsourcing, automatic classifier	English
(Vidgen and Yasseri, 2020)	Dataset	Twitter	Non-binary	not islamophobic, weakly islamophobic, strongly islamophobic	Experts	English
(Kralj Novak et al., 2022)	Dataset	Twitter, YouTube	Non-binary	appropriate, inappropriate, offensive, violent	Non-experts	Slovene, Italian, English

A.1. Automatic Hate Speech Detection Resources and Approaches

Table 12 continued from previous page

Publication	Resource type	Data source	Annotation scheme	Labels	Annotator type	Language
(Olteanu et al., 2018)	Dataset+lexica	Twitter, Reddit	Multi-level	favorable, unfavorable, commentary, neutral; Muslims, other religious groups, Arabs, ethnic groups, immigrants, other groups; promotes violence, intimidates, offends or discriminates; diagnoses causes, suggests solutions, both	Crowdsourcing	English
(Mandl et al., 2019)	Benchmark	Twitter, Facebook	Multi-level	non hate-offensive, hate and offensive; hate, offensive, profanity; targeted insult, untargeted	Non-experts	English, Hindi, German
(Corazza et al., 2019)	Dataset	Instagram	Binary	hateful, not hateful	Experts	Italian
(Pérez et al., 2023)	Dataset	Twitter	Multi-level	hateful, not hateful; racism, appearance, criminal, politics, women, class, LGBTI, disabled	Non-experts	Spanish
(Ljubešić et al., 2022)	Dataset	Twitter, YouTube	Non-binary	appropriate, inappropriate, offensive, violent	Non-experts	Slovene, Italian, English

Table 12 continued from previous page

Publication	Resource type	Data source	Annotation scheme	Labels	Annotator type	Language
(Akhtar et al., 2019)	Dataset	Twitter	Binary	homophobic, not homophobic	Non-experts	Italian
(Vigna et al., 2017)	Dataset	Facebook	Non-binary	strong hate, weak hate, no hate	Experts	Italian
(Sanguinetti et al., 2018)	Benchmark	Twitter	Non-binary	hate speech, aggressiveness, offensiveness, irony, stereotype	Experts, crowdsourcing	Italian
(Waseem and Hovy, 2016)	Dataset	Twitter	Non-binary	racism, sexism, both, neither	Experts	English
(Nobata et al., 2016)	Dataset	Yahoo!News, Yahoo!Finance	Multi-level	clean, abusive; hate speech, derogatory, profanity	Experts, crowdsourcing	English
(Chung et al., 2019)	Dataset	Other	Multi-level	affiliation, consequences, denouncing, facts, humor, hypocrisy, negative, other, positive, question; crimes, culture, economics, generic, islamization, rapism, terrorism, women	Non-experts	English, French, Italian

A.1. Automatic Hate Speech Detection Resources and Approaches

Table 12 continued from previous page

Publication	Resource type	Data source	Annotation scheme	Labels	Annotator type	Language
(Steinberger et al., 2017)	Dataset	News websites	Binary	flames, no flames	Experts	Czech, English, French, Italian, German
(Bohra et al., 2018)	Dataset	Twitter	Binary	HS, normal speech	Non-experts	English, Hindi
(Ishmam and Sharmin, 2019)	Dataset	Facebook	Non-binary	HS, inciteful, religious, hatred, communal hatred, religious comment, political comment	Experts	Bengali
(Fernquist et al., 2019)	Dataset	Web fora	Multi-level	-3: aggression, -2: insult, -1: dislike, 0: neutral	Non-experts	Swedish

Table 13.: Approaches to automatic hate speech detection

Publication	Classifiers	Features
(Burnap and Williams, 2014)	SVM, Decision Tree, Random Forest	n-grams, dependency relationships
(Davidson et al., 2017)	SVM, Logistic Regression	TF-IDF, POS, sentiment analysis information, hashtags, mentions, retweets, URLs, number of characters, n-grams
(Kwok and Wang, 2013)	Naive Bayes	n-grams
(Liu and Forss, 2014)	Naive Bayes	TF-IDF, n-grams, topic similarity, sentiment analysis information
(Djuric et al., 2015)	Logistic Regression	word embeddings, lexicon
(Agarwal and Sureka, 2017)	Decision Tree, Decision Tree, Random Forest, One-class classifier	topic modelling, sentiment analysis information, tone analysis, semantic analysis, contextual metadata
(Burnap and Williams, 2015)	SVM, Decision Tree, Random Forest, Bayesian Logistic Regression, Ensemble	n-grams, dependency relationships
(Nobata et al., 2016)	Skip-gram Model	n-grams, word embeddings, lexicon
(Xiang et al., 2012)	SVM, Decision Tree, Random Forest, Logistic Regression	LDA clusters
(Dinakar et al., 2012)	SVM, Decision Tree, Naive Bayes	sentiment analysis information; knowledge base information

Table 13 continued from previous page

Publication	Classifiers	Features
(Chen et al., 2012)	Lexical Syntactic Feature-based language model	lexical features, dependency relationships
(Waseem and Hovy, 2016)	Logistic Regression	meta-information about users
(Hosseinmardi et al., 2015)	Naive Bayes, SVM	multimodal information, n-grams
(Zimmerman et al., 2018)	CNN ensemble	word embeddings
(Zhang et al., 2018)	CNN-GRU	convolutional features, word order information
(Gambäck and Sikdar, 2017)	CNN	n-grams, word embeddings
(Sharma et al., 2017)	RNN	word embeddings
(Ni and Cao, 2020)	RNN, LSTM-GRU	word embeddings
(Badjatiya et al., 2017)	Logistic Regression, SVM, CNN, Fast-Text, Bi-LSTM, Gradient Boosted Decision Tree	n-grams, TD-IDF, word embeddings
(Vigna et al., 2017)	SVM, Bi-LSTM	morpho-syntactical features, sentiment analysis information, word embeddings
(Mossie and Wang, 2020)	GRU, RNN	n-grams, word embeddings
(Koufakou et al., 2020)	BERT	lexicon-based lexical features, word embeddings

Table 13 continued from previous page

Publication	Classifiers	Features
(Fortuna et al., 2021)	BERT, ALBERT, FastText, SVM, Random Forest	n-grams, word embeddings
(Glavaš et al., 2020)	BERT, ROBERTA, mBERT, XLM-R	word embeddings

A.2. Label Distribution of Monolingual Hate Speech Datasets

The following figures illustrate the distribution of labels for the monolingual datasets, i.e., the English, Italian and Slovenian datasets, presented in Kralj Novak et al. (2022) and employed in the creation of the multilingual datasets presented in this thesis.

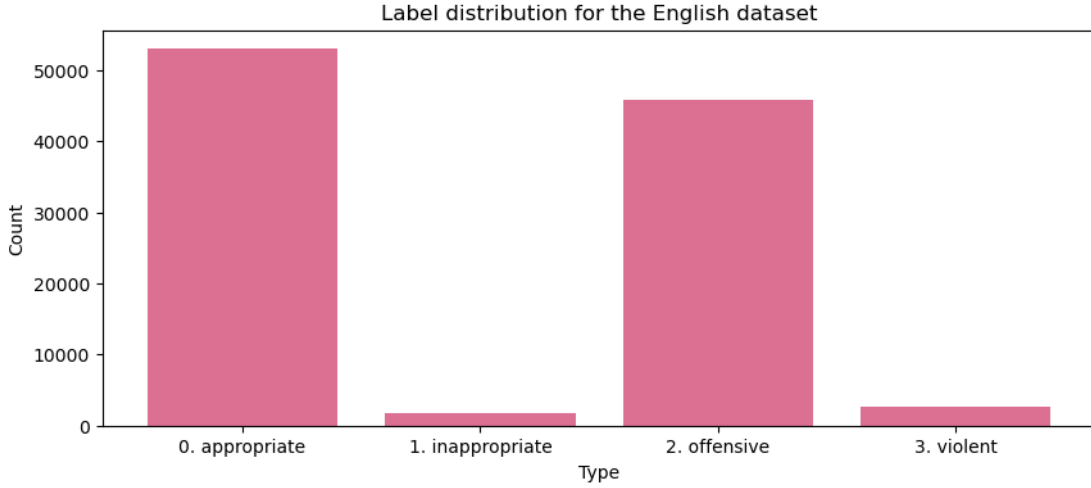


Figure 5.: Label distribution for the English dataset in Kralj Novak et al. (2022)

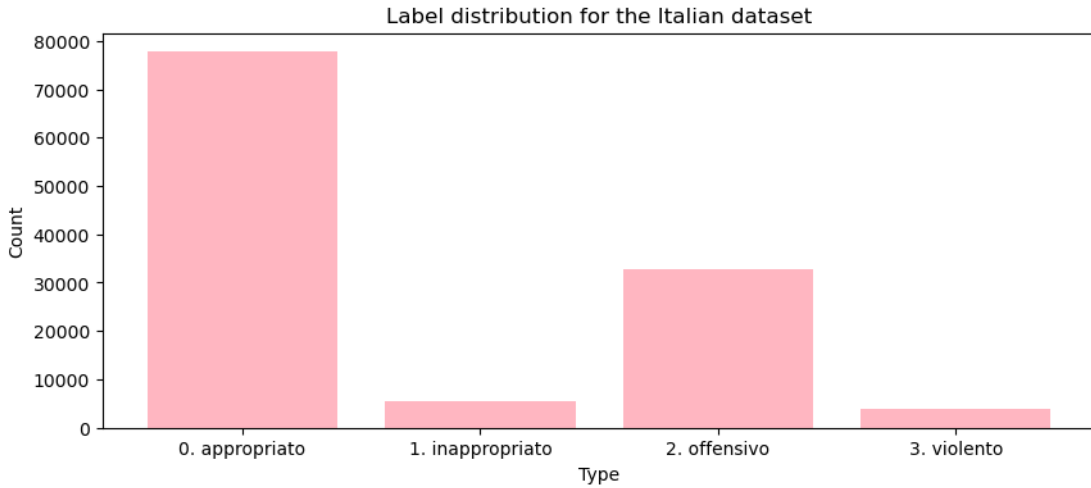


Figure 6.: Label distribution for the Italian dataset in Kralj Novak et al. (2022)

A. Appendix

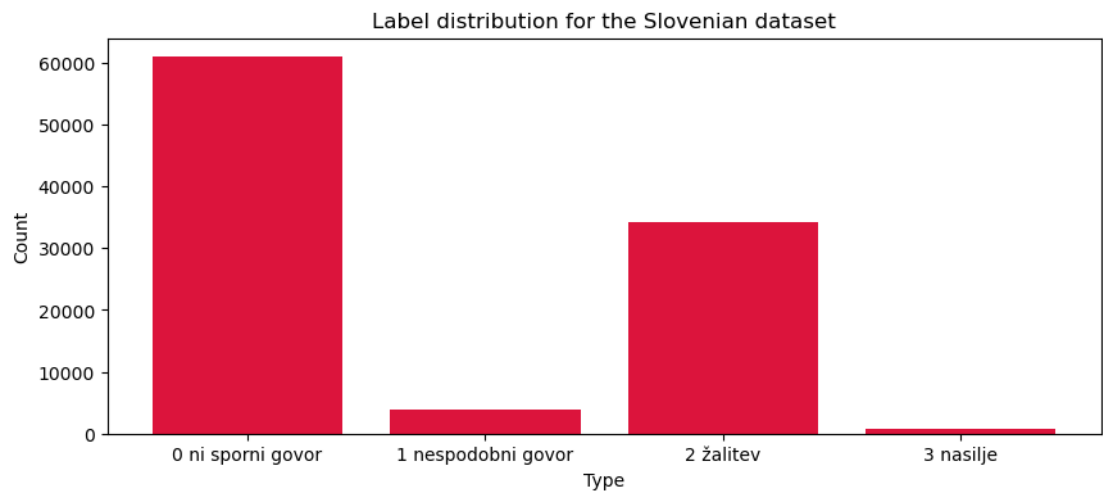


Figure 7.: Label distribution for the Slovenian dataset in Kralj Novak et al. (2022)